Feedback Control Systems

ANALYSIS, SYNTHESIS, AND DESIGN

J-C GILLE
M. J. PELLEGRIN
P. DECAULNE

Professors, École Nationale Supérieure
de l'Aéronautique, Paris

FOREWORD BY

GORDON S. BROWN

Head, Department of Electrical Engineering
Massachusetts Institute of Technology

ALBERT C. HALL

Director of Research, The Martin Company

McGRAW-HILL BOOK COMPANY, INC.

New York     Toronto     London

1959
FEEDBACK CONTROL SYSTEMS

Copyright © 1959 by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers. Library of Congress Catalog Card Number 58-9856

THE MAPLE PRESS COMPANY, YORK, PA.
FOREWORD

Within the last twenty years, rapid increases in control-system technology have resulted from the progressively larger demands placed upon this field. A thorough knowledge of this technology is a requirement of engineers in all fields. In this period, we have seen the number of engineers specializing in feedback controls grow from a few hundred to tens of thousands. As a direct result of the requirements of radar and fire control systems developed during the Second World War, automatic-control technology grew from an art to a science. Also, during the last war, guided-missile development was initiated, and this field has grown until, in the United States alone, the effort now is measured in an annual expenditure of many billions of dollars. Clearly, guided-missile development could not occur without rapid advances in control-system technology and its closely related field, systems engineering.

Today, with the initiation of space missions, being placed on control technology are requirements of another order of magnitude above those which the engineer is now meeting. The optimum control of space vehicles, particularly in the case of longer-range missions, will provide a tremendous saving in propulsion energy. However, such a saving requires that extremely close control be exercised in the early phases of the orbit. Other problems in space missions indicate that nonlinear techniques may provide fundamental improvements over linear control techniques.

The student of control-system technology can today look forward to participating in the solution of a number of fascinating problems in the guidance, control, and navigation of space vehicles.

As a result, first, of technology available because of the stimulation of military needs and, second, of the rapid increases in labor costs, studies are under way in most industries to find ways of reducing production costs by making use of automatic-control techniques. Indeed, even before 1930 the chemical and oil industries pioneered the use of automatic systems in continuous and batch process control. The last year has seen the promising use of numerically controlled machining operations in the aircraft industry and elsewhere. Continuous study is being given to the development of new manufacturing processes and, indeed, even to the control of businesses by techniques that draw heavily on feedback-control technology.

Thus we see automatic-control techniques being applied on the one hand to control of space vehicles and on the other to the reduction of manufacturing costs of consumer products. We believe that this book by French engineer-scientists represents a notable contribution to the organized and unified body of knowledge in the broad area of automatic
control and will provide a point of view that will be helpful to students interested in widely separated fields. The book provides an up-to-date presentation of contemporary American technology. But what is more important, it provides a very thorough English-language version of contemporary European and Russian work in both the linear and non-linear areas. We believe that this book will give an engineer a helpful measure of the European and Russian approach to control-system analysis. Not only have the authors had direct oral communication with their counterparts in most European countries, but they have studied, appraised, and extracted from almost every important published French, German, and Russian work in the field up to about 1957.

Another distinguishing quality of the work is that the philosophy and technique of presentation are European rather than American. This should afford refreshing reading and some valuable new learning for American engineers and scientists. Many ideas and concepts often become clearer and more broadly understood when they are presented in a way that differs from the way one originally encountered them. Engineering students, college professors, and scientists will recognize the more than average attention that is given to the pure and applied mathematics underwriting modern automatic-control theory and to the physical limitations of their use. All three authors received their early technical education in the French university system, in which the mathematical level is notably high. Their association with American education was brief, but lasting. It was initially in Cambridge, Massachusetts, where they followed American courses of instruction and saw some examples of American automatic-control practice. By their presence as students in the control courses taught by us in the Department of Electrical Engineering at MIT, our academic program was handsomely enriched. In recent years, the authors have organized one of the most elaborate programs for the teaching of feedback control systems in Europe. Furthermore, every year J-C Gille lectures on automatic control as a visiting professor at Laval University in Quebec.

Students and educators will appreciate the large number of illustrative problems included with the work, and they will be glad to learn that instructors can procure from the authors the solutions to many of them. The five-language glossary of automatic-control terms with lucid notes explaining German and Russian concepts is also an unusual component of a control book.

The interest of this publication has already been recognized throughout Europe, and it has been translated into three European languages: German, Russian, and Polish. We believe that it will also be appreciated by professors, engineers, and students in the United States.

Gordon S. Brown
Albert C. Hall
PREFACE

Conduire par ordre mes pensées, en commençant par les objets les plus simples et les plus aisés à connaître, pour monter peu à peu comme par degrés jusques à la connaissance des plus composés.

R. Descartes, Discours de la Méthode, 1637

This book is intended (a) as a servomechanism textbook for seniors and first-year graduate students, and (b) as a reference book for the engineer working in the automatic-control field. The aim of the authors is, by progressing from the elements of the subject, to provide in a single book a treatment both of the over-all theory of feedback control systems and of their components. The book is an adaptation of "Théorie et Technique des Asservissements" published in French by Dunod, Paris (first edition, 1956; second edition, 1958).

The work corresponds to lectures and problems given since 1950 at the École Nationale Supérieure de l'Aéronautique in Paris and since 1956 at the Faculté des Sciences of Laval University in Quebec. It is strictly progressive in nature, starting from the elements of the subject; has no discontinuity of treatment; and constitutes a self-contained entity. It is hoped that these features will help accentuate the present trend toward the teaching of at least the fundamentals of servomechanisms to seniors, and even to juniors, in electrical engineering, and to engineers in others fields besides electrical engineering. This would have been difficult a few years ago. At the present time, however, much has been written on all phases of the subject; the type of presentation offered here makes it possible to gain a thorough understanding of the different aspects of the subject with a minimum of prerequisites.

Prerequisite for the book are the following: in mathematics, familiarity with complex numbers, including complex exponentials (but complex variable theory, e.g., the residue theorem, is not a prerequisite), the concept of a differential equation (not the solution), and the partial-fraction expansion of polynomial fractions; in physics, knowledge of the basic law of dynamics $f = ma$ and Ohm's and Lenz's laws.

For those who would use this book as a text, it may be noted that the subject matter contained in it, though presented in a coherent sequence, can be divided into two categories: (1) material constituting a consistent textbook in the basic theory of servomechanisms and (2) material constituting additional data on particular or more advanced methods.

1. The chapters falling in the first category are, essentially, concerned with frequency-response methods for linear systems and describing-function analysis for nonlinear systems. The corresponding chapters and sections, taken as a
whole, constitute a consistent textbook which can be used for teaching seniors or as a review for graduate students. They are:

a. In Part 1 (general linear dynamics): Chaps. 2, 3, 4, and 5, Chap. 6 except Sec. 6.5, Chap. 7, Chap. 8 except Sec. 8.5, and Sec. 9.1.


c. In Part 3 (nonlinear servo systems): Chaps. 22 and 24 except Sec. 24.5.

Excluded from each of the chapters listed is the material printed in reduced type; this material is concerned with examples or additional data.

2. The remainder of the book is concerned with other problems and approaches. The corresponding sets of chapters and sections (listed below) have been made independent of each other, in order that the instructor may choose which he prefers to emphasize particularly for the more advanced, e.g., graduate students. The material falling in this second category is:


b. Transfer matrices and applications (impedance matching, multiple servo systems): Chaps. 10 and 21.

c. Statistical approach: Chap. 12, Secs. 17.4, 19.2, and 24.5.1.


e. Special nonlinear problems: most of Part 3, i.e., Chap. 23, Sec. 24.5, and Chaps. 25 to 28.

f. Components of servo systems: all of Part 4, i.e., Chaps. 29 to 33, with emphasis on servomotors (Chaps. 30 to 32).

For the engineer actually working in the automatic-control field it was believed desirable to fill a gap in existing servo literature by writing a book dealing with both the general theory and the components of servo systems.

In fact, the actual limitations to the improvement of servo systems are often to be found not in the methods available for analysis and synthesis, which are excellent and often optimum, but incorporated at the very outset of the design procedure—by the choice of the fixed components. Unfortunately, little of the existing literature is devoted to servocomponents in relation to their influence on over-all system performance. Most books on analysis and synthesis assume that the fixed components have been "properly" chosen. On the other hand, few works devoted to servocomponents are concerned primarily with the part the latter play as elements of the whole system. Accordingly, it has been the aim of the authors to help both the practicing engineer and the student by offering a systematic account which places equal emphasis on over-all theory and the problems pertaining to the component parts, and which links these two subjects intimately.

Also, it was believed worthwhile to introduce to the American public some recent European works on servo which are little known in English-speaking countries. In particular, recent Soviet advances concerning

---

1 In the author's teaching in Paris, two terms are devoted to teaching the items 1 (i.e., basic servo theory) and 2f (i.e., servo components). In Quebec, items 1 are taught in two undergraduate one-term courses (covering chapters from the first and second part of the book, respectively), whereas item 2e (nonlinear servomechanisms, i.e., Part 3 of the book) is the subject of a one-term graduate course.
nonlinear stability and on-off systems have been presented in terms consistent with American viewpoints and work in the field. A five-language glossary of important terms in automatic control has been placed at the end of the book. It is hoped that this will encourage English-speaking engineers to read foreign publications on the subject and will similarly help foreign experts to read literature in English.

Limiting the scope of the book was one of the most difficult problems the authors had to face. The criterion of limitation they have tried to follow consists in emphasizing problems in proportion to their practical importance for the man working in the servomechanism field. Thus, concepts that originate in the communication or circuit fields have been treated as servo concepts: i.e., stress has been laid on the new, often original, aspect they have taken on as they have been incorporated into the servo field. This applies especially to noise and nonlinearities, which are described from the point of view of the automatic-control field, rather than in more general or more theoretical terms. Moreover, only those concepts of information theory that are, at the present time, of direct aid to servomechanism engineers are encompassed.

This emphasis on the viewpoint of the servomechanism engineer has led to stressing some properties far more than is the case in most textbooks. In particular, physical limitations and pitfalls in use of theory have been given special attention. For instance, nearly as much is written about what transfer functions cannot be used for as about what they can be used for. Validity limitations at “high” frequencies, which, in practice, are often amazingly low; corrections for finite output impedance; etc., are often of primary interest for the man in the field. Accordingly, they have been given a correspondingly important treatment.

In conclusion, the authors gratefully acknowledge their indebtedness to all who cooperated in the preparation of the book, and especially of its edition in English.

First, to their former professors of the postwar years: Prof. Ph. LeCorbeiller at Harvard; the late Donald P. Campbell, Professors C. S. Draper, M. F. Gardner, Y. W. Lee, and R. C. Seaman at the Massachusetts Institute of Technology; and Prof. L. Gauthier of the Faculté des Sciences of Nancy. Special mention is due Prof. G. S. Brown, former Director of the MIT Servomechanism Laboratory, who not only initiated the authors in automatic control but also organized the first consistent teaching of servomechanisms, and to Dr. A. C. Hall, former Director of the MIT Dynamics and Automatic Control Laboratory (DACL), whose teaching of automatic-control theory and its application in aeronautics has had a profound and lasting influence on the authors’ subsequent work.

Thanks are also due to those who gave the authors an opportunity for developing research and teaching in the servo field: M. Decker, chief engineer in charge of guided-missile research and development at the Ministère de l’Air, and P. de Valroger, Director of the Ecole Nationale Supérieure de l’Aéronautique. Also, to all those who cooperated with the authors in their research or teaching activity, with special reference
to Dr. J. A. Hrones, Director of the DACL, and F. H. Raymond, Director of the Société d'Electronique et d'Automatisme at Courbevoie, France.

Finally, thanks are due to those who made the preparation of an edition in English possible. The greater part of this edition has been written by L. Mozart Boisvert, of Quebec; Paul Jaillard, Michel Lermoyez, and Mostyn Mowbray, of Paris; and the authors themselves. Chapters or parts thereof were also translated by Dr. R. Scanlan and J. Haas, of Paris; Dr. M. Boisvert, J. Dumas, and E. Vaillancourt, of the Laval staff; Y. Caron, R. Arsenault, P. Binet, M. Guérin, C. Lemyre, H. Rouleau, J. Saint-Onge, and J. Savard, of Quebec.

Since most of the translators were French-speaking, the problem of gaining a good style in English was extremely important. Helpful suggestions were obtained from and numerous corrections were made by Dr. G. Reethof, of Vickers, Inc.; R. Wilcox, of Hycon Co. and Northwestern University; Dr. B. C. Sharpe, of the University of Michigan; J. Kaiser, of MIT; Dr. P. Nikiforuk, of Canadair; D. Flemming, A. Paris, G. Spindler, and R. Wall, of the Canadian Armament Research and Development Establishment (CARDE). The help of P. Nikiforuk and A. Paris, who kindly reread, criticized, and greatly helped to unify many chapters in the course of the preparation of the book, was so important that it can be said, without exaggeration, that the edition is nearly as much their work as the work of the authors themselves.

Finally, acknowledgment is due to Prof. J. G. Truxal, with whom the authors had extremely stimulating exchanges and whose valuable suggestions led to substantial modifications with respect to the first French edition. Acknowledgment must also be made to Dr. T. J. Higgins, of the University of Wisconsin, who read and polished the whole of the translation in English.

J-C Gille
M. J. Pélegrin
P. Decaulne
# CONTENTS

Foreword .............................................. v
Preface ................................................ vii
List of Selected Figures ......................... xvi
List of Selected Tables ......................... xvii
List of Symbols ..................................... xviii

INTRODUCTION

Chapter 1. The concept of feedback control ........ 1
  1.1 Open-loop control systems .................... 1
  1.2 Feedback, or closed-loop, systems .......... 7
  1.3 Examples of servo systems. Philosophical considerations 11
  1.4 Introduction to servo problems. Outline of following chapters 20

PART ONE. DYNAMICS OF LINEAR SYSTEMS

Chapter 2. Deriving the equations of linear systems ........ 25
  2.1 General considerations ...................... 25
  2.2 A few notes on linearity .................... 26
  2.3 Electric lumped-constant systems .......... 29
  2.4 Other systems: analogy ................... 37

Chapter 3. Transients .................................. 43
  3.1 The concept of transients .................. 43
  3.2 The importance of transients ............. 46
  3.3 Time response of linear systems .......... 47

Chapter 4. The Laplace transform .................. 53
  4.1 Definition and fundamental properties .......... 53
  4.2 Other important properties ............... 56
  4.3 Impulse functions .......................... 58
  4.4 Other properties of the Laplace transform .. 61
  4.5 Basic concepts concerning Fourier transforms .. 62
  4.6 Obtaining the time response of a linear system . 63
      Appendix .................................. 64

Chapter 5. First-order systems .................. 68
  5.1 Transient response .......................... 68
  5.2 Harmonic response .......................... 72
  5.3 Response to an arbitrary input, transfer function . 77
  5.4 Examples of first-order systems .......... 78
  5.5 Lags and time constants .................. 80
## CONTENTS

### Chapter 6. Second-order systems
- 6.1 Frequency-response curves and loci ........................................ 83
- 6.2 Transfer functions and loci .................................................. 83
- 6.3 Transient response ............................................................. 87
- 6.4 Performance of second-order systems ..................................... 95
- 6.5 The poles of the transfer function, as located in the complex plane 109
- 6.6 Examples of second-order systems. Their importance .......... 112

### Chapter 7. Transfer functions
- 7.1 The concept of transfer function ............................................ 114
- 7.2 Application to steady state .................................................. 118
- 7.3 Application to transients ..................................................... 122
- 7.4 Philosophy of the transfer function: the two practical approaches 125
- 7.5 Determination of transfer functions ....................................... 126

### Chapter 8. The harmonic approach: transfer loci
- 8.1 Nyquist loci and equivalent representations .......................... 132
- 8.2 Nichols loci. Practical construction of transfer loci (Bode plots) 139
- 8.3 Relation to the time response ............................................... 143
- 8.4 Principle of experimental techniques for the determination of transfer loci ................................................. 150
- 8.5 Relation between amplitude and phase .................................... 154
- 8.6 Synthesis: conclusion on the harmonic approach ................... 158

- 9.1 The pole configuration and stability ...................................... 159
- 9.2 Algebraic stability criteria .................................................. 163
- 9.3 The pole-zero configuration ................................................ 169
- 9.4 Concluding remarks on the pole-zero approach ....................... 176

### Chapter 10. Transfer matrices
- 10.1 Transfer matrices ........................................................... 178
- 10.2 Cascade-system synthesis with predetermined transfer functions 187
- 10.3 Impedance matching for maximum power transfer ................... 190

### Chapter 11. Linearity domain
- 11.1 Amplitude linearity domain ............................................... 193
- 11.2 Frequency linearity domain .............................................. 195
- 11.3 Synthesis: linearity domain .............................................. 198

### Chapter 12. Statistical considerations
- 12.1 Introduction ................................................................. 200
- 12.2 Concept of frequency spectrum .......................................... 201
- 12.3 Fundamental relation, applications ..................................... 206
- 12.4 Fundamentals of information theory .................................... 210

---

### PART TWO. LINEAR SERVO SYSTEMS

#### Chapter 13. Transfer functions and transfer loci of feedback control systems 215
- 13.1 Open-loop and closed-loop transfer functions ....................... 215
- 13.2 Harmonic approach. Graphical interpretation of $H = KG/(1 + KG)$ 223
- 13.3 Open-loop gain adjustment .............................................. 229
- 13.4 Influence of time lags .................................................... 234
CONTENTS

  14.1 The concept of root locus 237
  14.2 Construction of root loci 239
  14.3 Stability of feedback control systems 253

Chapter 15. The steady state of feedback control systems 260
  15.1 Position error 260
  15.2 Generalization 262
  15.3 Application to automatic piloting 265

Chapter 16. Graphical stability criteria for feedback control systems 268
  16.1 Nyquist’s criterion 268
  16.2 Various applications 276
  16.3 The left-hand criterion 279
  16.4 Practical concept of stability 280

Chapter 17. Performance criteria 284
  17.1 Ideal control systems and practical specifications 284
  17.2 A set of general-purpose criteria 286
  17.3 Evaluating the effect of time lags 289
  17.4 More elaborate criteria 290
  17.5 Practical use of performance criteria 294

Chapter 18. Compensation of feedback control systems 295
  18.1 Concept of compensation. Compensating networks 295
  18.2 Phase-lead compensation, or derivative control 297
  18.3 Phase-lag compensation, or integral control 307
  18.4 Combination of lead and lag compensation 312
  18.5 Generalization 314

Chapter 19. Servo-system synthesis 318
  19.1 Synthesis by means of pole-zero configuration 319
  19.2 Statistical approach to servo-system synthesis 326
  19.3 Conclusion on servo-system synthesis 334

Chapter 20. Sampled-data systems 335
  20.1 Linear sampled-data systems 336
  20.2 z transforms 341
  20.3 Frequency analysis of sampled-data systems 352
  20.4 Remarks on sampled-data servo synthesis 361

Chapter 21. Multiple feedback control systems 364
  21.1 Definition, examples 364
  21.2 Analog computers and simulators 366
  21.3 Some remarks on general theory 370

PART THREE. NONLINEAR SERVO SYSTEMS

Chapter 22. General remarks on nonlinear systems 374
  22.1 General remarks on linearity and nonlinearity 374
  22.2 Definition and classification of nonlinear systems 379
22.3 Nonlinearities most frequently encountered in servo systems .......... 381
22.4 Method of investigation .................................. 389

Chapter 23. Transients in nonlinear servo systems ............... 394
23.1 Time response of on-off servo systems .................. 394
23.2 Time response of servo systems with one nonlinear element .... 398
23.3 Conclusion ...................................................... 402

Chapter 24. The first-harmonic approximation ...................... 403
24.1 The concept of the describing function .................... 403
24.2 Describing functions of common nonlinear elements .......... 406
24.3 Stability of servo systems with one nonlinear element ...... 414
24.4 Performance and compensation ............................ 421
24.5 Additional comments on describing functions .............. 423

Chapter 25. The Poincaré approach in the phase plane .......... 432
25.1 The phase plane ............................................. 432
25.2 Applications to some simple nonlinear servo systems ...... 436
25.3 Optimum on-off servo systems ............................ 443

Chapter 26. Oscillations of on-off control systems ............... 446
26.1 Introduction .................................................. 447
26.2 Self-oscillations of on-off regulators ..................... 449
26.3 Forced oscillations in on-off servo systems .............. 475
26.4 Stability of periodic states ............................... 477

Chapter 27. Additional methods applicable to nonlinear systems .. 484
27.1 Poincaré's theorems ........................................ 484
27.2 Limit cycles, stability ...................................... 491

Chapter 28. Linearization. Final remarks ......................... 498
28.1 Linearization ................................................ 498
28.2 Concluding remarks on nonlinear systems ................. 501

PART FOUR. COMPONENTS OF SERVO SYSTEMS

Chapter 29. Error-sensing devices ............................... 503
29.1 Some remarks on sensing devices and noise ................ 503
29.2 Classification of sensing devices .......................... 511
29.3 Variable-impedance sensing devices ........................ 513
29.4 Generators as sensing devices ............................. 523
29.5 Vacuum tubes as sensing devices ........................... 531
29.6 Gyrosscopic sensing devices ............................... 532
29.7 Accelerometers .............................................. 535
29.8 Choice of a sensing device ................................. 537

Chapter 30. Determination and design of servomotors .......... 539
30.1 The problem .................................................. 539
30.2 Discussion of inertias and gear ratios .................... 542
30.3 Determining the transfer function of a motor from its characteristic curves ...................... 549
30.4 Points for comparison of the various types of motors .... 552
CONTENTS

30.5 Positioning the characteristic and design calculations of the motor in compliance with the specifications. ........................................ 554

Chapter 31. Electric servomotors ........................................ 569
  31.1 General features of motors ........................................ 569
  31.2 Effect of inductance on time constants ............................ 576
  31.3 Field-controlled motors .......................................... 581
  31.4 Armature-controlled motors ...................................... 586
  31.5 Alternating-current motors ...................................... 591
  31.6 Numerical data for electric servomotors ......................... 600

Chapter 32. Hydraulic servomotors .................................... 604
  32.1 Description of hydraulic motors ................................ 604
  32.2 Summary of the fundamental laws of hydrodynamics .............. 613
  32.3 Transfer function of a hydraulic transmission ................. 619
  32.4 Equations of valve-controlled hydraulic motors ............... 623
  32.5 Servocontrols or boosters ...................................... 628
  32.6 Additional remarks .............................................. 631
  32.7 Numerical data sheets for hydraulic servomotors ............. 635

Chapter 33. Amplifiers and preamplifiers ............................. 641
  33.1 The general problem ............................................. 641
  33.2 Vacuum-tube amplifiers ......................................... 644
  33.3 Relay amplifiers .................................................. 645
  33.4 Rotating amplifier. Ward-Leonard system ....................... 648
  33.5 Machines derived from the Ward-Leonard system ............... 655
  33.6 Mixed electronic-tube amplifiers ............................... 660
  33.7 Magnetic amplifier ............................................... 665
  33.8 Rectifying, demodulating, filtering ............................ 668
  33.9 Example .......................................................... 675
  33.10 Transistors .......................................................... 681
  33.11 Comparative table of the principal types of amplifier ....... 696

PART FIVE. BASIC DESIGN OF A SERVO SYSTEM

Chapter 34. Two examples of basic system design ....................... 698
  34.1 Some major principles .......................................... 698
  34.2 First example: positional servo for a measuring system ....... 699
  34.3 Second example: autopilot for a guided missile ................ 710

Problems ............................................................................. 719

Five-language Glossary of Important Automatic-control Terms .......... 757

Alphabetical List of Glossary Words ..................................... 768

Complementary Notes to Glossary ....................................... 771

Selected Bibliography ....................................................... 775

Index .................................................................................. 783
LIST OF SELECTED FIGURES

1-32. Block diagrams of a feedback control system .......................... 11
1-47. Components of a servo system .............................................. 21
5-6. Response of first-order systems to \( u(t) \) and \( 1 - u(t) \) ........... 70
5-12. Bode plots of first-order systems ........................................ 75
5-13. Alignment chart for first-order systems .................................. 76
6-7. Frequency response of second-order systems ............................. 90
6-8. Nyquist loci of second-order systems ..................................... 91
6-9. Bode plots of first- and second-order systems (twofold reduction of
Charts 1 and 2 at back of the book). Nichols loci of first- and
second-order systems .............................................................. 92–94
6-11. Unit-step response of second-order system ............................. 97
6-14. Resonance ratio vs. damping ratio ....................................... 101
6-16. Overshoot vs. damping ratio ............................................... 102
6-17. Resonant and natural frequencies vs. damping ratio .................. 102
6-20. Cutoff frequencies of second-order systems ............................. 104
6-21. Response time of second-order systems .................................. 105

6-23 to 6-25. Determining the transfer function of a second-order system from
the step response ................................................................. 107–109

8-1 to 8-6. Nyquist locus of airplane in longitudinal motion ............... 133–136
8-16. Nichols loci for \( 1/s(1 + 2zs + s^2) \) ....................................... 140
8-21. Chart for determining step response from Bode plots ................. 146, 147
13-4. Nyquist loci for \( 1/s(1 + s)(1 + as) \) ....................................... 217
13-5. Nichols loci for \( 1/s(1 + s)(1 + as) \) ....................................... 218

13-13b and c. The Hall chart and the Nichols chart (twofold reduction of
Chart 3 at the back of the book) .................................................. 225
18-12. Frequency response of \( (1 + azs)/(1 + rs) \) .............................. 300
18-13. Nichols loci of \( (1 + azs)/(a(1 + rs)) \) .................................... 301
21-2. Block diagram of multiple servo system .................................. 365
24-8. Describing function of on-off element with dead zone and
hysteresis .................................................................................... 410
24-9. Critical locus of on-off element with dead zone and hysteresis .... 411
24-12. Describing function of linear system with saturation and dead
zone .......................................................................................... 412
24-14. Chart for obtaining describing function of any frequency-inde
pendent nonlinear element without hysteresis ................................. 414
26-8. Hamel loci for \( L(s) = 1/s(1 + s)(1 + as) \) ............................... 456
26-9. Hamel loci for \( L(s) = 1/s(s^2 + 2\zeta s + 1) \) ........................... 457
26-10. Hamel loci for \( L(s) = 1/s^2(1 + s)(1 + as) \) .......................... 458
26-18. Tsypkin loci for \( L(s) = 1/s(1 + s)(1 + as) \) ............................ 465
26-19. Tsypkin loci for \( L(s) = 1/s(s^2 + 2\zeta s + 1) \) ........................ 466
26-20. Tsypkin loci for \( L(s) = 1/s^2(1 + s)(1 + as) \) ......................... 467
26-24. Generalized Tsypkin loci for \( L(s) = 1/s(1 + s)(1 + as) \) when
on-off element has dead zone .................................................. 470–474
30-16. Envelope of operative servomotor characteristic .................... 557
32-20. Pressure loss in a convergent for hydraulic fluids ................. 618
Charts at the back of the book

1. Amplitude response of first- and second-order systems (Bode plots; 1 cm corresponds to 2 db)
2. Phase response of first- and second-order systems (Bode plots; 1 cm corresponds to 10°)
3. Nichols chart (1 cm corresponds to 2 db and 10°)
4. Chart for determining servomotor characteristic

LIST OF SELECTED TABLES

<table>
<thead>
<tr>
<th>Table No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-1</td>
<td>Laplace transforms of common functions of time</td>
<td>54</td>
</tr>
<tr>
<td>10-1</td>
<td>Quadripole connections</td>
<td>183</td>
</tr>
<tr>
<td>15-2</td>
<td>Steady-state error for servo systems with 0, 1, 2, and 3 integrations when disturbance is applied at the input</td>
<td>264</td>
</tr>
<tr>
<td>20-1</td>
<td>Laplace and z transforms</td>
<td>344</td>
</tr>
<tr>
<td>31-1 to 31-6</td>
<td>Numerical data for commercial electric servomotors</td>
<td>600-603</td>
</tr>
<tr>
<td>32-1 to 32-2</td>
<td>Numerical data for commercial hydraulic servomotors</td>
<td>636</td>
</tr>
<tr>
<td>33-2</td>
<td>Comparison of different types of amplifier</td>
<td>687</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS

(Only those symbols used consistently in several chapters are listed here.)

\( a \) \hspace{1cm} \text{time-constant ratio of phase-lead network}

\( A(\omega) \) \hspace{1cm} \text{amplitude response, magnitude of } H(j\omega) \text{ or } L(j\omega)

\( b \) \hspace{1cm} \text{time-constant ratio of phase-lag network}

\( B(x_1) \) \hspace{1cm} \text{magnitude of describing function}

\( c \) \hspace{1cm} \text{(as subscript)} \hspace{1cm} \text{critical; cutoff}

\( C \) \hspace{1cm} \text{capacitance; torque (also } \Gamma, \tau)\)

\( C_a \) \hspace{1cm} \text{acceleration constant}

\( C_v \) \hspace{1cm} \text{velocity constant}

\( d(t), D(s) \) \hspace{1cm} \text{disturbance}

\( e \) \hspace{1cm} \text{base of natural logarithms (also } e)\)

\( e(t), E(s) \) \hspace{1cm} \text{input, command}

\( e_m \) \hspace{1cm} \text{threshold}

\( e_M \) \hspace{1cm} \text{saturation}

\( f \) \hspace{1cm} \text{friction coefficient}

\( G(s) \) \hspace{1cm} \text{frequency-dependent part of open-loop transfer function } KG(s)

\( h \) \hspace{1cm} \text{hysteresis of on-off element}

\( h(t) \) \hspace{1cm} \text{unit-impulse response}

\( H(s) \) \hspace{1cm} \text{transfer function}

\( i(t), I(t) \) \hspace{1cm} \text{current}

\( \text{Im} \) \hspace{1cm} \text{imaginary part of}

\( j \) \hspace{1cm} \sqrt{-1}

\( J \) \hspace{1cm} \text{moment of inertia}

\( k \) \hspace{1cm} \text{constant}

\( K \) \hspace{1cm} \text{gain of linear system (especially, open-loop gain of servo system)}

\( K_a \) \hspace{1cm} \text{gain of linear system with two integrations}

\( K_v \) \hspace{1cm} \text{gain of linear system with one integration}

\( \mathcal{L} \) \hspace{1cm} \text{Laplace transform}

\( \mathcal{L}^{-1} \) \hspace{1cm} \text{inverse Laplace transform}

\( L(s) \) \hspace{1cm} \text{function combining linear blocks of servo system (see Sec. 22.4)}

\( m \) \hspace{1cm} \text{mass}

\( M \) \hspace{1cm} \text{magnitude of output of on-off element}

\( p_a \) \hspace{1cm} \text{pole of transfer function}

\( q(t) \) \hspace{1cm} \text{unit-step response}

\( Q \) \hspace{1cm} \text{resonance ratio}
LIST OF SYMBOLS

\( R \) resistance; Reynolds number
\( r(t), R(s) \) response, output
\( \text{Re} \) real part of
\( s \) Laplace variable
\( t \) time
\( T \) time lag, time constant; period
\( u \) nondimensional frequency \((\omega/\omega_n \text{ or } T\omega)\)
\( u(t) \) unit-step function
\( u_1(t) \) unit-impulse function [also \( \delta(t) \)]
\( U(\omega) \) real part of \( L(j\omega) \)
\( v(t) \) velocity
\( V(t) \) voltage
\( V(x,y) \) Liapunov function
\( V(\omega) \) \( jV(\omega) \) is imaginary part of \( L(j\omega) \)
\( w(t) \) output of nonlinear element
\( x \) time-dependent variable
\( x_1 \) magnitude of harmonic input to nonlinear system
\( y \) time-dependent variable
\( z \) damping ratio (also \( \xi \)); \( e^{\tau} \)
\( Z \) impedance
\( z_i \) zero of transfer function

\( \alpha \) real part of \( s \); gear ratio
\( \alpha(t) \) angle
\( \gamma(t), \Gamma(s) \) acceleration [also \( a(t) \)]
\( \Gamma \) torque (also: \( C, \tau \))
\( \delta(t) \) unit-impulse Dirac function [also \( u_1(t) \)]
\( \delta(t), \Delta(s) \) deflection of aircraft control surface
\( \Delta \) dead zone of on-off element
\( \varepsilon \) positive quantity; base of natural logarithms (also \( e \))
\( \varepsilon(t), \varepsilon(s) \) error
\( \varepsilon_1 \) magnitude of \( \varepsilon \) when sinusoidal
\( \xi \) damping ratio (also \( z \))
\( \eta(t), H(s) \) shaped error signal
\( \theta(t), \Theta(s) \) angular position
\( \theta_m(t) \) angular position of motor shaft
\( \theta_r(t) \) output angular position
\( \lambda \) parameter
\( \Lambda(\omega) \) expression for Tsypkin locus
\( \tau \) lag, time constant; torque (also \( C, \Gamma \))
\( \varphi \) phase angle
\( \varphi(\tau) \) autocorrelation function
\( \varphi_{st}(\tau) \) crosscorrelation function
\( \Phi \) magnetic flux
\( \Phi(\omega) \) phase response, argument of \( H(j\omega) \) or \( L(j\omega) \); frequency spectrum
\( \psi \) angle in \( s \) plane (see Fig. 6-27)
LIST OF SYMBOLS

$\Psi(x_1)$ argument of describing function

$\omega$ angular frequency

$\omega(t)$ angular velocity

$\Omega$ angular velocity; sampling frequency $2\pi/T$

The Laplace and $z$ transforms of $f(t)$ are generally denoted by $F(s)$ and $F^*(z)$, respectively.
INTRODUCTION

CHAPTER 1

THE CONCEPT OF FEEDBACK CONTROL

Summary

1. Open-loop control systems.
2. Feedback, or closed-loop, systems. Servo systems and regulators.
4. Introduction to servo problems. Outline of following chapters.

A servo system is a control system possessing these two features: (1) power amplification and (2) feedback comparison. The meaning of these defining features will first be explained; then examples of servo systems will be given, and some of the problems relevant to such systems will be presented. This discussion will lead into the general plan of the book, which will subsequently be outlined.

1.1. OPEN-LOOP CONTROL SYSTEMS

1.1.1. Control Systems without Power Amplification. The concept of control can be understood intuitively and can be explained to a first approximation by some simple examples. Consider a crank handle linked through various sets of gears to a shaft holding a mechanical part. The angular position of the crank controls that of the mechanical part. This

- Figure 1-1.

means that, provided the backlash in the gear mechanism is very small, a single-valued relation exists between the input (the angular position of the crank) and the output (the angular position of the mechanical part). This fact is diagrammatically illustrated by Fig. 1-1, which expresses the existence of a relationship between input and output. For this case, the mathematical relationship is of the form output = K \times input, provided the system has very great torsional stiffness. In this diagram, called a block diagram, the block represents the transmission properties of the physical system. The arrows entering and leaving the block represent physical quantities, namely, the controlling input and the resulting output, respectively.

Other physical systems can be considered in a similar manner. The angular position of the steering wheel in an automobile controls the steering of the front wheels. Also, in an aircraft, the assembly of the control itself (stick, rudder control, etc.), the connecting rods, and the control surface constitute a control system, the position of the stick controlling the deflection of the aileron or elevator. This is diagrammatically represented by Fig. 1-2, and provided negligible backlash.
and considerable stiffness are assumed, it is expressed by the relation \( output = K \times input \).

In the preceding examples, all the power delivered at the output is supplied by the input. Such control systems involve no power amplification.

1.1.2. Control Systems with Power Amplification. Consider a separately excited d-c motor whose field is adjustable by means of a potentiometer. The potentiometer controls the motor. In fact, there is a single-valued relationship between the position of the potentiometer and the torque \( C \) on the output shaft of the motor. Diagrammatically, this is shown in Fig. 1-3, the controlling element being the potentiometer and the system to be controlled consisting of the motor and its load. It is to be noted that, if the load is constant or if the load variations take place at a very low power level compared to the power output of the motor, the potentiometer can be considered as controlling the speed of rotation \( N \) of the motor under load (Fig. 1-4).

Similarly, if a switch whose position determines whether or not the motor is excited is provided, the switch is said to control the motor. This means that a relationship exists between the position of the switch and the speed of the motor (Fig. 1-5). The only difference between this and the case in which the motor is controlled by means of a potentiometer is that the relation between the field and the position of the controlling element is continuous in the case of a potentiometer and discontinuous in the case of a switch (Fig. 1-6).

Similarly, for a room heated by an electric radiator, the switch or the adjusting resistance of the radiator controls the room temperature
THE CONCEPT OF FEEDBACK CONTROL

(Fig. 1-7). There is a relation between the switch position and the heat supplied by the radiator. Therefore, for a room with constant heat-transfer characteristics, the switch position controls the room temperature.

In the above examples, contrary to those of Sec. 1.1.1, it may be noticed that the power delivered at the output of the controlled system is not supplied by the controlling element. The energy used to actuate the control itself is supplied by the operator, whereas the energy developed in the controlled system is delivered by a power source (battery, power supply, etc.) feeding the motor or radiator. Very frequently, a small amount of energy at the input is sufficient to control phenomena which involve considerable power. Such systems are for this reason called, in general terms, control systems with power amplification.

It is important to note that the difference between control systems with and without power amplification does not appear in the block diagrams shown.

1.1.3. Extension of the Concept of Control. More generally, when a system subjected to the action of a variable input produces an output which is a function of the input, it is said that the input controls the output. For example, temperature affects the column of mercury in a thermometer. This can be represented by a schematic diagram, as shown in Fig. 1-8. The system has no power amplification. The relation between input and output is given by the calibration curve of the instrument.
Again, an integrating element in an electronic computer can be considered as a system controlled by the input function, which it transforms into its integral, as shown in Fig. 1-9 or expressed by

\[ r = \int_0^t e(t) \, dt \]

In such a case the system has a power amplification, the output energy coming from the high-voltage source feeding the amplifier tubes.

Generalizing, the concept of control is almost identical to that of a relation between two functions (input and output) in a physical system. The system can be considered as controlled by the input, the latter being transformed into the output function. Such a relation is represented by Fig. 1-10, which expresses the existence of a functional relation

\[ F(e, r) = 0 \]

between the input \( e \) and the output \( r \). Frequently, the functional relationship comprises a complicated relation between input, output, and their derivatives and integrals with respect to time:

\[ F\left( e, \frac{de}{dt}, \frac{d^2e}{dt^2}, \ldots, \int e \, dt, \ldots, r, \frac{dr}{dt}, \ldots \right) = 0 \]

Control systems can still be considered and block diagrams for them drawn even though the relation between input and output is too complex to be expressed mathematically, or even estimated. A noteworthy example is the case of control systems involving human operators. For instance, a policeman, by switching the traffic lights from red to green, can control the motion of an automobile with power amplification, as represented in the block diagram of Fig. 1-11, in which the center block represents the "personal equation" of the driver.

1.1.4. Secondary Inputs, Disturbances. The preceding examples are concerned with systems having single inputs and outputs. In actual fact, however, every system is affected by several input quantities. Besides, if a system description is characterized by several parameters, as many outputs as parameters can be considered. To speak of one input and one output is, therefore, a simplification. The choice of output components called for always depends upon the point of view taken in analyzing the system.
1. With regard to the outputs, the most important one is selected as the main output. For example, for a galvanometer the output normally considered is the coil deflection due to the current flowing through the coil, rather than the simultaneously occurring heating, or Joule effect, (Fig. 1-12) unless, for some exceptional reason, emphasis is to be placed on the thermal behavior of the galvanometer.

Similarly, consider the symmetrical motion of an airplane. The elevator deflection is the input variable, while the output is generally (a) the horizontal and vertical components of the forward velocity (Fig. 1-13) if the interest is focused on the motion of the center of gravity or (b) the angular position of the airplane with respect to the horizontal reference (pitch angle $\theta$) or the velocity vector (angle of attack $i$) (Fig. 1-14) if the interest is focused on the motion of the airplane about its center of gravity.

![Fig. 1-12. Galvanometer diagram](image)

![Fig. 1-13. Aircraft diagram](image)

![Fig. 1-14. Aircraft diagram](image)

![Fig. 1-15. Instrument diagram](image)

2. With regard to the inputs, one main input or more is frequently kept separate from the secondary inputs, the former being of major importance for the experimenter. For example, in the case of a measuring instrument the input is the quantity to be measured, the secondary inputs being the external conditions (such as temperature) which can affect the relation between input and output (Fig. 1-15). In the case of a control system, the command, or control input, i.e., the input quantity that should determine the output, is generally considered as the main input.

The effect of the secondary inputs is detrimental to the accurate functioning of the apparatus. Because of this, they are usually called disturbances or unwanted inputs.

For example, the transverse motion of an aircraft is controlled by the aileron and rudder deflections $\alpha$ and $\gamma$. This motion is also influenced by atmospheric disturbances, which consist mainly of forces and torques on the aircraft produced by gusts (Fig. 1-16). On the other hand, if interest is focused on the aircraft’s response to
atmospheric gusts, one is led to consider the control-surface deflections as secondary inputs (Fig. 1-17). As shown by this last example, the viewpoint taken essentially determines the number of secondary inputs which must be considered in addition to the main inputs in the block-diagram representation of the system.

**Fig. 1-16.**

Remark. In some cases, the manner in which the disturbances affect the system is sufficiently well known to enable their introduction into

**Fig. 1-18. Introduction of disturbance.**

**Fig. 1-19. Introduction of disturbance (general case).**

the block diagram with accuracy. Consider, for example, the field-controlled d-c motor shown in Fig. 1-3. Its speed of rotation \( N \) is also a function of the load on the output shaft. Should this load suddenly change, the motor speed will vary although the control current remains constant. This happens, for instance, in a metal shaper when the tool hits the part to be machined. This situation can be represented by

**Fig. 1-20.**

**Fig. 1-21.**
Fig. 1-18, where the symbol appearing between the motor and the load represents, with the + signs, a summation operator. This diagram is a more accurate representation than the general one of Fig. 1-19.

In the same manner, the block diagram describing the roll control of an aircraft subjected to gusts (Fig. 1-20) can be more fully represented, for the particular case in which disturbances produce roll torques, by utilizing the operator as shown in Fig. 1-21.

1.2. FEEDBACK, OR CLOSED-LOOP, SYSTEMS

1.2.1. Inadequacy of Open-loop Control, Necessity for Closed-loop Control. The systems considered in the preceding section have a straightforward relationship between two variables: the input, or control variable, and the output, or controlled variable. In reality, as has just been seen, the relationship between the input and the output is modified by secondary inputs, or disturbances, which are always present. Because of this, one is never certain in practice that a given control input will give the desired output.

1. For example, in the case of the temperature-control system described above, a given position of the potentiometer corresponds to given room temperature only if the temperature-sensitive parameters of the latter are constant. Opening a window obviously modifies the temperature $T$ even though the position of the rheostat remains unchanged.

Consequently, for the effective control of a given output, it is not sufficient to set the control to a certain value. It is necessary (as in everyday life) to check the way in which the command is carried out, and if necessary to modify the command accordingly. This is, in effect, what is done by an operator whose job it is to maintain the temperature of the room at a given value $T_0$. He observes the real temperature $T$ of the room and actuates the control switch as soon as he notes a difference between $T$ and $T_0$.

Automatic-control mechanisms act in the same way. If the room contains a temperature-sensitive bimetal strip arranged to close the switch when $T < T_0$ and open it when $T > T_0$ (Fig. 1-22), then a device exists which makes it possible not only to control the temperature of the room but also to check the working and maintain effective control, despite disturbances that may arise. As will be explained below, $T$ has been slaved to $T_0$, or $T$ has been regulated to $T_0$. The block diagram of such a system may be represented as in Fig. 1-23. The symbol $\otimes$ designates, with its corresponding algebraic signs, a differencing operator which takes the difference of the control input $T_0$ and the output $T$. Physically, this differencing operator takes the form of a comparator component (the bimetal strip) with two inputs $T$ and $T_0$ and whose output is $T_0 - T$. 

![Fig. 1-22.](image)
2. Similarly, consider the electric motor referred to above, whose speed of rotation $N$ depends on the control setting $N_0$ and also on the output load (Figs. 1-18 and 1-19). If it is desired to keep $N_0$ (the running speed) constant, the field voltage regulator can be controlled by the difference between the required speed $N_0$ and the effective speed $N$, the latter being determined by means of a tachometer. The running speed of the motor is thus sure to be controlled effectively, despite external disturbances. In other words, the speed of the motor has been slaved. This is sometimes expressed by saying that the motor itself has been slaved. The corresponding block diagram is shown in Fig. 1-24.

![Fig. 1-23.](image)

![Fig. 1-24.](image)

3. In a similar way, a man's standing position is maintained by means of a certain tonus of muscles known as posture muscles. This necessitates checking; for if this tonus were maintained by open-loop control, a puff of wind would blow the man down. The checking in question is effected through the cerebellum, which constantly receives information on the posture muscles in the form of nerve impulses supplied by strain-gauge-like detectors located in the muscles themselves. If the body leans backward, these detectors signal to the cerebellum, indicating the shortening of the back muscles and the lengthening of the muscles in the front of the body, and the cerebellum orders a new distribution of tonus to counter the threat to equilibrium. In short, the muscular tonus is slaved to the value necessary for maintenance of the upright position. Since this regulation is normally unconscious, it may be considered an automatic control. The mechanism is represented diagrammatically in Fig. 1-25.

![Fig. 1-25.](image)

---

1 It should be noted that Fig. 1-25 includes a number of simplifications: (a) In fact, there is not simply one "tonus" variable, but several variables, namely, the tonus of the different posture muscles involved. (The physiological unit is the muscle fiber!) (b) The maintenance of the upright position is at the same time ensured by sensory
1.2.2. Conclusion and Definitions. The above examples are control systems which exercise a control on the carrying out of the command so that the proper relationship is maintained between the input and the output, despite disturbances which may arise. They are automatic-control systems. Their diagrams, compared with those of Sec. 1.1, show a new structure characterized by the presence of a feedback loop (an input-output comparator). They no longer have the simple forward-path form; instead, they have loops, and they are known as feedback, or closed-loop, systems. The systems considered in Sec. 1.1 are known as open-loop systems.

Note that power amplification may or may not be included in closed-loop systems. When power amplification is present (the system is then a servo system), it does not appear explicitly in the block diagrams.

Fig. 1-26. General block diagram of feedback control system.

Fig. 1-27. Forward path.

Fig. 1-28. Feedback path.

1.2.3. Remarks on the Definition of a Servo System. 1. Servo systems have been previously defined as systems having two features: (1) power amplification and (2) feedback comparison. The feedback comparison introduces a closed loop in the block diagram, and the power stage is controlled by the error between the input $x$ and the output $y$, as detected by the error-sensing device. It can be said, in short, that the forward path transmits the power, while the feedback path pertains to the accuracy of the system, so that such a system provides a compromise between the usually contradictory requirements of power and accuracy.

2. The definition just given is accepted by the majority of authors. However, it is well to bear in mind that the terminology employed does not have unanimous approval. In particular, certain authors do not include in the definition the proviso "with power amplification," thus making no distinction between servo systems and feedback control systems. Others add to the definition of servo systems the restriction that the presence of a static error $\varepsilon$ results in a constant output rate $dr/dt$ or an impressions supplied by the eye and the inner ear.

The fact is that all physiological regulations are complex, and the diagrams representing them are greatly simplified. The utility of these diagrams is sometimes questionable. Here they seem to correspond to the real physiological mechanism in question, but more often they must be considered as simplified representations. The present knowledge of physiology is, in general, not sufficient to enable us to state whether or not diagrams of this type represent the effective functioning of the "human machine."
acceleration $d^2r/dt^2$, that is, that there must be at least one integration in the forward path (Chap. 15).

A detailed consideration of the reasons given by the proponents of one definition or another seems out of place, especially in an introductory chapter. In any case, the interest of a formal definition is somewhat academic, for the very concept of a servo system is by no means absolute (see Sec. 1.3.5). But once a definition is adopted, it is of the utmost importance to remember that other authors may not necessarily use it, and this should always be borne in mind.

The term servomechanism is sometimes used as an approximate synonym of servo system. However, “servomechanism” should preferably be used only with reference to servo systems whose final stages involve mechanical parts. Thus, in the thought of the authors, a turbine regulator or a gun-turret servo system is a servomechanism, whereas an AVC\(^1\) or a temperature control is merely a servo system.

3. So far, the servo systems dealt with have consisted of one control input and one output with unity feedback, i.e., comparison of the input with the output and not with a function of the output. In due course, the concepts already developed will be generalized and will include concepts of feedback control systems with transfer functions in the feedback loops (Chap. 13) and of multiple feedback control systems (Chap. 21). These generalizations raise no conceptual difficulties.

1.2.4. Regulators and Servo Systems. As control systems, all servo systems with unity feedback tend to equalize input and output. But depending on their applications, the following distinctions are made:

1. Regulators, which are control systems where the control input is constant ($e = e_0$) or varies by steps, are designed to make the output

![Fig. 1-29. Regulator.](image)

![Fig. 1-30. Servo system (in the strict sense).](image)

equal to the input despite disturbances in the system. Typical examples are a Watt regulator for a steam engine and an aircraft-heading stabilizer. The purpose of a regulator is to maintain the error $\varepsilon = e - r$ (or, if $e_0$ is taken as zero, $\varepsilon = -r$) as close to zero as possible. The disturbances are often considered as the main inputs (Fig. 1-29).

2. Servo systems in the strict sense have inputs which vary constantly with time and are designed to make input and output equal, whatever the time variations of the input; automatic tracking radars, reproducing lathes, and control-surface servos are all “follow-up” systems. The input to a servo system is the command, or control input. Either the response or the error $\varepsilon = e - r$ can be considered as output (Fig. 1-30).

The difference between a regulator and a servo system is not an absolute one, for two reasons:

a. Depending on the variables of interest, a particular system can be considered either as a regulator or as a servo system. For example, a servo system can be considered as a regulator the aim of which is to maintain zero error, despite the time-varying control input which, in this case, is considered to be the disturbance.

\(^1\) Automatic volume control.
b. In fact, almost all servo systems pose problems from both the regulator and servomechanism points of view. This concept will now be developed, supplementary to the examples already cited. In the case of the temperature regulator (Sec. 1.2.1) it is desirable that, for an abrupt change in input \( T_0 \), the output \( T \) follow rapidly, which condition is the basic requirement of a servo system. This constitutes the servo-system problem. Conversely, for a tracking radar, the parabolic antenna must not only obey the input orders but must do so regardless of its orientation with respect to the vertical, wind effects, etc. This latter is a regulation problem.

In conclusion, rather than specify systems as regulators or servo systems, it is preferable to refer to them as servo systems which can be considered, depending on their applications, from either the regulator or the servo-system point of view. Moreover, in the majority of practical cases, both aspects must be taken into account (Fig. 1-31).

![Fig. 1-31.](image)

![Fig. 1-32.](image)

For the case in which the disturbances can be introduced at specific points in the loop, the block diagrams can be represented in either of the two forms shown in Fig. 1-32a and b. The diagram of Fig. 1-32a expresses the servomechanism viewpoint, since the command, or control input, is considered as the main input. Figure 1-32b is more typical of the regulator approach (\( e = 0 \), hence \( e = -r \)).

1.3. EXAMPLES OF SERVO SYSTEMS. PHILOSOPHICAL CONSIDERATIONS

1.3.1. Reproducing Shaper. In order to reproduce a given shape to a different scale or in harder material, a reproducing shaper (Fig. 1-33) can be used. The master piece \( A \) and the workpiece \( X \) are rigidly held in a frame \( T \) which is allowed to move in the horizontal plane. A small sensing device \( E \) is so applied to the master
piece that it follows its contour when $T$ moves. The displacement of the tool $C$ is governed by a motor $G$ through a mechanical transmission. A differencing system (represented by a pulley and a potentiometer) controls the motor to zero the error between the vertical displacements $x$ and $y$ of $E$ and $C$, respectively. This is a servomechanism whose block diagram is shown in Fig. 1-34. This block diagram represents a servo system, the purpose of which is to cause the cutting tool $C$ to follow the displacements of the feeler $E(r = e)$, whatever the shape of the master $A$.

![Fig. 1-33. (After H. Lauer, R. Lesnick, and L. E. Matson, “Servomechanism Fundamentals,” McGraw-Hill, New York, 1947, by kind permission of the authors.)](image)

**Fig. 1.34.** Servo block diagram for automatic milling machine.

![Fig. 1-35. Regulator block diagram for automatic milling machine.](image)

**Fig. 1-35.** Regulator block diagram for automatic milling machine.

The regulator aspect becomes evident if one considers that the equality $r = e$ has to be maintained in spite of torque variations on the tool. Figure 1-34, the block diagram of the system, shows the disturbing torque added to the driving torque developed by the motor. Thus the two aspects must be kept in mind simultaneously, since a sudden change in the input $E$ results in a sudden variation of metal thickness to be ground and therefore also in a variation of load torque on the motor. However, the case would reduce to that of a regulator for a flat master and a rough workpiece (Fig. 1-35).

1.3.2. Autopilot. Consider a longitudinal autopilot designed to keep the pitch angle $\theta$ (angle of the aircraft to the horizontal) equal to a constant value $\theta_0$ ($\theta_0 = 0$ in horizontal flight).

The error between the desired pitch angle $\theta_0$ and the actual pitch angle $\theta$ is detected
by means of an error-sensing device, e.g., a pitch gyroscope fitted with a potentiometer whose zero corresponds to \( \theta_0 \). In such a case, the current \( i \) is proportional to \( \theta_0 - \theta \). The current \( i \) is then modified and amplified before being introduced to the control motor. The modifications in question cannot be gone into in detail in this first chapter (they usually consist of an addition to \( i \) of currents proportional to its successive derivatives and integrals, this operation being effected by a compensating network). The current is often amplified by an electronic amplifier whose output is the control current of the servo motor.

With such a system of control, the servomotor produces a torque \( C \) from which results, via the control surface and the aircraft, the pitch angle \( \theta \). This is represented by the block diagram in Fig. 1-36, where the disturbances, essentially the forces and torques induced by atmospheric turbulence, are represented as secondary inputs. In the particular case where the effect of gusts produces primarily pitch torques \( C_2 \) acting in addition to the controlling torque \( C_1 \), the block diagram of Fig. 1-37 applies.

This feedback control system can be considered as a regulator as well as a servo system. From the regulator aspect, it is required to maintain the pitch angle of the aircraft equal to the constant command value (horizontal flight, for example), despite the effects of atmospheric disturbances, variations in the centering of the aircraft, and so on.

From the servo aspect proper, it is required that the pitch of the aircraft should conform as closely as possible to the command input, however rapid and unforeseen these commands may be (sudden changes in the angle of climb, quick maneuvers, etc.). For this particular case, the regulator viewpoint is more relevant than the servo viewpoint. Incidentally, such an autopilot is often called a stabilizer because of its function.

\textit{Note.} In some stabilizers, the gyroscope controls the servomotor not by means of a potentiometer adjusting the current in proportion to the error, but rather by means of a two-point contact whose output is a current of constant magnitude and of polarity which changes sign with a change in sign of the error \( \theta_0 - \theta \) (Fig. 1-38). In this case, we have what is known as an on-off servo system, as opposed to a servo system with continuous or linear sensing.
Fig. 1-39. Block diagrams for autopilot and human pilot.
1.3.3. Manual Piloting. Consider an aircraft whose pilot wishes to maintain a constant pitch $\theta_0$. To do so, the pilot operates the controls in such a way as to counteract the differences he observes between $\theta_0$ and the actual pitch $\theta$. This constitutes a servo system whose block diagram is very similar to that of the preceding example (Fig. 1-39), with the difference that the pilot replaces the sensing device, the compensating network, the amplifier, and the servomotor. More precisely, the pilot's sense organs are the analog of the error-sensing device, and his muscles correspond to the amplifier and the servomotor. The compensating network is to be found in the pilot's brain, in the form of the modification of his reactions resulting from his training as a pilot.\footnote{This point will become clearer later in the book (Chap. 18).}

Here again there are two aspects: the regulator aspect, i.e., the maintaining of a constant pitch despite atmospheric disturbances, and the servo aspect proper, the control of the pitch of the aircraft. In practice, the two are, as a rule, inseparable.

1.3.4. Control and Checking. In the army, the civil service, and like groups, it is well known that, before an order is carried out, it must be followed up closely. This is because of "secondary inputs," in effect, mainly accounted for by the universal tendency to avoid responsibility whenever possible. The effect is marked in proportion to the length of the command chain (Fig. 1-40), whence stems the necessity to scrutinize the carrying out of the order by comparing it, as it is passed along the chain of command, with the original order and adjusting it as and when deemed necessary (Fig. 1-41).

![Fig. 1-40](image1)

![Fig. 1-41](image2)

The block diagram has the same form as the preceding ones; therefore we are concerned with a feedback system. Moreover, it obviously involves power amplification (in the case of an army corps, the amplification factor may amount to several thousands), and thus we have a servo system.

1.3.5. The Relative Character of the Servo-system Concept. Nothing absolute is involved in the concept of a servo system; that is to say, when a control system is designated as a servo system, it simply means that it is convenient to consider it as such in relation to a given problem. It may well be, however, that for other problems it is preferable to overlook the fact that it is a feedback system and consider it as an ordinary control system.

Consider, as an example, a remotely controlled gun turret. From the operator's viewpoint, it is of little consequence whether or not the system incorporates a feedback with an error-sensing device. To the operator, only the input, the output, and the manner in which they are linked (precision, accuracy, sensitivity to external disturbances, etc.) are of interest. One is thus justified in representing the apparatus by a diagram of the type shown in Fig. 1-42, which does not show a closed loop.
This does not alter the fact that the system in question is seen, on examination of its internal workings, to be a typical servo system. The servo-system characteristics only become evident, however, if its internal structure is investigated. The system, therefore, can be regarded as a servo system or not, depending on the viewer and his perspective.

The same is true in the case of the example in Sec. 1.3.2 for the pitch-control autopilot. Consider, first, the viewpoint of an engineer who is to place a bombsight into an aircraft equipped with this type of automatic pilot. His prime concern is with the degree of aircraft stability; he is little concerned with the manner in which

![Diagram](image_url)

**Fig. 1-42.** Gun-turret servo system as ordinary control system.

![Diagram](image_url)

**Fig. 1-43.** Autopilot as ordinary control system.

![Diagram](image_url)

**Fig. 1-44.** Autopilot as servo system.

![Diagram](image_url)

**Fig. 1-45.** Block diagram of autopilot, showing superposed loops.

this stability is achieved. To him the aircraft and its automatic pilot merely constitute a system whose output $\theta$ is linked to the input $\theta_0$ in a manner which determines stability, insensitiveness to external disturbances, and the like. The block diagram will thus be of the type shown in Fig. 1-43.

Consider, now, the viewpoint of the engineer who wishes to study the principles of autopilot operation, without, however, going into details of how the equipment is constructed. For him, the block diagram will be as shown in Fig. 1-44, and it will include a closed loop.

As for the autopilot specialist himself, he will consider it fundamental that the control-surface deflection $\delta$ be slaved to the output $j$ of the compensating network, and his block diagram will include two superposed loops (Fig. 1-45).

Many more such examples could be given. In general, the diagram of a fairly complex system includes an increasing number of closed loops in proportion to how much the person analyzing the system desires to

\[1\] In the type of piloting known as position piloting (see Sec. 15.3).
THE CONCEPT OF FEEDBACK CONTROL

delve into the details of its internal functioning and is well versed in servo-system concepts. This explains the tendency of servo specialists to "see servo systems everywhere."

The important point that is to be remembered is that servo systems and non-servo systems must not be regarded as two separate and distinct entities. Servo systems are control systems with power amplification that are designed to achieve as accurately as possible the equality of two quantities: a low-power input and a higher-power output. The fact that servo systems incorporate feedback is a distinguishing feature—and, admittedly, a very important feature—but it is one which is of interest only to those concerned with the internal operations of the system.

1.3.6. General Characteristics of Servo Systems. The concept of servo systems described above is very general, and the examples which may be cited to illustrate it are innumerable and extremely varied. This is understandable in view of the varied nature of the quantities that can be regulated—mechanical, electrical, thermal, and so on—as well as the diverse techniques which can be employed, the intermediate forms of energy and different orders of magnitude involved, and the degrees of system complexity.

It may, however, be said, in general terms, that servo systems possess three essential properties, each of which can be regarded as a slightly different way of expressing the fact that the systems are feedback systems. The properties are:

1. The capacity to adapt actively to unforeseen circumstances
2. The capacity to function independently of certain external conditions
3. The ability to obtain a high degree of accuracy in the neighborhood of zero

We shall now explain these three properties, which are most important because they lead to a proper understanding of the problems posed by servo systems.

1. Insofar as adaptation to unforeseen circumstances is concerned, it is sufficient to bear in mind that feedback control is present when a control system takes cognizance of the result of its operation and uses this knowledge to modify its operation.

In this sense, almost all systems in which a human operator is involved are feedback systems. For example, the control of traffic by means of traffic lights which change color at predetermined intervals is an open-loop control system. But if the traffic is controlled by a policeman, who takes account of variations in the flow of vehicles, then control is by means of a closed loop. The difference between the two rests essentially on the fact that in the latter case human intelligence is used, thus making it possible to judge the effects of unforeseen circumstances and adapt the behavior of the system accordingly.

Feedback systems which are entirely electromechanical have, in a sense, this same characteristic of "intelligence." This is exemplified by comparing a shell as fired by

---

1 Special care must be taken not to refer, as is sometimes done, to the special dynamics of servo systems, as opposed to conventional mechanics based on the equation $F = ma$. 

an ordinary antiaircraft gun (open-loop system) with a ground-to-air guided missile (closed-loop system). The trajectory of the former is entirely determined by the laws of mechanics and the initial conditions. If a storm occurs, or if the target aircraft takes evasive action, the shell may miss. Conversely, because of its guidance system, the guided missile is able to correct its course while in the air, so as to hit its target despite evasive action by the latter and despite external disturbances.¹

It is this property of adaptation that is referred to when servo systems are called “systems which think,” or when they are considered as systems capable of replacing a human operator for certain tasks. The property is a direct consequence of the fact that servo systems are controlled by the deviation between the input and the output, and it is simply a statement of the fact that they are feedback systems.

2. With regard to the capacity to function independently of certain external conditions (regulation), it must be remembered that the feedback loop is introduced precisely in order to make the system independent of external disturbances so that it may reliably obey the control input. This property is implied by saying that a servo system is, in the general sense of the term, a filter which eliminates the secondary inputs, while faithfully obeying the control input. This is a very fruitful concept, and should always be borne in mind by engineers who are studying or designing servo systems. Again, the property is sometimes expressed by saying that a servo system possesses homeostasis, a physiological term designating the capacity of living beings to maintain a constant internal state despite changes in external conditions, e.g., to regulate body temperature.

3. Finally, concerning accuracy, feedback systems may be considered as systems which, in order to maintain equality between input and output, take as their principal variable the difference between these two quantities, so as to minimize this difference.

The usefulness of this technique becomes apparent after considering an open-loop control system consisting of considerable amplification, e.g., in several successive stages. An error drift, or bias, in one of the initial stages has an effect, in amplified form, on the following stages. The introduction of a feedback loop feeds back the component of output due to the undesirable input within the system. Since the system operates on the basis of maintaining the difference between input and output as equal to zero, a correction is made, hence reducing residual errors to second-order effects. The error is, therefore, the main variable, which means that a nulling process is used to effect control.

Numerous physiological regulating mechanisms are opposing effects; physiologists call them self-regulating systems. An example is the regulation of the content of certain substances such as sugar and calcium in the bloodstream by means of two

¹ The difference is sometimes explained as follows: the behavior of the shell is determined by the laws of mechanics and the initial conditions, whereas the behavior of the guided missile is not determined at the outset. The servo system may, therefore, be considered as a mechanism that is “freed” from the determinism of physical laws, and so possesses its own autonomy. The authors consider it superfluous to stress the puerility of reasoning of this type.
opposing hormones. Another example is that of voluntary muscular movement; Beausoliel has shown that, in voluntary movement, the agonistic and antagonistic muscles possess a certain tonus from which the movement results. (This is the principle of reciprocal innervation.)

13.7. The Philosophy of Servo Systems. Much has been written on this subject during the last few years. The nineteenth century, during which engineering was primarily concerned with large quantities such as kilowatts and thousands of tons, stands in marked contrast to the twentieth century, in which enormous forces are controlled by small, delicate "mechanisms which think." One speaks of the "age of robotics," and a connection has been sought between the technique of servomechanisms and the behavior of living beings. The preceding remarks provide a sufficiently clear indication of the scope and the limits of such considerations.

1. Servo systems, as indicated, are systems that think, in the sense that they can replace, at least partially, a human operator. This fact has important and obvious repercussions in the technical, industrial, and economic spheres. On the other hand, from a philosophical point of view, it is not possible, without the abuse of language, to speak of "freedom" or "artificial thinking." The reason for this is that a servo system, like any electromechanical system, even if it is extremely complex and involves many internal loops, has a predetermined function, and can only follow the course specified by its designer.

2. Secondly, servo systems possess the property of homeostasis, in which respect they resemble living beings. Some of them are even capable of modifying their internal functioning in order to achieve optimum adaptation to circumstances. The most celebrated example of this is Ashby's homeostat. Mention may also be made in this respect of autopilots which modify their functioning according to certain deviating factors. These additional properties introduce complexities and additional loops in their block-diagram representations.

However, the analogy must not be carried too far. Homeostats and electronic "animals" only possess reflexes of a certain nature determined by their designers' application of physical laws. Animal and human behavior have other potentialities!

3. Servo systems have also things in common with self-regulating systems encountered in physiology, biology, and the human sciences in general. This fact has led to a technique of electronic models, especially in relation to the mechanical representation of economic and social phenomena. This technique makes it possible to study quantitatively the consequences of assumptions involving calculations too complex for solution by conventional methods.

It must not be forgotten, however, that the simplifications adopted in order to express human factors, which are always numerous and complex, in the form of equations often result in gross distortions. Furthermore, the stumbling block in writing the equations of economic and social phenomena almost always consists in departing from the domain of formal analogy only and giving the equations numerical coefficients in the hope of leading to results that conform to reality. In point of fact, electronic models have no more philosophical significance than the mechanical models of the last century. In particular, when the behavior of the model apparently reproduces real, observed physiological or social phenomena, there is no justification for deducing a functional identity of the mechanisms involved.

13.8. The Use of Servo Systems. Servo systems can thus perform tasks which an ordinary control system cannot. It must not be forgotten, however, that this is at the expense of greater complexity.

On the contrary, in the case of a swift, jerky movement, the antagonistic muscles are relaxed. This is an example of open-loop control.
Consequently, just as it is undesirable to employ a man to perform completely routine work, it is a priori inadvisable to use feedback control systems under conditions not involving unforeseen factors, such as in controlling valves of a piston engine which can be controlled simply by a camshaft. In principle, feedback control systems have their uses only under working conditions where random elements are present.

1.4. INTRODUCTION TO SERVO PROBLEMS. OUTLINE OF FOLLOWING CHAPTERS

1.4.1. The Elements of a Servo System. Any servo system can be schematically represented by a closed-loop block diagram (Fig. 1-46) which essentially consists of a forward path with power amplification and a feedback path. The forward path itself consists of the following basic elements (Fig. 1-47):

1. At the right-hand extremity of the forward path is the system to be controlled (the cutting head in Sec. 1.3.1, the aircraft in Secs. 1.3.2 and 1.3.3) preceded by a motor (Secs. 1.3.1 and 1.3.2) or its equivalent (muscles in Sec. 1.3.3). This motor is called a servomotor and may be electrical, hydraulic, or pneumatic. Very frequently it operates at high power (some hundreds of watts in Sec. 1.3.1, 10 to 30 watts in Sec. 1.3.2) and even when it does not, it always operates at power that is higher than that of the extreme left of the forward path.

2. At the junction of the feedback loop and the forward path is the error-sensing device. This is an element with two inputs, one of which is the servo-system control input and the other the servo-system output obtained through feedback. Its output is proportional to the error, i.e., to the difference between the input and the output

\[
\varepsilon(t) = e(t) - r(t)
\]

The error-sensing device actually consists of a detector or sensing device. It may be a differencer or its equivalent as in Sec. 1.3.1, a gyroscopic mechanism as in Sec. 1.3.2, or sense organs as in Secs. 1.3.3 and 1.3.4. It is followed by a transducer which supplies the error signal from the quantity picked up. Depending on the nature of the error signal, the transducer is electrical (either d-c—a potentiometer, for example—or a-c, such as a synchro), mechanical, or of some other form (the centripetal extension of the neurone in Sec. 1.3.3). The sensor always operates at low power. This power is generally very small in relation to the power of the output stage and is usually only a small fraction of a watt.

As a general rule it is the sensor that incorporates the basic idea or
Fig. 1-47. Components of a servo system.
stratagem of a servo system. Important inventions in the field are nearly always centered around the sensing device.

3. Between the sensing device and the servomotor is the power-amplification stage. It may be electrical, hydraulic, mechanical, or of some other form. An example of the latter is the control of a muscle by a motor nerve (Sec. 1.3.3). The power-amplification stage itself is usually subdivided into two more stages. Thus, for example, in electronic amplification there is usually a preamplifier and an amplifier proper. If amplification is by relays, then it is produced first by using small sensitive relays and then large power relays.

4. The fourth basic element is the compensating network, which is usually inserted between the sensor and the preamplifier. It therefore works at low power. Its function is to process the error signal with the object of improving the performance, both static (for accuracy) and dynamic (for damping), of the servo system. It is also frequently used for stabilizing the system. The compensating network is generally electrical, consisting of resistances and capacitances whose values have previously been calculated and subsequently adjusted in the laboratory; it may, however, also be mechanical.

The power stage of a servo system is in general the term used to designate the entire second stage of amplification, including the servomotor. In many industrial servo systems, the power stage is the most costly part of the entire apparatus.

All the points outlined above will be dealt with in detail later on. However, for convenience, the essential points to be remembered for the present are contained in condensed form in Fig. 1-47.

1.4.2. The Study of Servo Systems: The Theory of Feedback Control. The study of any physical system involves both theoretical and experimental considerations. Initially, theoretical considerations are carried out and the laws of Newtonian mechanics (\( F = ma \)) and of electricity (Maxwell's laws and their particular cases: Ohm's and Kirchhoff's laws) are applied to the system in question. Following this initial theoretical investigation is the experimental, or testing, stage in which the system is studied under varied conditions, either in the laboratory or in the field. This procedure equally applies to servo systems. They must first be studied theoretically, in the light of physical laws, and then experimentally.

The application of the laws of mechanics and electricity to servo systems leads, however, to the elaboration of a certain number of new concepts which are applicable to all servo systems and which make possible a simplified description of their operation. These new concepts are contained in what is called feedback-control theory. Interest in this theory lies in its unification, or synthesis, since it enables the development of concepts and methods common to all problems encountered in servo systems, however disparate they may first appear, and makes clear their underlying unity.

Feedback-control theory is also influential in the experimental study of servo systems. As will be seen, it is generally on the basis of servo
theory that systematic tests of servo systems and their components are
carried out. Finally, servo theory exerts a fundamental role in the
synthesis, as well as in the study, of servo systems. This aspect will be
dealt with below.

Servo systems present to the engineer\(^1\) extremely varied problems. One
of the most important of these problems, and the one that forms the
basis of this book, is that of synthesis.

In practice, one is always faced with problems of partial synthesis,
which may be generalized as follows: Given certain elements of a servo
system, how is one to design the other elements so that the servo system
as a whole will satisfy given standards of performance? For example,
given an aircraft, it is required to design an autopilot from the sensing
device to the servomotor and to ensure that the piloted aircraft will have
a degree of stabilization that conforms to given technical standards.
Or again, given a turbine, it is required to design a regulator for it that
will ensure specific working requirements. In any servo problem, as the
number of elements that must be determined increases, the problem
more closely approaches that of total synthesis.

Thus, in designing an autopilot for a guided missile, the missile itself cannot be
considered as an element that is known. The missile may be just as likely designed
around its stabilization and guidance as vice versa. Consequently, nearly all the
elements of the servo system must be determined. This is, therefore, a problem that
very closely approaches that of total synthesis; in practice it is dealt with by a method
of successive approximations, the missile being modified to accommodate the servo
system contemplated for it. Conversely, if, for example, it is merely desired to sta-
bilize an aircraft by means of an autopilot which has already been perfected for a
different aircraft, the basic components of the autopilot (sensing device, amplifier, and
servomotor) are already given, and only the compensating network and a few minor
adjustments must be determined. This is, then, merely a problem of adaptation.

In any case, solutions to problems of synthesis are complex. They
are attained by successive stages of theoretical calculations and practical
testing. At this point what may be considered the fundamental axiom
of servo-system design will be given, an axiom of an importance that
cannot be too strongly stressed: a servo system constitutes an entity, and
each stage in its design must be considered with respect to the performance of
the entire system. In other words, servo theory provides basic guiding
principles not only for the design of each separate stage, but also for the
over-all design of the servo system.\(^2\)

The last part of this book deals with the consequences of this axiom.
Failure to recognize it is, in the final analysis, frequently responsible for
a great many technical failures in the servo field. For the time being,

\(^1\) Much more varied and wider in scope, obviously, are the problems which servo
systems pose for more or less competent philosopher-cyberneticians and more or less
informed journalists. In this connection, the authors have nothing to add to the
remarks made in Sec. 1.3.7.

\(^2\) The only exceptions are very simple servo systems, or those derived from similar
existing systems.
however, it will only be stressed that, even during the preliminary planning stage, it is essential to calculate briefly, on the basis of servo theory, the over-all performance of the system. One of the difficulties that still remain at this stage is to state with accuracy what is expected of the system, and to express it in the language of servo systems (the choice of a criterion, see Chap. 17). In view of what has already been said about the random nature of feedback-control-system inputs, it is seen that the problem is one that is essentially of a statistical character.

1.4.4. Outline of Following Chapters. The preceding considerations were necessary in order to give an outline of the following chapters. The aim is to explain systematically the manner in which a servo system can be designed and its principal components defined starting from given technical data. It is now seen why the chapters that deal with the component parts must be preceded by a treatment of the general theory. This book has, therefore, been planned according to the following general pattern:

Servo theory.................... Parts 1, 2, and 3
Servo-system components....... Part 4
Servo-system design............ Part 5

1. In dealing with servo theory, the conventional practice followed has been to discuss the subject under two main headings: the theory of linear servo systems, which includes the important particular case of servo systems defined by linear equations with constant coefficients, and the theory of nonlinear servo systems. This division is justifiable from at least two viewpoints:

a. Mathematical. The theory of linear servo systems involves a coherent set of mathematical concepts centering around the theory of analytical functions, whereas the theory of nonlinear servo systems derives from much more disparate mathematical notions.

b. Historical. Linear servo theory, which was developed principally during World War II, has at the present time reached such a stage of maturity that it constitutes a classic and unified subject in itself. Conversely, nonlinear servo theories are the result of more recent research. They are less unified and often less well known because of military secrecy surrounding the problems relating to them.

Part 1 of this book, under the heading "Dynamics of Linear Systems," introduces basic concepts that are in constant use in the study of servo systems, especially so far as linear theory is concerned. It is important to understand thoroughly the concepts given in this part of the book before proceeding to the remainder of the book. Part 2 is concerned with the theory of linear servo systems and constitutes the main body of the book so far as current applications are concerned. Part 3 brings together a certain number of notions that are related to nonlinear servo systems.

2. Servo-system components are then systematically treated from the design standpoint. This constitutes Part 4.

3. The fifth and final part is devoted to servo-system design and contains two illustrative numerical examples.
PART ONE

DYNAMICS OF LINEAR SYSTEMS

CHAPTER 2

DERIVING THE EQUATIONS OF LINEAR SYSTEMS

Summary

1. General considerations.
2. A few notes on linearity.
3. Case of lumped-parameter electric systems.
4. Other systems. Notion of analogy.

2.1. GENERAL CONSIDERATIONS

2.1.1. The Importance of Deriving Equations. The mathematical study of a physical system, whether it is mechanical, electrical, acoustic, etc., involves the following steps:

1. Specify the assumptions to be made.
2. Derive the system equations by applying the physical laws that govern the system.
3. Study the behavior of the system from the equations. This step does not always require that a complete solution of the equations be derived, since information can often be obtained directly from the equations.

Contrary to what is sometimes thought, the derivation or writing of the system equations (step 2) is more important than the study of the equations themselves (step 3). Indeed, experience shows that the majority of errors arise from inexact equations rather than from faulty solutions. In addition, the equations, once written, can be solved and studied through the use of a computer, but no calculating machine can write the equations of the system under study.

2.1.2. The Technique. The derivation of the equations of a system involves application of the physical laws that govern it, i.e., the laws of mechanics and of electricity. In theory, all methods of applying the laws of mechanics are equivalent, the principle of virtual work and the Lagrange equations being only special formulations of the fundamental law of dynamics, \( F = ma \). However, for problems encountered in practice one method may be preferable to another from the viewpoint of the work involved in obtaining the desired equations.

In practice, the choice of the method best adapted to the problem is very important. For the great majority of problems are so complex that it is necessary for the engineer to carry out a systematic analysis. For example, it is obvious that the study of systems like those shown in Figs. 2-1 and 2-2 is impossible if they are approached haphazardly. Although the complexity of these two examples is equivalent to the problems given to students in undergraduate work, they are, neverthe-
less, not so complex as the majority of the systems an engineer must study in practice.

The general methods used for deriving the equations are beyond the scope of this book. The objective is simply to indicate for the special case of linear systems, which are especially important in the feedback-control field, a systematic method for the choice of variables and for the derivation of the equations. After defining the notion of a linear system, this technique will be presented for electrical systems; afterward, it will be extended to include certain types of mechanical systems. The similarity of procedure gives rise to the concept of electromechanical analogy.

2.2. A FEW NOTES ON LINEARITY

2.2.1. Definition. A physical system is said to be linear when the equations governing it are linear differential equations with constant coefficients. This definition is the one that is most widely adopted in engineering. It has the inconvenience of being more restrictive than the mathematical notion of a linear system which is founded on the principle of superposition.¹ This engineering definition excludes systems that

¹The principle of superposition may be stated as follows: (1) Additivity. If \( r_1(t) \) and \( r_2(t) \) are the responses of a system when it is separately subjected to two inputs \( e_1(t) \) and \( e_2(t) \), respectively, its response to the input \( e_1(t) + e_2(t) \) is \( r_1(t) + r_2(t) \). (2) Homogeneity, i.e., proportionality between cause and effect. If \( r(t) \) is response to the input \( e(t) \), the response to \( ke(t) \), where \( k \) is a constant, will be \( kr(t) \). It can be shown that the principle of superposition is characteristic of the systems governed by linear differential equations.
are governed by linear differential equations, the coefficients of which are functions of time, whereas these systems are linear in the mathematical sense. To avoid confusion, the latter are called linear systems with variable coefficients.

Actually there are no perfectly linear systems. Whenever a system is said to be linear, it means that assumptions have been made and that the result of these assumptions is that the system may be correctly represented by a linear scheme. Very often the assumptions made include:

- a. Neglecting the effect of certain discontinuous phenomena such as coulomb friction.
- b. Replacing continuous functions of variables by the first terms of their series expansions. This presumes that the following terms are negligible. A typical example of this is small-signal theory in electronics.
- c. Assuming that the response of a multiple input system is the sum of its response to each input (principle of superposition).

These assumptions will not be dealt with here. A more detailed discussion will be given in Chap. 11, when the reader has acquired a better knowledge of linear systems.

2.2.2. Comparison of Linear Systems with General Systems. An often forgotten assumption, which is fundamental for the notion of linearity, is that the operating frequencies are low. In the case of electric systems, it can easily be seen that the assumption of linearity presumes that the frequencies $N$ of the phenomena under consideration are not too high. To be more precise, one assumes that the associated wavelength $\lambda = c/N$, where $c$ is the velocity of light, is very large with respect to the dimensions of the system.

<table>
<thead>
<tr>
<th>Table 2-1. Maxwell equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>General case (waveguides)</td>
</tr>
<tr>
<td>Linear systems</td>
</tr>
<tr>
<td>(Kirchhoff’s laws)</td>
</tr>
<tr>
<td>$x \ll \lambda$</td>
</tr>
<tr>
<td>Free propagation</td>
</tr>
<tr>
<td>(Poynting’s vectors)</td>
</tr>
<tr>
<td>$x \gg \lambda$</td>
</tr>
</tbody>
</table>

Considered in this manner, it is seen that the class of linear systems stems from one of the two main simplifications of Maxwell’s equations, the second case, diametrically opposed, being that of propagation in a free field. In fact, if $x$ represents the dimensions of the system under consideration, the free-propagation simplification corresponds to $x \gg \lambda$ and the linear-systems simplification to $x \ll \lambda$. The general case is the one in which neither $x \gg \lambda$ nor $x \ll \lambda$. Maxwell’s equations must then be used in all their complexity (for example for waveguide problems; see Table 2-1).
The transition from the special case of free propagation to the general case occurs when \( x \) ceases to be very large with respect to \( \lambda \), as, for example, in the presence of obstacles to the propagation. The properties of the electromagnetic field in this case can sometimes be deduced from those corresponding to the free-propagation case by using interference methods. The transition from the particular case of linear systems to the general case occurs when \( \lambda = c/N \) is no longer large compared to \( x \), or when the period, \( T = 1/N \), of the phenomena under consideration is not large compared to the propagation time \( x/c \) of an electromagnetic wave in the system. This applies to high-frequency circuits which operate as if the propagation were not instantaneous. Thus, there is a space variable. To clarify this last point, consider the simple example of two rectilinear parallel conductors excited by the same current. These conductors will attract each other (Fig. 2-3). If the current \( i \) is sinusoidal, that is, \( i = i_0 \sin \omega t \), the conductors attract each other as long as \( \omega \) is not too high. However, as \( \omega \) increases, the time \( x/c \) required for the electromagnetic field created by the conductor 1 to reach conductor 2 becomes comparable to the period of the current. At the instant \( t \), a current \( i_0 \sin \omega t \) will flow through conductor 2. At this moment, conductor 2 will be subjected to the field created by conductor 1, \( x/c \) sec earlier, at which instant the current flowing through it was \( i_0 \sin \omega(t - x/c) \). It is obvious that, as soon as \( \omega x/c \) stops being small compared to \( \pi \), the resulting phase shift ceases to be negligible. For example, if \( \omega x/c = \pi \), the conductors repel instead of attract each other. Therefore, the created force is a function of the distance \( x \). This variation of force with distance is in addition to the normal \( 1/x \) variation. This is an apt description of the effect of the appearance of a space variable or of propagation that is not instantaneous.

The transition from linear circuits to very-high-frequency circuits is accomplished in the following two stages (see Table 2-2):

a. As long as the frequency is not too high, the system can be approximately represented by linear equations with constant coefficients by introducing supplementary terms, e.g., unwanted inductances of wound resistors or unwanted capacitances of tubes.

b. If the frequency is still higher, such a correction becomes insufficient and Maxwell's equations must be applied in one form or in another, as is the case for distributed-parameter circuits.

<table>
<thead>
<tr>
<th>Linear systems (Kirchhoff's laws)</th>
<th>Distributed-parameter circuits</th>
<th>Waveguides</th>
<th>Free propagation (Poynting's vectors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circuits with undesired inductances and capacitances</td>
<td>Propagation with obstacles</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analogous considerations are valid in mechanics. It is permissible to represent by linear equations with constant coefficients numerous mechanical systems composed of masses, springs, and dashpots. However, when the frequencies of the phenomena under consideration become sufficiently high, bodies cannot be considered as infinitely rigid, nor can the inertia of springs be neglected. This is analogous to the appearance of unwanted capacitances and inductances in the previous example. If the frequencies become even higher, it is not permissible to analyze the phenomena with lumped parameters. One must now resort to the mechanics of continuous structures, which are not linear and which are somewhat similar to electrical circuits with distributed parameters. The frequencies at which linear mechanics gives way to the mechanics of continuous structures frequently are of the order of a few cps.

These considerations will be stated again and summarized in Chap. 11. They have been discussed in this chapter because the implicit condition of linearity, that the frequencies should not be too high, is one of those most easily forgotten in practice. Therefore, in the study of linear systems, it is important to make certain that, as the frequency increases, the assumptions of linearity are not violated.

In order to show by means of introductory examples how differential equations for physical systems can be derived, the following section is devoted to the technique of writing the equations of simple lumped-constant electric networks, which are typical examples of linear systems.

### 2.3. ELECTRIC LUMPED-CONSTANT SYSTEMS

*Note. This section is an elementary introduction to the equations of simple electric circuits. It is written for the benefit of other than electrical engineers and should be ignored by readers who have already had some training in electric circuits.*

#### 2.3.1. Fundamental Assumption. All electric systems considered in this section are assumed to be small compared to the wavelength associated with the phenomena under study. If this is the case (Sec. 2.2.2), there is no space variable; the only independent variable is time. For example, the discussions that follow apply to a usual laboratory network, with dimensions of the order of a few inches or feet, if the frequencies involved do not exceed 100 cps, since the associated wavelength is 3,000 km. But they would not apply if the frequencies were of the order of 100 Mcps, that is, if the wavelength were 3 m.

#### 2.3.2. Electric-network Components. The components of electric networks without amplifiers are listed in Table 2-3 with the symbols commonly used to represent them. It is customary to differentiate between active and passive elements.

1. **Active Elements.** The active elements are voltage and current sources (Fig. 2-4a and b). A *voltage source* has a fixed potential difference $V(t)$ across its two terminals; that is, $V(t)$ is characteristic of the source

![Fig. 2-4.](image)
<table>
<thead>
<tr>
<th>Active elements</th>
<th>Meshes</th>
<th>Symbol</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signs indicate polarity when ( V(t) ) is positive</td>
<td>( V = I R )</td>
<td>Resistance ( R )</td>
<td>( I = \frac{V}{R} )</td>
</tr>
<tr>
<td>Voltage ( V(t) )</td>
<td>Dissipation of energy</td>
<td>Conductance ( G = \frac{1}{R} )</td>
<td></td>
</tr>
<tr>
<td>Passive elements</td>
<td>( V = L \frac{dI}{dt} )</td>
<td>Inductance ( L )</td>
<td>Inverse inductance ( \frac{1}{L} )</td>
</tr>
<tr>
<td>Storage of electromagnetic energy</td>
<td></td>
<td></td>
<td>( I = \frac{1}{L} \int V , dt )</td>
</tr>
<tr>
<td>( V = \frac{1}{C} \int I , dt )</td>
<td>Capacitance ( C )</td>
<td></td>
<td>( I = C \frac{dV}{dt} )</td>
</tr>
<tr>
<td>Inductance ( S = \frac{1}{C} )</td>
<td>Storage of electrostatic energy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kirchhoff's equations</td>
<td>( \sum V_k(t) = 0 ) along mesh</td>
<td>( \sum I_k(t) = 0 ) at node</td>
<td>Currents directed from a node are considered positive</td>
</tr>
</tbody>
</table>

*Adapted from M. Gardner and F. Barnes, "Transients in Linear Systems," pp. 36-37, Wiley, New York, 1942, by kind permission of the authors and publisher.

and does not depend on the rest of the circuit, but the current \( I(t) \) that flows through it is not fixed. A current source, on the contrary, forces a fixed current \( I(t) \) into the circuit to which it is connected, but the voltage \( V(t) \) across its terminals depends on the rest of the circuit.

In practice, an emf generator \( E \) can be regarded as a voltage source \( E \) when its internal resistance \( r \) is negligible with respect to the resistance \( R \) of the rest of the circuit (Fig. 2-5) or when its internal impedance is negligible with respect to the impedance of the rest of the circuit. The same emf generator can be regarded as a current source \( E/\tau \) when its internal resistance \( r \) (or impedance) is much greater than the resistance \( R \) (or impedance) of the external circuit (Fig. 2-5).

2. Passive Elements. The passive elements are inductances, capacitances, and resistances. An inductance \( L \) is a storage element for electromagnetic energy \( \frac{1}{2}LI^2 \), where \( I \) is the current that exists in the induct-
The presence of the current $I$ results in an induced voltage $-L \frac{dI}{dt}$; hence, the voltage drop $V$ across the inductance is given by

$$ V = L \frac{dI}{dt} \quad \text{or} \quad I = \frac{1}{L} \int_0^t V \, dt + I(0) $$

where $I(0)$ is the initial current that exists in the inductance at $t = 0$. In these equations, $L$ is evaluated in henrys if $V$, $I$, and $t$ are in volts, amperes, and seconds. For brevity the expression for $I$ is often written $(1/L) \int V \, dt$.

A capacitance $C$ is a storage element for electrostatic energy $\frac{1}{2}CV^2$. It is defined by the equations

$$ V = \frac{1}{C} \int_0^t I \, dt + V(0) $$

or

$$ I = C \frac{dV}{dt} $$

where $C$ is in farads. The expression for $V$ is often written for brevity as $(1/C) \int I \, dt$. The inverse capacitance $1/C$ is commonly termed elastance.

A resistance $R$ merely dissipates electric energy at the rate of $RI^2$ watts. It is defined by

$$ V = RI \quad \text{or} \quad I = \frac{V}{R} $$

where $R$ is in ohms. The inverse resistance $1/R$ is called conductance.

Mutual inductances $M$ between two circuits with currents $I_1$ and $I_2$ can also exist. They are storage elements for electromagnetic energy $\frac{1}{2}MI_1I_2$, and they are defined by

$$ V_2 = M \frac{dI_1}{dt} \quad V_1 = M \frac{dI_2}{dt} $$

Actual circuit components are represented more or less accurately by one or more of the above ideal elements. Potentiometers are represented by resistances guaranteed to within 0.1 per cent. Capacitors are best represented by a capacitance $C$ with a resistance $R$ in parallel, which accounts for the leakage and dielectric losses. At a given frequency $F = \omega/2\pi$ a quality factor $Q = RC\omega$ can be defined; it is often of the order of $10^4$ for precision capacitors.
Inductors are more difficult to represent by ideal elements. Roughly speaking, they correspond to an inductance \( L \), a resistance \( R \) in series, and an unwanted capacitance which becomes important at high frequencies. A quality factor \( Q = L\omega/R \) can also be defined, with typical values between 10 for very low frequencies and 200 for ferrite material at high frequencies. It should be noted that \( Q \) is nearly independent of frequency when the latter is high, say above 50 kcps, because of skin effect which causes \( R \) to increase at the same time as \( \omega \).

2.3.3. Equations for Electric Networks. The equations for electric networks with lumped parameters are most easily obtained by applying to them Kirchhoff’s laws, which can be applied as mesh equations or as mode equations.

The following definitions will be recalled. A branch is a two-terminal element, or a series connection of two or more elements which may be either passive active. A node is the terminal of an element, with the convention that a terminal common to two elements is considered as a unique node. A mesh is the closed path made up of several branches in series, or of one branch closing on itself.

The mesh equations state that the sum of the potential drops \( V_K \) around a mesh is zero when the mesh is traced out, say clockwise:

\[
\sum_{K} v_K(t) = 0
\]

The node equations state that the sum of the currents \( I_K \) leaving one node is zero:

\[
\sum_{K} I_K(t) = 0
\]

Note. When writing the mesh equations, it is in general preferable to write them in terms of mesh currents, the current flowing in a branch being the difference of the mesh currents in the two meshes separated by the branch considered. See the following example.

Example. The circuit shown in Fig. 2-6 represents a simplified relaxation oscillator (Sec. 5.4, Example 1) in which \( \rho \) is the resistance of a neon tube and the capacitor \( C \) discharges into \( R_1 \) and \( R_2 + \rho \) alternately. This circuit consists of a voltage source \( e(t) \), a capacitance \( C \), and three resistances \( R_1 \), \( R_2 \), and \( \rho \), that is, five elements, one active and four passive. There are three branches \( eR_1, C, \) and \( R_2 \rho \) and four nodes \( eR_1, R_1C\rho, \rho R_2, \) and \( CR_2 \). Three meshes can be considered \( eR_1C, C\rho R_2, \) and \( eR_1(\rho R_2) \), but only the first two are independent, the third being the “sum” of them (Sec. 2.3.5).

1. Writing the mesh equations (Fig. 2-7). Let \( I_1 \) and \( I_2 \) be the mesh currents in meshes \( eR_1C \) and \( C\rho R_2 \), respectively. \( I_1 \) is the current forced by \( e \) through \( R_1 \); \( I_2 \) is
the current that exists in \( R_2 \) and \( \rho \). Thus the current that exists in the branch \( C \) is \( I_1 - I_2 \) (positive downward). The mesh equations are

\[
\begin{align*}
\frac{e - \gamma}{C} &= R_1 I_1 + \frac{1}{C} \int_0^t (I_1 - I_2) \, dt \\
\frac{\gamma}{C} &= \frac{1}{C} \int_0^t (I_2 - I_1) \, dt + (\rho + R_2) I_2
\end{align*}
\]

where \( \gamma \) is the initial charge of capacitor \( C_1 \) with the sign indicated in Fig. 2-7.

2. Writing the node equations (Fig. 2-8). There are three independent node pairs. If the negative terminal of the voltage generator is taken as potential reference, the second nodes of these three pairs will be the positive terminal of the voltage generator, with a known potential \( e \); the terminal common to \( R_1 \), \( C \), and \( \rho \), with an unknown potential \( V_1 \); and the node common to \( \rho \) and \( R_2 \), with an unknown potential \( V_2 \). Thus the node equation is of no interest when applied to the node where the potential is known. Applying it to the nodes \( V_1 \) and \( V_2 \) yields, respectively,

\[
\frac{1}{R_1} (V_1 - e) + C \frac{dV_1}{dt} + \frac{1}{\rho} (V_1 - V_2) = 0
\]

\[
\frac{1}{\rho} (V_2 - V_1) + \frac{1}{R_1} V_2 = 0
\]

2.3.4. Initial Conditions. In order to write the mesh equations, the initial values (at \( t = 0 \)) of the voltages across the capacitances must be specified. Solving them requires that the initial currents in the inductances, which appear as terms in \( dI/dt \), be specified. If the node equations are used, it is necessary to specify the initial currents in the inductances for writing the equations, and the initial voltages across the capacitors for solving them. Thus in both cases it is necessary to specify the initial energy state, both electromagnetic and electrostatic, of the system.

In many cases one is interested in studying the behavior of the system after a disturbance has occurred at time \( t = 0 \), the condition of the system for \( t < 0 \) being known. The disturbance in question may be suddenly applying an additional voltage, closing or opening a switch, etc. In this case, specifying the initial conditions is facilitated by noting that the current that exists in an inductance and the voltage drop across a capacitance cannot, in general, experience discontinuities for \( t = 0 \). That is, in many practical problems

\[
I_L(0^+) = I_L(0^-) \quad V_C(0^+) = V_C(0^-)
\]
where $0^-$ is the condition before the event at $t = 0$ has happened and $0^+$ is the condition immediately after. Since the condition of the system for $t < 0$ is known, the quantities $I_L(0^-)$ and $V_c(0^-)$ are known. Hence the values $I_L(0^+)$ and $V_c(0^+)$ are obtained from the above equations.

**Example.** The network shown in Fig. 2-9, in which $e$ is a constant-emf generator, is assumed to have reached its equilibrium condition with switch $K$ closed. It is desired to study the effect of opening the switch $K$ at $t = 0$.

The mesh equations are easily obtained:

\[
\frac{\gamma_1}{C_1} = \frac{1}{C_1} \int_0^t I_1 \, dt + L \frac{dI_1}{dt} + R_1 I_1 + R_4(I_1 - I_4) \\
- \frac{\gamma_2}{C_2} = R_4(I_2 - I_1) + \frac{1}{C_2} \int_0^t I_2 \, dt
\]

In these equations $I_1$ and $I_2$ are the two mesh currents shown in Fig. 2-9 (the current that flows through $R_1$ is $I_1 - I_2$ and is taken as positive in the downward direction).

![Fig. 2-9.](image)

![Fig. 2-10.](image)

and $\gamma_1$ and $\gamma_2$ are the initial charges across $C_1$ and $C_2$, respectively, as shown in the figure.

The essential initial conditions are the values at $t = 0^+$ of the voltages $\gamma_1/C_1$ and $\gamma_2/C_2$ (necessary for writing the equations) and of the current $I_1$ (necessary for solving them). Before the switch was opened, the network was in an equilibrium condition. Under this condition the inductance acts as a short circuit, since an inductance has an induced emf only in response to variations in current, the term $L \frac{dI_1}{dt}$ vanishing when $I_1$ is constant. The capacitance, on the contrary, acts as an open circuit, since no current flows through a capacitance under d-c conditions. Thus the network for $t < 0$ is equivalent to the simplified network shown in Fig. 2-10, for which one can immediately write

\[
I_1(0^-) = \frac{e}{R_1 + R_2} \quad \frac{\gamma_1}{C_1} = e \quad \frac{\gamma_2}{C_2} = \frac{R_2 e}{R_1 + R_2}
\]

Thus the behavior of the system for $t = 0^+$ is completely defined from the following equations together with the pertinent initial conditions:
\[ e = L \frac{dI_1}{dt} + R_1 I_1 + R_2 (I_1 - I_2) + \frac{1}{C_1} \int_0^t \! I_1 \, dt \]

\[ R_2 (I_1 - I_2) + \frac{1}{C} \int_0^t \! I_2 \, dt + \frac{R_{ae}}{R_1 + R_2} = 0 \]

\[ I_1(0^+) = \frac{e}{R_1 + R_2} \]

If the node equations had been written, \( I_1(0^+) \) would be incorporated into the equations, and the other conditions

\[ \frac{\gamma_1}{C_1} = e \quad \frac{\gamma_2}{C_2} = \frac{R_2}{R_1 + R_2} e \]

would have to be added to them.

### 2.3.5. Choosing the Mesh or the Node Equations.

If the mesh equations are written, the unknown variables are the *mesh currents*. If the node equations are written, the unknowns are the *potential drops* across the node pairs. It is generally advisable to write the equations that involve the *smaller number of unknown variables*. The number of unknown variables can be read from the geometry of the network. If a network has \( e \) elements and \( n \) nodes, it can be shown that there are

\[ m = e + 1 - n \quad \text{independent meshes} \]

\[ p = n - 1 \quad \text{independent node pairs} \]

As a result, writing the mesh equations leads to a number of unknown quantities equal to \( m \) minus the number of current sources. Writing the node equations leads to a number of unknowns equal to \( p \) minus the number of current sources.

In the example of Sec. 2.3.3, \( e = 5, n = 4 \), hence there are two independent meshes and three independent node pairs. There are two unknowns, whether using the mesh or the node equations. If the voltage across \( C_1 \) is considered to be the output of interest, it is preferable to write the node equations, as these yield it directly.

### 2.3.6. Dual Networks

1. *Introductory Example.* Consider the two electric networks shown in Fig. 2-11. The first has one independent mesh with a mesh equation

\[ E = L_1 \frac{dI}{dt} + R_1 I + \frac{1}{C_1} \int I \, dt \]

1 If the network is made up of \( s \) separate parts connected by mutual inductances, the 1 in these equations should be replaced by \( s \).
The second has one independent node pair with a node equation

\[ I = C_2 \frac{dV}{dt} + \frac{1}{R_2} V + \frac{1}{L_2} \int V \, dt \]

If the numerical values of the elements are chosen in order that

\[ L_1 = C_2 \quad R_1 = \frac{1}{R_2} \quad C_1 = L_2 \]

it is seen that these two networks are described by the same equation, which is a mesh equation for the first and a node equation for the second. They are said to be reciprocal, or dual.

2. Definition. Two electric systems are said to be reciprocal, or dual, if they are described by the same differential equations, the mesh equations of one of them being identical with the node equations of the other. The following elements are reciprocal:

- Mesh currents and potential drops across node pairs
- Voltage sources and current sources
- Conductances and resistances
- Capacitances and inductances

3. Obtaining the Reciprocal of a Given Network. Not all networks have reciprocals. If one exists, it can be obtained by the following technique. On the diagram of the original network, place one point inside each mesh and one point outside the circuit. These points will be the nodes of the dual network. Then connect the points, obtained two by two, by as many branches as there are elements common to the two corresponding meshes. Finally, place on each branch the reciprocal of the element crossed.

Example. The network shown in Fig. 2-9 is made up of

- A voltage source \( e(t) \)
- An inductance \( L \) and a resistance \( R_1 \) connected in series
- A capacitance \( C_1 \)
- A resistance \( R_2 \) and a capacitance \( C_2 \) placed in parallel

![Fig. 2-12. Obtaining the reciprocal of a network.](image)

![Fig. 2-13. Reciprocal network of Fig. 2-9.](image)
The dual network obtained as explained by Fig. 2-12 is shown in Fig. 2-13. It consists of:

- A current source of $e(t)$ amp
- A capacitance $L$ and a resistance $1/R_1$ placed in parallel
- An inductance $C_1$
- A resistance $1/R_2$ and an inductance $C_2$ placed in series

The node equations for the dual network, after the switch $K'$ has been closed, are

$$i_1 = \frac{1}{C_1} \int_0^t V_1 \, dt + L \frac{dV_1}{dt} + R_1 V_1 + R_2 (V_1 - V_2)$$
$$-i_2 = R_2 (V_2 - V_1) + \frac{1}{C_2} \int_0^t V_2 \, dt$$

in which $V_1$ and $V_2$ are the node potentials with respect to zero, $i_1$ and $i_2$ are the initial currents in the inductances $C_1$ and $C_2$. The initial voltage across the capacitance $L$ should be added to these equations.

These node equations with switch $K'$ closed are the mesh equations of the original circuit with switch $K$ open. Conversely, the mesh equations of the circuit shown in Fig. 2-13 are the node equations of the original circuit.

2.4. OTHER SYSTEMS: ANALOGY

2.4.1. Introductory Example. Consider the mechanical system of Fig. 2-14, in which a mass $m$ is vertically displaced with respect to a fixed reference frame by a force $F(t)$. The mass is assumed to be restrained to the frame by a spring with a stiffness coefficient $k$. Furthermore, a viscous friction is assumed to be present, developing a force $fv$ proportional to the velocity $v$ of the mass with respect to the frame. Applying the law of dynamic force equilibrium yields

$$m \frac{dv}{dt} + fv + k \int v \, dt = F(t)$$

This equation has the same form as that of the electric circuits of Sec. 2.3.6. If, also,

$$m = L_1 = C_2 \quad f = R_1 = \frac{1}{R_2} \quad k = \frac{1}{C_1} = \frac{1}{L_2}$$

the mechanical system is governed by the same equations as the two electric systems. It is said to be their mechanical analog.

2.4.2. Definition. Two physical systems are said to be analogs of each other when they are governed by identical differential equations. According to their physical nature they are termed electromechanical, electro-hydraulic, or electroacoustical analogies. In what follows only the electromechanical analogy will be studied.\(^1\)

\(^1\) For the electrohydraulic analogy, see Probs. 13, 14, and 15. A somewhat different type of "hydraulic analogy" will be discussed in Sec. 9.3.7. For the electroacoustic analogy, see P. Morse and R. Bolt, "Sound Waves in Rooms," Rev. Modern Phys., 18(2): 69–150, 1944.
Remark. A few authors consider that two systems are analogs of each other when
the form of their differential equations (and not necessarily the numerical values of
their coefficients) are identical.

2.4.3. Electric Analogs of a Mechanical System. If a mechanical
system is linear, its electric analog can be drawn by using the following
technique.\(^1\) First, write the equations of the mechanical systems, using
forces and velocities as variables. Then consider in these equations (Table
2-4) either (a) forces as currents and velocities as voltages \((v \sim V\) or
\(F \sim I\) analogy) or (b) forces as voltages and velocities as currents
\((F \sim V\) or \(v \sim I\) analogy) and draw the electric circuit governed by these
equations. The two electric circuits thus obtained, (a) and (b), are said
to be the \(v \sim V\) and the \(F \sim V\) analogs of the mechanical system,
respectively.

<table>
<thead>
<tr>
<th>Table 2-4. Electric Analogs for Translational Mechanical Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Variables</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Elements</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. As an aid to memory, the \(V \sim v\) analogy preserves the form of the diagrams
and the references (fixed space, ground). In the same analogy, a spring \(k\) becomes an inductance \(L = 1/k\).

If the mechanical system is rotational, the law of dynamics will be
written as \(\tau = J \frac{d\omega}{dt}\), where \(\tau\) is a torque and \(\omega\) is an angular velocity,
and not as \(F = ma\). The two analogies are then (Table 2-5) (a) the
\(\omega \sim V\), \(\tau \sim I\) analogy and (b) the \(\omega \sim I\), \(\tau \sim V\) analogy.

Example. Electrical Analog of a Railroad Train. Consider a train made up of a
locomotive and two identical coaches. Let \(v(t)\) be the velocity of the locomotive.

\(^1\) Certain information concerning the conditions under which a mechanical or
electromechanical system has an electric analog can be found in M. Denis-Papin and
1951.
Table 2-5. Electric Analogs for Rotational Mechanical Systems

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mechanical</th>
<th>Electric analog ( V \sim \omega ) or ( I \sim \omega )</th>
<th>Electric analog ( V \sim \tau ) or ( I \sim \omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torque ( \tau )</td>
<td>( \tau = J \frac{d\omega}{dt} )</td>
<td>Current ( I = \tau ) ( V = \omega )</td>
<td>Voltage ( V = \tau ) ( \text{Current } I = \omega )</td>
</tr>
<tr>
<td>Angular velocity ( \omega )</td>
<td>( \text{Capacitance } C = J ) ( I = C \frac{dV}{dt} )</td>
<td>Inductance ( L = m ) ( V = L \frac{dI}{dt} )</td>
<td></td>
</tr>
<tr>
<td>Moment of inertia</td>
<td>( \tau = f\omega )</td>
<td>Resistance ( R = 1/f ) ( I = \frac{1}{R} V )</td>
<td>Resistance ( R = f ) ( V = RI )</td>
</tr>
<tr>
<td>Rotational mechanical resistance ( f )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torsional stiffness ( k )</td>
<td>( \tau = k f \omega dt )</td>
<td>Inductance ( L = 1/k ) ( I = \frac{1}{L} \int V , dt )</td>
<td>Capacitance ( C = 1/k ) ( V = \frac{1}{C} \int I , dt )</td>
</tr>
</tbody>
</table>

Note. As an aid to memory, the \( V \sim \omega \) analogy preserves the form of the diagrams and the references (fixed space, ground). In the same analogy a torsional spring \( k \) becomes an inductance \( L = 1/k \).

Fig. 2-15.

The couplings are identical and are assumed to be adequately represented by a spring \( k \) and a viscous friction \( f \), so that a force

\[ k(x_1 - x_2) + f \frac{d}{dt} (x_1 - x_2) \]

is developed between the two coaches when their abscissas are \( x_1 \) and \( x_2 \). Applying the law of dynamics to each coach gives

\[ m \frac{dv_1}{dt} = k \int (v - v_1) \, dt + f(v - v_1) + k \int (v_2 - v_1) \, dt + f(v_2 - v_1) \]

\[ m \frac{dv_2}{dt} = k \int (v_1 - v_2) \, dt + f(v_1 - v_2) \]

where \( v_1 \) and \( v_2 \) are the velocities of the coaches.

Considering velocities as voltages and forces as currents (\( v \sim V \) analogy) and using Table 2-4 yields

\[ C \frac{dV_1}{dt} = \frac{1}{L} \int (V - V_1) \, dt + \frac{1}{R} (V - V_1) + \frac{1}{L} \int (V_2 - V_1) \, dt + \frac{1}{R} (V_2 - V_1) \]

\[ C \frac{dV_2}{dt} = \frac{1}{L} \int (V_1 - V_2) \, dt + \frac{1}{R} (V_1 - V_2) \]
These are the node equations of the system shown in Fig. 2-16, where \( C = m, L = 1/k, \) and \( R = 1/f. \) This system has four meshes and three independent node pairs. It incorporates a voltage source. It is defined by the two equations given.

Similarly, using the \( v \sim I \) analogy, i.e., considering velocities as currents and forces as voltages, and using Table 2-4, leads to the equations

\[
L \frac{dI_1}{dt} = \frac{1}{C} \int (I_1 - I_1) \, dt + R(I_1 - I_1) + \frac{1}{C} \int (I_2 - I_1) \, dt + R(I_2 - I_1)
\]

\[
L \frac{dI_2}{dt} = \frac{1}{C} \int (I_1 - I_2) \, dt + R(I_1 - I_2)
\]

which are the mesh equations of the system shown in Fig. 2-17, provided \( L = m, C = 1/k, \) and \( R = f. \) This system has three meshes and four independent node pairs; it incorporates a current source. It is defined by the two equations given. Note that the systems of Figs. 2-16 and 2-17 are reciprocal.

2.4.4. Mechanical Networks. Those mechanical systems that have an electric analog can be represented as mechanical networks similar to electric networks. Mechanical circuit elements are shown in Table 2-6. Diagrams are drawn by connecting all the terminals that move together. (Masses or inertias are represented as elements with two terminals, one of which is connected to the reference point.) It is then possible to consider meshes and nodes and to extend to such mechanical systems the discussions of Secs. 2.3.3 and 2.3.4.

Examples. Figures 2-18 and 2-19 show in network form the mechanical systems of Figs. 2-14 and 2-15, respectively. The former incorporates a force source, the latter a velocity source.

2.4.5. Comparison of the Two Analogies. In general, the \( v \sim V \) analogy is more convenient than the \( v \sim I \) for the following two reasons:¹

¹ An interesting discussion of the comparison between the two analogies can be found in J. Loeb and G. Cahen, "Des réseaux électriques aux transmissions mécaniques," Association Technique Maritime et Aéronautique, 47: 107–145 (1948).
### Table 2-6. Mechanical-network Components

<table>
<thead>
<tr>
<th>Active elements</th>
<th>Translational</th>
<th>Rotational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Symbol</td>
<td>Name</td>
</tr>
<tr>
<td>Force source</td>
<td><img src="image" alt="F" /></td>
<td>Torque source</td>
</tr>
<tr>
<td>Velocity source</td>
<td><img src="image" alt="v" /></td>
<td>Angular velocity source</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Passive elements</th>
<th>Translational</th>
<th>Rotational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Symbol</td>
<td>Name</td>
</tr>
<tr>
<td>Mass $m$</td>
<td>$F = m \frac{dv}{dt}$</td>
<td>Moment of inertia $J$</td>
</tr>
<tr>
<td>Mechanical resistance $f$</td>
<td>$F = fv$</td>
<td>Mechanical resistance $f$</td>
</tr>
<tr>
<td>Stiffness</td>
<td>$F = k \int v dt$</td>
<td>Torsional stiffness</td>
</tr>
</tbody>
</table>

![Fig. 2-18. Mechanical network for system of Fig. 2-14.](image)

![Fig. 2-19. Mechanical network for railroad train of Fig. 2-15.](image)
1. In the $v \sim V$ analogy the mechanical and electrical diagrams have the same form.

2. In the $v \sim V$ analogy the ground (origin of potential) corresponds to a fixed reference in space.

**2.4.6. Interest of the Concept of Analogy.**

1. The *theoretical* interest of analogy is very *small*. The concept of analogy merely exposes the fact that, if similar assumptions are made (essentially, linear superposition), different physical systems are described by similar differential equations.

2. On the other hand, the concept of analogy is of *considerable* interest from the *practical* viewpoint. In fact:

   a. This concept enables results concerning a certain system to be immediately extended to its analogs; thus the theory of feedback amplifiers has been extended to servomechanisms, which are governed by analogous\(^1\) equations.

   b. When it is desired to study a mechanical (or hydraulic or acoustic) system, it is possible, instead of solving its differential equations or testing it in the field, to build its electrical analog, with which experiments can be effected in the laboratory. This is the basic idea underlying the technique of use of *analog computers* and *simulators* (Sec. 21.2).

\(^1\) Actually, the equations of feedback amplifiers and those of servomechanisms have the same form but are not analogous in the strict sense defined in Sec. 2.4.2. The major reason is that the orders of magnitude of the time scales in question are very different: a few cps for servomechanisms, much higher values for feedback amplifiers.
CHAPTER 3
TRANSIENTS

Summary
1. The concept of transients.
2. Their importance.
3. Response of a linear system to typical inputs.

3.1. THE CONCEPT OF TRANSIENTS

3.1.1. Introductory Example. Let us consider a moving-coil galvanometer. The current $I$ which flows through it is measured by the coil deflection $\alpha$ (Fig. 3-1). For a constant $I$ the galvanometer deflection is obtained by equating the electromagnetic torque $GI$ and the restraining torque $k\alpha$, $k$ being the torsional stiffness of the coil suspension,

$$k\alpha = GI \tag{3-1}$$

But, if the current through the galvanometer varies with time, the deflection angle $\alpha$ will not be given at each instant by $k\alpha(t) = GI(t)$. Its determination requires that the laws of dynamics be applied to the system. This necessitates adding to the left side of Eq. (3-1) an inertial term and a viscous-friction term:

$$J \frac{d^2\alpha}{dt^2} + f \frac{d\alpha}{dt} + k\alpha = GI(t) \tag{3-2}$$

where $J$ is the moment of inertia, $f$ is the damping constant, and $f \, d\alpha/dt$ is a damping torque proportional to the angular velocity. In other words, the study of the behavior of the galvanometer for a variable input requires that the complete differential equation be considered; the static equation (3-1) is not sufficient. For example, let us consider the two following cases.

Case 1. Suppose that a constant current $I_1$ flows through the galvanometer and that the latter has reached its equilibrium position. If at time $t = 0$ the galvanometer is suddenly subjected to a constant current $I_2 \neq I_1$, its behavior will be a solution of the differential equation (3-2) with $GI_2$ at the right-hand side. The initial conditions will be $\alpha(0) = (G/k)I_1$ because $\alpha(t)$ is necessarily continuous, and $(d\alpha/dt)(0) = 0$ because of the inertia. [In fact, if $(d\alpha/dt)(0)$ were not zero, $(d^2\alpha/dt^2)(0)$ would have to be infinite, which would not be compatible with the differential equation, since $I_2$ is finite.]
It is well known that the behavior of the galvanometer is, or is not, oscillatory depending on its degree of damping. Figure 3-2 shows two typical responses. Such responses are termed the transient of the galvanometer between its $I = I_1$ state and its $I = I_2$ state. Once the transient has vanished, the galvanometer is said to have reached its forced state, or steady state, that is, $\alpha(t) = (G/k)I_2$.

It is obviously desirable that a galvanometer have a satisfactory transient performance (Fig. 3-3); that is, that the transient (a) be such that the system responds rapidly and (b) be sufficiently damped.

**Case 2.** If the galvanometer, initially at rest ($\alpha = 0$), is suddenly subjected to an input $I = I_0 \sin (\omega t + \varphi)$ (sinusoidal or harmonic input), the response will be the sum of two terms:

a. A forced response which is the particular integral of

$$J \frac{d^2\alpha}{dt^2} + f \frac{d\alpha}{dt} + k\alpha = GI_0 \sin (\omega t + \varphi) \quad (3-3)$$

It can be determined by writing it a priori in the form

$$\alpha = \alpha_0 \sin (\omega t + \psi) \quad (3-4)$$

and finding $\alpha_0$ and $\psi$ by substituting the expression (3-4) into Eq. (3-3).

b. A term involving damped exponentials, which may be oscillatory or not ($f^2 \gtrless 4JK$). It can be obtained from the general solution of

$$J \frac{d^2\alpha}{dt^2} + f \frac{d\alpha}{dt} + k\alpha = 0$$

the integration constants being evaluated from the initial conditions. This term corresponds to the transient of the galvanometer between the two steady states

$$\alpha_1 = 0 \quad \alpha_2 = \alpha_0 \sin (\omega t + \psi)$$

The following is a physical explanation.

a. In time, the input $I = I_0 \sin (\omega t + \varphi)$ will force the galvanometer to follow its oscillations: this will result in the forced or steady-state oscillations, given by

$$\alpha = \alpha_0 \sin (\omega t + \psi)$$

which is the particular integral of the complete equation.
b. But, at the instant when the input is no longer zero and becomes \( I_0 \sin (\omega t + \varphi) \), the initial conditions \( \alpha(0) = 0 \) and \( (d\alpha/dt)(0) = 0 \) are not compatible with the forced response. As a result, it will take a certain time for the galvanometer to reach its steady state. The way in which it will do so is essentially dependent (1) on the initial conditions, more or less removed from the steady state and (2) on the inherent reaction of the galvanometer, dependent on the coefficients \( J, f, \) and \( k \).

![Transient performance diagrams](image)

**Fig. 3-3.** Satisfactory and unsatisfactory transient performances.

This is the *transient*. It is now physically clear why it is mathematically expressed by the general solution of the galvanometer equation, the constants resulting from the initial conditions.

**Note.** It is sometimes said that the transient corresponds to the state wherein occurs the dissipation of the energy the system had stored at \( t = 0 \).

3.1.2. Other Examples. Let us consider a *position-controlled gun turret* which is made to follow a prescribed angle in azimuth (Fig. 3-4). When a change in orientation is commanded, a certain time is required (because of inertia, etc.) before the turret can reach its new prescribed position. It is obviously desirable that that time be as short as possible.

Similarly, let us assume that the angular velocity of a hydraulic motor is to be...
controlled by hydraulic transfer. A sudden change in the command results in a transient. If the transient is oscillatory, dangerous increases in oil pressure may occur.\(^1\)

Numerous other examples could be cited, ranging from meteorology (lightning is a typical transient phenomenon) to mechanics (including all variable-speed motors), acoustics (sound of a piano string), and radio. It is now possible to specify the concepts common to all these examples by giving a few definitions.

3.1.3. Definitions. 1. A permanent input to a system is defined as an input whose expression as a function of time is \(a\) a constant, \(b\) a linear (or parabolic) function, or \(c\) a periodic function.

2. Most systems, when subjected to a permanent input for a sufficient length of time, finally yield a response that is \(a\) a constant, \(b\) a linear (or parabolic function, or \(c\) a periodic function of time.

The system will then be said to be in a permanent, or steady, state. The particular case of a steady state consisting of a constant input and a constant output is an equilibrium.

Examples. Examples are an electric motor when it rotates at constant speed (the output can be considered as a constant angular velocity or as a position which is a linear function of time), a relaxation oscillator under normal conditions, an airplane when it climbs at constant rate of climb, etc.

3. When a system in a permanent state (steady state I) is subjected at \(t = 0\) to another permanent input, it generally takes time for the system to achieve its new permanent condition (steady state II). During this time the system is said to be in a transient condition.

3.2. THE IMPORTANCE OF TRANSIENTS

The consideration of the transient response of a system is very important. Three reasons at least can be given for this:

1. Practical convenience. It is obviously objectionable for a galvanometer to oscillate for example, 3 min, before its indication can be read; this holds, however excellent its static performance.

2. Safety. The previously cited example of hydraulic transfer has indicated how transients can give rise to dangerous increases in oil pressure. These first two reasons alone are sufficient to explain why the transient performance of a system always deserves careful investigation.

3. But there is still a third reason, which is most fundamental and is also a basic factor in the understanding of what is to follow in this book. There exist a great number of systems which are never in a steady-state condition.

To be specific, let us consider a recorder used to measure and record the pitching rate of a stabilized airplane. The input as a function of time may well have the form shown in Fig. 3-5, which is taken from an

\(^1\) The transient response of hydraulic transfer will be studied in Prob. 14. If the angular velocity of the controlled motor is taken as the output, the system is the hydraulic analog of the galvanometer of Sec. 3.1.1. Oil compressibility is analogous to inertia; leakage is analogous to viscous damping. See Secs. 4.6.3 through 4.8.5 in the first French (Dunod, 1956) or the German (Oldenbourg, 1959) edition of this book.
actual flight test. Obviously, the most important matter is not that 
the recording instrument have an accuracy of $10^{-5}$ under static conditions; 
of much more importance is the question whether its transient perform-
ance will enable it to follow the minute variations of the input. For 
example, if the transient response of the recording system is slow, the 
recorder will not be able to follow the more rapid fluctuations of the 
input. These will be smoothed out, and the resulting output will look 
somewhat like Fig. 3-6. If the transient response is insufficiently 

damped, an unwanted oscillation will be superimposed, yielding some-
thing like Fig. 3-7.

Thus, any system that is subjected to unexpected inputs constantly 
works under transient conditions. This is the case for all servo systems, 
because their inputs are inherently random functions. In conclusion, 
servo systems practically always work under transient conditions. As a 
consequence, their dynamic performance (speed of response, sufficient 
damping) is at least as important as their static behavior.

3.3. TIME RESPONSE OF LINEAR SYSTEMS

3.3.1. Response of a System to Typical Inputs. The following cases 
are especially important in practice.

Case 1. Unit-step response. Steady state I is defined by the perma-
nent input $e(t) = 0$, steady state II by $e(t) = 1$. 

![Fig. 3-5. Airplane pitching rate as a function of time.](image)

![Fig. 3-6. Recording of Fig. 3-5 when instrument transient is too slow.](image)

![Fig. 3-7. Recording of Fig. 3-5 when instrument transient is insufficiently damped.](image)
This is expressed by saying that the system is subjected to the input
\[ e(t) = u(t) \]
where the function \( u(t) \), or unit-step function (Fig. 3-8), is defined as follows for \( t \neq 0 \):
\[
\begin{align*}
  u(t) &= 0 & t < 0 \\
  u(t) &= 1 & t > 0
\end{align*}
\]
It is not usually defined for \( t = 0 \), where a discontinuity takes place.

The corresponding response is called the unit-step response. Obtaining the unit-step response for a system is the transient condition most commonly used in practice.

Case 2. Ramp response. Steady state I is defined by \( e(t) = 0 \), steady state II by \( e(t) = t \).

The system is said to be subjected to a unit-slope ramp input
\[ e(t) = tu(t) \]
and its response is called the unit-ramp response (Fig. 3-9).

Case 3. Unit-impulse response. This is the response of the system to a unit impulse. An ideal impulse is well approximated by a function of time whose duration is very short compared to the time constants of the system it is applied to but whose magnitude is great enough for its effect to be noticeable. For example, the limit of the function shown in Fig. 3-10, as \( K \) approaches infinity, defines a unit impulse (Sec. 4.3). The unit-impulse response is a transient response between the steady states.

No. I: \( e(t) = 0 \) \( t < 0 \)
No. II: \( e(t) = 0 \) \( t > 0 \)

Case 4. Harmonic response (Fig. 3-11). If a system, initially at rest, is subjected to an input
\[ e(t) = Au(t) \sin (\omega t + \phi) \]
\[ \dagger \]
\[^{\dagger}\text{We recall the following elementary definitions: } A \text{ is the amplitude or magnitude of the harmonic input, } \omega \text{ is the angular frequency (in radians per second), the frequency is } \omega/2\pi \text{ (in cps), } \phi \text{ is the phase (in radians) with reference to } \sin \omega t.\]
the response is the sum of (a) a transient component, which generally vanishes in time and (b) a steady-state component which is periodic at $2\pi/\omega$. (It is sinusoidal if the system is linear.)

When a system is subjected to a harmonic test, one is generally interested in the forced state. This is an important difference as compared with Cases 1 to 3, where one is generally interested in the transient performance.

Obtaining harmonic responses of the system for various values of $\omega$ gives the frequency response of the system.

3.3.2. Figures of Merit for the Steady-state Performance of Control Systems. When the system $S$ under consideration is a control system, it is desired that the response should follow the input without appreciable error. This requirement is generally stated for a step and for a ramp input.

1. For a step input, the steady-state response of the system may or may not be identical with the input. If it is not, the system is said to involve a steady-state error for a step input, or more briefly, a position error (Fig. 3-12a). The term position error, initially used for position-control systems, has been extended by usage to control systems which are not positional. If the steady-state response is identical with the input (Fig. 3-12b), the system is said to involve zero position error.

2. For a ramp input, if the steady-state response is not identical with the input, the system is said to involve a steady-state error for a ramp, or constant-velocity, input, or more briefly, a velocity error (Fig. 3-13a). The term velocity error has been extended by usage to control systems which are not positional. If the steady-state response is identical with the input (Fig. 3-13b), the system is said to have zero velocity error.
3. Similarly, an acceleration error can be defined (Fig. 3-14) as the steady-state error of the system subjected to an acceleration input.

4. The forced response of a system to a harmonic input is, in general, a periodic function that is different from the input. If the system is linear, its response will be sinusoidal. One can then specify by how much the output amplitude and phase differ from the input amplitude and phase.

3.3.3. Figures of Merit for the Transient Performance of Control Systems. As already pointed out, it is generally required that the transient response of a control system be sufficiently fast and sufficiently damped. These requirements are usually specified as follows:

1. The condition that the transients be sufficiently damped is generally expressed by stipulating that the system, when subjected to a step or ramp input, will not go too far beyond its final steady-state position. This is usually expressed by the condition that the first overshoot of the step response, defined as shown in Fig. 3-15, should not exceed a certain value expressed as a percentage of the final value.

2. The condition that the transient response be fast enough is usually expressed by stipulating that the response time of the system be sufficiently small. The response time is usually defined as the time that the system takes to achieve 95 per cent of its final output for a unit-step input and thereafter remain between 95 and 105 per cent, i.e., the time after which the system approximates a steady-state condition within ±5 per cent\(^1\) (Fig. 3-16).

When the transient is oscillatory, its duration can also be characterized by the period of the transient oscillations, provided their damping (i.e.,

---

\(^1\) The response time is a way of expressing the duration of the transient. Mathematically speaking, the duration of the transient is infinite. For practical purposes, however, the transient can be considered to have disappeared when it has decayed to a sufficiently small percentage of its initial magnitude. The usual values are 5 or 2 per cent; they define a response time to within 5 and 2 per cent, respectively.
the rate at which the oscillations decrease in magnitude) is known. The frequency of the transient oscillations is termed the natural frequency of the system.

3.3.4. A Set of Usual Performance Specifications for Control Systems. The quantities just defined are very important when evaluating the performance of a control system. In particular, at the very outset of the design procedure, specifications concerning the performance of the system are stipulated. Usual specifications concern both the steady-state and the transient response for typical inputs.

1. Steady-state Specifications. These are usually:
   a. The system should have no position error.
   b. The velocity error should be sufficiently small, and zero if possible.
   c. The acceleration error should be sufficiently small.

When these three specifications happen to conflict, condition c is generally considered less important than conditions a and b.

d. Finally, it is also often stated that the magnification and phase shift should not exceed certain values in given frequency ranges.

2. Transient Specifications. Transient specifications are most commonly stated for a step input.
   a. The overshoot should not exceed a certain value.
   b. The response time should be sufficiently short, or in other words, the natural frequency should be sufficiently large.

It will be seen later in the book how such specifications can be applied to feedback control systems.

3.3.5. Studying the Time Response. If a system $S$ (Fig. 3-17) is represented by a linear differential equation with constant coefficients

$$A_n \frac{d^n r}{dt^n} + \cdots + A_1 \frac{dr}{dt} + A_0 r(t) = e(t)$$

the response for $t > 0$ for each of the typical cases of Sec. 3.3.1 is obtained by solving the differential equation with

$$e(t) = 1 \quad \text{for Case 1}$$
$$e(t) = t \quad \text{for Case 2}$$
$$e(t) = 0 \quad \text{for Case 3}$$
$$e(t) = A \sin(\omega t + \varphi) \quad \text{for Case 4}$$

and taking into account the corresponding initial conditions. This leads generally to

$$r(t) = C_1 + g_1(t) \quad \text{for Case 1}$$
$$r(t) = C_2 + g_2(t) \quad \text{for Case 2}$$
$$r(t) = g_3(t) \quad \text{for Case 3}$$
$$r(t) = C_4 \sin(\omega t + \psi_4) + g_4(t) \quad \text{for Case 4}$$

in other words to the sum of:

a. A function (the particular integral) which has the same form as the input. It is the steady state.
b. Exponentials (the general solution for the equation with zero in the right-hand side) whose exponents depend on the roots of the characteristic equation

\[ A_n \lambda^n + \cdots + A_1 \lambda + A_0 = 0 \]

It is the transient.

In practice, the application of this method\(^1\) has the following limitations: (1) As system complexity increases, computation becomes involved and tedious. (2) It is necessary to write the differential equation of the system. Thus, it cannot be applied when it is desired to study the transient performance of physical systems whose differential equation cannot be formulated.

It is the purpose of the following chapters to outline an approach that does not have these limitations. This approach enables the characterization of system performance without recourse to the complete response, since figures of merit can be obtained directly. Furthermore, it can be applied to systems for which data are obtained experimentally, and not from explicit differential equations. The basis for this approach is the Laplace transform, treated in the next chapter.

\(^1\) It should be pointed out that such a method is somewhat artificial. To solve a particular problem, a much more general solution must be developed and the coefficients painfully identified. The initial conditions of the problem, which physically are the primum movens of the transient, seem to be related to it solely by means of algebraic formal relations.
CHAPTER 4

THE LAPLACE TRANSFORM

Summary

1. Definition and fundamental properties.
2. Other important properties.
3. Impulse functions.
4. Other properties of the Laplace transform.
5. Basic notions concerning Fourier transforms.
6. Obtaining the time response of a linear system.
7. Appendix to Chap. 4.

The purpose of this chapter is to present the fundamentals of Laplace-transform theory. The Laplace transform is defined by means of a correspondence between functions of time and functions of the complex variable $s$. Its most important properties will be stated. It is stressed that knowledge of the present chapter, although it does not involve a great deal of advanced mathematics, is quite sufficient to understand, apply, and fully master the contents of this book. The reader interested in the mathematical proofs of the properties stated is referred to the appendix at the end of this chapter and to works listed in the Bibliography.

4.1. DEFINITION AND FUNDAMENTAL PROPERTIES

4.1.1. Definition. Let $f(t)$ be a function of time that is zero (or that is not defined) for $t < 0$. Under certain very general conditions (see Chapter 4 appendix) a function of the complex variable $s$ that is denoted by $F(s)$ or $\mathcal{L}f(t)$ can be defined; it is said to be the Laplace transform of $f(t)$. The Laplace transforms of the functions $f(t)$ most frequently encountered in the servo field are given in Table 4-1. The function of time $f(t)$ is said to be the inverse Laplace transform of $F(s)$ and is written $f(t) = \mathcal{L}^{-1}F(s)$.

4.1.2. Superposition: Additivity and Homogeneity. If two functions of time $f(t)$ and $g(t)$ have Laplace transforms, the function $f(t) + g(t)$ has a Laplace transform given by

$$\mathcal{L}[f(t) + g(t)] = \mathcal{L}f(t) + \mathcal{L}g(t)$$

1 Similarly, a complete knowledge of the properties of the exponential function is not necessary for using a table of logarithms. A thorough knowledge of Laplace-transform theory is not necessary to servo-system engineers. One reason is that the Laplace transforms encountered in the servo field are usually algebraic functions, sometimes multiplied by exponentials, which present no mathematical difficulty from the viewpoint of convergence or singularities.
Table 4-1. Laplace Transforms of Common Functions of Time

<table>
<thead>
<tr>
<th>$F(s) = \mathcal{L}f(t)$</th>
<th>$f(t)$ for $t &gt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{s}$</td>
<td>$1^*$</td>
</tr>
<tr>
<td>$\frac{1}{s^2}$</td>
<td>$t^\dagger$</td>
</tr>
<tr>
<td>$\frac{1}{s^3}$</td>
<td>$\frac{t^2}{(n-1)!}$</td>
</tr>
<tr>
<td>$\frac{1}{s+a}$</td>
<td>$e^{-at}$</td>
</tr>
<tr>
<td>$\frac{1}{(s+a)^2}$</td>
<td>$te^{-at}$</td>
</tr>
<tr>
<td>$\frac{s}{s^2+\omega^2}$</td>
<td>$\cos \omega t$</td>
</tr>
<tr>
<td>$\frac{\omega}{s^2+\omega^2}$</td>
<td>$\sin \omega t$</td>
</tr>
<tr>
<td>$\frac{s+a}{(s+a)^2+\omega^2}$</td>
<td>$e^{-at}\cos \omega t$</td>
</tr>
<tr>
<td>$\frac{\omega}{(s+a)^2+\omega^2}$</td>
<td>$e^{-at}\sin \omega t$</td>
</tr>
<tr>
<td>$\frac{\alpha s + \beta}{(s+a)^2+\omega^2}$</td>
<td>$Ae^{-at}\cos(\omega t + \varphi)$</td>
</tr>
</tbody>
</table>

* Note that this function (zero for $t < 0$, unity for $t > 0$) is the unit-step function defined in Sec. 3.3.1, Case 1.

† Note that this function is the ramp function $tu(t)$ defined in Sec. 3.3.1, Case 2.

Similarly, if a function of time $f(t)$ has a Laplace transform, the function $Kf(t)$, where $K$ is a constant, has a Laplace transform given by

$$\mathcal{L}Kf(t) = K\mathcal{L}f(t)$$

These two properties are termed additivity and homogeneity, respectively.

Application. The inverse Laplace transform of a rational function of $s^\dagger$ can be obtained from Table 4-1 by performing the partial-fraction expansion and applying the additivity and homogeneity properties.

4.1.3. Differentiation. If a function $f(t)$ and its time derivative $df/dt$ have Laplace transforms, then

† If the degree of the numerator of the fraction is the same as or is larger than that of the denominator, the partial-fraction expansion exhibits terms of the form $A, Bs, Ce^t$, etc., which are not found in Table 4-1. For such terms, see Secs. 4.3.2 and 4.3.3.
where \( f(0) \) is the value of \( f(t) \) for \( t = 0 \). If \( f(t) \) has a discontinuity at \( t = 0 \), the value taken for \( f(0) \) should be \( f(0^+) \), that is, the value of \( f(t) \) as \( t \) approaches zero through positive values. In particular, if \( f(0) = 0 \), then

\[
\mathcal{L} \frac{df}{dt} = s \mathcal{L}(f(t))
\]

that is, \( s \) appears as an operator for differentiation.

**Generalization.** Applying the differentiation theorem to \( df/dt \) gives the Laplace transform of the second derivative \( d^2f/dt^2 \):

\[
\mathcal{L} \frac{d^2f}{dt^2} = s^2 \mathcal{L}(f(t)) - sf(0) - \frac{df}{dt}(0)
\]

Similarly, for the third derivative:

\[
\mathcal{L} \frac{d^3f}{dt^3} = s^3 \mathcal{L}(f(t)) - s^2f(0) - sf(0) - \frac{d^2f}{dt^2}(0)
\]

### 4.1.4. Application to Solving Differential Equations.

The following steps are involved in integrating a linear differential equation with constant coefficients:

1. Using the table of Laplace transforms, write the Laplace transforms of both sides of the differential equation.
2. Solve the algebraic equation obtained for the Laplace transform of the unknown function.
3. Using the table of Laplace transforms in reverse direction, find the unknown function.

**Example.** Solve

\[
\frac{dy}{dt} - y(t) = u(t)
\]

with the initial condition \( y(0) = 3 \).

**Solution.** 1. Equating the Laplace transforms of both sides gives

\[
s \mathcal{L}(y) - 3 - \mathcal{L}(y) = \frac{1}{s}
\]

2. Solving for \( \mathcal{L}(y) \) and using partial-fraction expansion gives

\[
\mathcal{L}(y) = \frac{1 + 3s}{s(s - 1)} = -\frac{1}{s} + \frac{4}{s - 1}
\]

3. Find the unknown function

\[
y = -1 + 4e^t
\]

The same method applies to integrating simultaneous linear differential equations with constant coefficients.
4.2. OTHER IMPORTANT PROPERTIES

4.2.1. The Time-lag Theorem. If \( f(t) \) has a Laplace transform \( F(s) \), the Laplace transform of the function obtained by delaying \( f(t)u(t) \) by a time \( \tau \) is

\[ e^{-\tau s}F(s) \]

where \( e^{-\tau} \) is called delay factor, or lag factor.

![Fig. 4-1. Function \( u(t - \tau) \).](image)

![Fig. 4-2. Function \( u(t) - u(t - \tau) \).](image)

The function of time whose Laplace transform is \( e^{-\tau s}F(s) \) is

\[ f(t - \tau)u(t - \tau) \]

that is

\[
\begin{align*}
0 & \quad t < \tau \\
f(t - \tau) & \quad t > \tau
\end{align*}
\]

**Example.** The Laplace transform of \( u(t - \tau) \) (Fig. 4-1) is \( e^{-\tau}(1/s) \).

The Laplace transform of

\[ u(t) - u(t - \tau) \]

(Fig. 4-2) is

\[
\frac{1}{s} - e^{-\tau}\frac{1}{s} = \left(1 - e^{-\tau}\right)\frac{1}{s}
\]

**Important Consideration.** When applying the time-lag theorem, it must be borne in mind that it is applied to the function \( f(t)u(t) \). Correspondingly, note that \( e^{-\tau F(s)} \) is the Laplace transform of

\[ f(t - \tau)u(t - \tau) \]

More convincing than a lengthy explanation is Fig. 4-3a, b, and c: the Laplace transform of the first function is \( 1/(s^2 + 1) \), the transform of the second (not of the second) is \( e^{-\tau/2}/(s^2 + 1) \).

![Fig 4-3. Remark concerning the lag theorem.](image)

4.2.2. Application: Transform of Periodic Function. If \( g(t) \) is a function of time which is zero outside the interval \((0,T)\) and has a Laplace transform \( G(s) \) (Fig. 4-4), the periodic function \( g_T(t) \), which consists of
the repetition of \( g(t) \) at intervals of time \( T \) (Fig. 4-5) has a Laplace transform

\[
G_T(s) = G(s) + e^{-sT}G(s) + e^{-2sT}G(s) + \cdots \\
= G(s) \frac{1}{1 - e^{-sT}}
\]

The Laplace transform of \( g_T(t) \) is thus obtained by multiplying the Laplace transform of \( g(t) \) by \( 1/(1 - e^{-sT}) \).

**Fig. 4-4.**

**Fig. 4-5.**

**Fig. 4-6.** Sequence of positive pulses.

**Fig. 4-7.** Sequence of alternately positive and negative pulses.

**Example 1.** The Laplace transform of an infinite sequence of positive pulses (amplitude \( M \), duration \( \tau \)) following one another at regular intervals \( T \) (Fig. 4-6) is

\[
\frac{M \left( 1 - e^{-s\tau} \right)}{s \left( 1 - e^{-sT} \right)}
\]

**Example 2.** The Laplace transform of an infinite sequence of alternately positive and negative pulses (amplitude \( M \), duration \( \lambda T/2 \)) following each other at constant intervals \( T \) (Fig. 4-7) is the product of

\[
\frac{M}{s} \left( 1 - e^{-\lambda T/2} - e^{-sT/2} + e^{-\left(\lambda + 1\right)sT/2} \right)
\]

by the factor \( 1/(1 - e^{-sT}) \).
If \( \lambda = 1 \), the sequence of pulses takes the aspect of a square wave; the Laplace transform becomes

\[
\frac{M}{s} \left( 1 - e^{-T/s} \right) = \frac{M}{s} \tanh \frac{sT}{4}
\]

Conversely, if a periodic function \( g_T(t) \) with a period \( T \) (Fig. 4-8) has a Laplace transform \( G_T(s) \), the Laplace transform of its first cycle (Fig. 4-9) is

\[
G(s) = G_T(s) \left( 1 - e^{-sT} \right)
\]

**4.2.3. The Initial- and Final-value Theorems.** The initial value \( f(0^+) \) and the final value \( f(\pm \infty) \) of a function \( f(t) \), if they exist, can be obtained from the function’s Laplace transform \( F(s) \) by using the relations

\[
f(0^+) = \lim_{s \to \infty} sF(s)
\]

\[
f(\pm \infty) = \lim_{s \to 0} sF(s)
\]

There are no special validity restrictions for the initial-value theorem. The final-value theorem is valid only if no pole of \( F(s) \) [i.e., value of \( s \) for which \( F(s) \) is infinite] has a positive or zero real part.

*Example.* The function whose Laplace transform is

\[
\frac{s + a}{s(s^2 + bs + c^2)} \quad b > 0
\]

has a final value

\[
\lim_{s \to \infty} \frac{s + a}{s^2 + bs + c^2} = \frac{a}{c^2}
\]

Its initial value is

\[
\lim_{s \to 0} \frac{s + a}{s^2 + bs + c^2} = 0
\]

and the initial value of its first derivative is

\[
\lim_{s \to 0} \frac{s(s + a)}{s^2 + bs + c^2} = 1
\]

**4.3. IMPULSE FUNCTIONS**

**4.3.1. General.** An ideal impulse is well approximated by a function of time whose duration is very short compared to the time constants of the system it is applied to but whose amplitude is great enough for the function to have a noticeable effect. An example is a force or a torque.
f(t) that is exerted for very short duration Δt but whose magnitude is sufficient for the integral \( \int_0^{\Delta t} f(t) \, dt \) to be finite and not negligible. In what follows, \( f(t) \) will be, for the sake of simplicity, considered as constant in the interval \((0, \Delta t)\).

### 4.3.2. Definition and Laplace Transform

Consider the function (Fig. 4-10) defined as follows:

\[
\begin{align*}
  f(t) &= 0 & t < 0 \\
  f(t) &= K & 0 < t < \frac{1}{K} \\
  f(t) &= 0 & t > \frac{1}{K}
\end{align*}
\]

This function has a Laplace transform which is (see Sec. 4.2.1):

\[
F(s) = \frac{K}{s} \left(1 - e^{-s/K}\right)
\]

One may question what happens when \( K \) becomes infinite, i.e., when the crosshatched rectangle of Fig. 4-10 becomes infinitely narrow while its area remains unity. Since

\[1 - e^{-s/K}\]

is equivalent to \( s/K \), it is seen that \( F(s) \) becomes unity.

The function of time toward which \( f(t) \) tends as \( K \to \infty \) is termed the **unit-impulse**, or **Dirac delta, function**. It is denoted by \( u_1(t) \) or \( \delta(t) \) and is usually defined by specifications:

\[
\begin{align*}
  u_1(t) &= 0 & t \neq 0 \\
  u_1(t) &= "\text{infinite}" & t = 0 \\
  \int_{-\infty}^{+\infty} u_1(t) \, dt &= 1
\end{align*}
\]

Its Laplace transform is

\[\mathcal{L}u_1(t) = 1\]

More generally, the Laplace transform of an impulse with strength (area) \( A \) is \( \mathcal{L}A\delta(t) = A \), which shows that impulses are the functions of time whose Laplace transforms are constants.

### 4.3.3. Remark

The impulse function \( Au_1(t) \) can be considered, intuitively, as the time derivative of the step function \( Au(t) \) (Fig. 4-11). This verifies

\[\mathcal{L}Au_1(t) = s\mathcal{L}Au(t) = A\]
More generally, it is possible to define a second impulse function \( Au_2(t) \) which is the time derivative of \( Au_1(t) \) (Fig. 4-12) and whose Laplace transform is

\[
\mathcal{L}Au_2(t) = As
\]

A third impulse function may also be defined (Fig. 4-13):

\[
u_3(t) = \frac{d}{dt} u_2(t) \quad \mathcal{L}Au_3(t) = As^2\]

4.3.4. Interest. Impulse functions are used to express schematically a function of time the duration of which is very short referred to the time scale of the phenomena under study.

Fig. 4-12. Second impulse function (see Fig. 4-11).

Fig. 4-13. Third impulse function (see Fig. 4-11).

Fig. 4-14.

Fig. 4-15.

For example, if an airplane with a natural period of 3 sec is subjected to a gust which acts as a torque, say, with a magnitude of 75,000 kg for a duration of 0.2 sec (Fig. 4-14), the gust may be represented by the impulse function:

\[
\tau(t) = 15,000 u_1(t) \text{ kg}
\]

\[
2\tau(t) = 15,000 \text{ kg-sec}
\]
When, in a feedback control system, the system to be controlled is unstable, the correction signal at the input of the system to be controlled frequently has the shape indicated in Fig. 4-15a and b, where \( \gamma_0 \) is the new steady-state value to be attained after a disturbance has taken place. The function shown in Fig. 4-15a roughly corresponds to a correction signal whose Laplace transform is \( \gamma_0/s + \gamma_1 \), that of Fig. 4-15b corresponds to \( \gamma_0/s + \gamma_1 s \). These figures indicate the control-surface deflection of a stabilized airplane subject to a sudden disturbance. Figure 4-15a is the case of a conventional unstable airplane; Fig. 15b corresponds to the case of certain types of helicopters, which are "twice as unstable."

4.4. OTHER PROPERTIES OF THE LAPLACE TRANSFORM

4.4.1. Change of Time Scale. If \( a \) is a quantity that does not depend on \( t \) and \( s \), then

\[
\mathcal{L}f \left( \frac{t}{a} \right) = aF(as)
\]

4.4.2. Translation in the \( s \) Plane. If \( a \) is a constant with nonnegative real part, then

\[
F(s + a) = \mathcal{L}e^{-at}f(t) \\
F(s - a) = \mathcal{L}e^{at}f(t)
\]

In other words, changing \( s \) to \( (s + a) \) corresponds in the time domain to multiplying \( f(t) \) by the exponential \( \exp(-at) \). This makes it possible, for example, to find \( \mathcal{L}e^{-at}\cos \omega t \) directly from \( \mathcal{L} \cos \omega t \) (Table 4-1).

This property may be compared with the time-lag theorem, which corresponds to a translation in the time domain and can be stated as follows: changing \( t \) to \( (t + T) \) corresponds to multiplying \( F(s) \) by \( \exp(-sT) \).

4.4.3. Integration. Let \( \mathfrak{F}(t) \) be an integral of \( f(t) \):

\[
\mathfrak{F}(t) = \mathfrak{F}(0) + \int_0^t f(t') \, dt'
\]

Its Laplace transform is

\[
\frac{\mathfrak{F}(0)}{s} + \frac{\mathcal{L}f(t)}{s}
\]

In particular, if \( F(0) = 0 \)

\[
\mathcal{L} \int_0^t f(t') \, dt' = \frac{\mathcal{L}f(t)}{s}
\]

that is, \( 1/s \) is an integrating factor.

4.4.4. Derivation in the \( s \) Plane. The following relation holds:

\[
\mathcal{L}tf(t) = -\frac{d}{ds}F(s)
\]

This enables \( \mathcal{L}t \) to be obtained directly from \( \mathcal{L}u(t) \), etc. (Table 4-1).

4.4.5. Convolution Integral, or Borel's (Duhamel's) Theorem, or Complex Multiplication Theorem. If \( f_1(t) \) and \( f_2(t) \) have Laplace transforms \( F_1(s) \) and \( F_2(s) \), respectively, then

\[
\mathcal{L} \int_0^t f_1(t - \tau)f_2(\tau) \, d\tau = F_1(s)F_2(s)
\]
This theorem can be used for finding $f_1(t)$ from $F_1(s)$ by choosing $f_2(t)$ and $F_2(s)$ properly. When $F_1(s)$ is an algebraic fraction, however, partial-fraction expansion leads to simpler computation. Applications of the convolution theorem will be given in Secs. 7.3.3 (Example 3 and Note 1) and 23.2.

4.5. BASIC CONCEPTS CONCERNING FOURIER TRANSFORMS

4.5.1. Fourier-series Expansion for a Periodic Function. Any periodic function $p_T(t)$ with a period $T$ (Fig. 4-16), i.e., with an angular frequency $2\pi/T$, can be expanded in the form of a Fourier series, i.e., written in the form:

$$p_T(t) = \frac{a_0}{T} + \left( a_1 \cos \frac{2\pi}{T} t + b_1 \sin \frac{2\pi}{T} t \right) + \left( a_2 \cos \frac{4\pi}{T} t + b_2 \sin \frac{4\pi}{T} t \right) + \cdots$$

$$p_T(t) = \frac{a_0}{T} + \sum_{n=1}^{\infty} \left( a_n \cos \frac{2\pi n}{T} t + b_n \sin \frac{2\pi n}{T} t \right)$$

where the integer $n$ is the rank of the harmonic, and the coefficients $a_n$ and $b_n$, giving the relative amplitude of the $n$th harmonic, are

$$a_0 = \int_{-T/2}^{T/2} p_T(t) \, dt \quad a_n = \int_{-T/2}^{T/2} p_T(t) \cos \frac{2\pi n}{T} t \, dt$$

$$b_n = \int_{-T/2}^{T/2} p_T(t) \sin \frac{2\pi n}{T} t \, dt$$

**Example 1.** The square-wave function shown in Fig. 4-17, with an amplitude $M$ and a period $T = 2\pi/\omega$, can be written as

$$p_T(t) = \frac{4M}{\pi} \left( \sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \cdots \right) = \frac{4M}{\pi} \sum_{n=1}^{\infty} \frac{\sin n\omega t}{n}$$

where $\sum$ indicates summation for all positive odd values of $n$, the even harmonics being all zero.

![Fig. 4-16. Example of periodic function.](image)

![Fig. 4-17. Square wave.](image)
Example 2. The sequence of alternately positive and negative pulses shown in Fig. 4-18, with an amplitude $M$ and a period $T = 2\pi/\omega$, can be written as

$$p_T(t) = \frac{4M}{\pi} \sum_{\lambda=1}^{\infty} \sin \frac{n\pi}{2} \cos \left( \omega t - \frac{\pi}{2} \right)$$

where $\sum_{\lambda}$ indicates summation for all odd values of $n$.

4.5.2. Application to Nonperiodic Functions. The equations quoted in Sec. 4.5.1 can be applied to a nonperiodic function $f(t)$ for a given value of $T$. The result is a series the sum of which is a function $f_T(t)$ which is (a) identical to $f(t)$ in the interval $(-T/2, +T/2)$ (b) periodic, of period $T$.

![Fig. 4-18. Sequence of pulses.](image)

4.5.3. Fourier Transform. If $T$ tends toward infinity, it can be shown that\(^1\) the series $f_T(t)$ goes into an integral

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{j\omega t} \, d\omega$$

and that the expressions for the coefficients $a_n$, $b_n$ merge into a continuous function

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} \, dt$$

This function $F(\omega)$ is termed the Fourier transform of $f(t)$ and is denoted by $\mathcal{F}f(t)$. Note that this definition does not require that $f(t)$ be zero for $t < 0$.

If $f(t)$ is zero for all negative values of $t$, the integral which defines $F(\omega)$ can be taken from 0 to $+\infty$. It is, then, easy to show (Chap. 4 appendix, Sec. 1) that the Fourier transform has the same functional form as the Laplace transform (if both exist), $s$ being replaced by $j\omega$.

4.6. OBTAINING THE TIME RESPONSE OF A LINEAR SYSTEM

4.6.1. General. Obtaining the time response $r(t)$ of a linear system for a given input $e(t)$ usually involves the following steps:

a. Writing the equations of the system.

b. Specifying the initial conditions pertaining to the problem.

c. Obtaining the Laplace transform $R(s)$ of the response. It is desirable to check the value $r(0)$ by means of the initial-value theorem. The

\(1\) This happens if the function $f(t)$ fulfills certain conditions that concern, essentially, the behavior of $f(t)$ as $t$ approaches infinity.
final value $r(\infty)$ and the initial values of $dr/dt$, $d^2r/dt^2$ can usually be obtained directly from the Laplace transform.

d. Determining the poles of $R(s)$. This indicates the mathematical nature of the different terms of $r(t)$.

e. Expanding $R(s)$ by the method of partial fractions, obtaining $r(t)$ by inverse transformation, and plotting $r(t)$ by graphical addition of its component terms.

Note. Although it is not absolutely necessary to obtain the initial and final values from $R(s)$ directly, it is advisable to do so before proceeding to step d. This provides a check on the computation.

4.6.2. Specifying the Initial Conditions. The most difficult step in the whole procedure is frequently to state the initial conditions pertaining to the problem. To do this, it is necessary to analyze the situation in the neighborhood of $t = 0$ by reasoning physically and taking the equations into account. When discontinuities are involved, considerable difficulty may arise.\(^1\)

Fortunately, for all cases in which the system starts from rest, the Laplace transform of the response can be obtained immediately, i.e., without having to analyze the initial conditions in detail (Sec. 7.3.4). This is made possible by using the concept of transfer functions, which will be outlined in the following chapters.

APPENDIX

The purpose of this appendix is to provide a mathematical complement for the subject covered in Chap. 4. The mathematical definition and some additional properties of the Laplace transform are given. Mathematical proof for the fundamental properties is presented without detailed treatment of the problems of convergence. Such aspects usually do not cause difficulties in the servo field. The reader especially interested in them is referred to the Bibliography.

1. Mathematical Definition. If $f(t)$ is a function of time which is zero (or which is not defined) for $t < 0$, the Laplace transform $F(s)$ of $f(t)$ is defined as

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st}f(t)\,dt$$

for all the values of $s$ for which the integral converges. Generally this integral converges when the real part of $s$ is greater than a certain number $\sigma$ or $\sigma_c$ termed the abscissa of convergence, i.e., when the complex number $s$ lies to the right of the straight line $\text{Re}\,s = \sigma_c$ in the complex plane (Fig. 4-19).

This definition enables one to find the Laplace transforms of the functions listed in Table 4-1, by means of integration by parts. The abscissa of convergence is equal to:

a. Zero when \( f(t) \) remains finite or becomes equivalent, as \( t \) tends toward infinity, to a finite power of \( t \), for example, when \( f(t) \) for \( t > 0 \) is one of the following functions: 1, \( t \), \( t^n \), \( \cos \omega t \), \( \sin \omega t \), \( t \cos \omega t \), etc.

b. a when \( f(t) \) behaves, as \( t \) approaches infinity, like \( \exp (-at) \), where \( a \) is an algebraic quantity, e.g., when \( f(t) \) is one of the following functions: \( e^{-at} \), \( te^{-at} \), \( e^{-at} \cos \omega t \), \( te^{-at} \sin \omega t \), etc. Note that functions of time that increase faster than an exponential, for example, \( f(t) = \exp (t^2) \), have no Laplace transform, since it is impossible to find suitable values of \( \sigma \) which assure the convergence of \( \int_0^\infty e^{-\sigma t}f(t) \, dt \).

2. Proof of Some Properties. a. Differentiation and Integration. The differentiation theorem can be proved by integrating by parts the integral that defines \( F(s) \). Letting \( u = f(t) \) and \( dv = e^{-at} \, dt \), the familiar formula \( u \, dv = uw - v \, du \) yields

\[
F(s) = \left[-\frac{1}{s} e^{-at}f(t)\right]_0^\infty + \frac{1}{s} \int_0^\infty e^{-at} \frac{df}{dt} \, dt
\]

The first term on the right is \( f(0^+) / s \), when \( f(0^+) \) is the limit of \( f(t) \) as \( t \) decreases toward zero. Hence, after multiplication by \( s \):

\[
\mathcal{L} \frac{df}{dt} = sF(s) - f(0^+)
\]

The integration theorem can be similarly proved by integration by parts.

b. Final and Initial Value. Suppose \( F(s) \) has no singularity in the right half plane or on the imaginary axis. The equation

\[
\mathcal{L} \frac{df}{dt} = \int_0^\infty \frac{df}{dt} e^{-at} \, dt = sF(s) - f(0^+)
\]

becomes, as \( s \) approaches zero,

\[
\int_0^\infty \frac{df}{dt} \, dt = \lim_{s \to 0} [sF(s) - f(0^+)]
\]

where the left-hand side can be written

\[
\lim_{t \to \infty} \int_0^t \frac{df}{dt} \, dt = \lim_{t \to \infty} [f(t) - f(0^+)]
\]

whence the theorem.

The initial-value theorem can be proved in a similar manner by letting \( s \) approach infinity.

c. Convolution Integral. Borel's theorem

\[
\mathcal{L} \int_0^t f_1(t - \tau)f_2(\tau) \, d\tau = F_1(s)F_2(s)
\]

can be proved as follows: Let

\[
f(t) = \int_0^t f_1(t - \tau)f_2(\tau) \, d\tau
\]

The function \( f(t) \) has a Laplace transform

\[
\int_0^\infty \int_0^t f_1(t - \tau)f_2(\tau) \, d\tau e^{-st} \, dt = \int_0^\infty \int_0^\infty f_1(t - \tau)f_2(\tau)u(t - \tau) \, d\tau e^{-st} \, dt
\]

which can be written by changing the order of integration

\[
\int_0^\infty f_2(\tau) \, d\tau \int_0^\infty f_1(t - \tau)u(t - \tau)e^{-st} \, dt
\]
The second integral is \( F_1(s) \exp(-st) \), as can be proved by letting \( \lambda = t - \tau \) and integrating from \( \tau \) to \( \infty \), which does not change its value; hence,

\[
\mathcal{L}f(t) = F_1(s) \int_0^\infty f_2(\tau)e^{-st}d\tau = F_1(s)F_2(s)
\]

3. The Inverse Laplace Transform. A function \( f(t) \) can be obtained from its Laplace transform by the relation

\[
f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} e^{-st}F(s)\,ds
\]

where \( c \) is a real constant greater than the real part of any singularity of \( F(s) \). It is interesting to compare this equation with that for inverting the Fourier transform (Sec. 4.5.3).

4. Conditions of Existence for the Laplace and the Fourier Transforms. (Fig. 4-20a, b, and c). In most usual cases, a function which is zero for all negative values

![Diagram](image)

**Fig. 4-20.** (a) Both Laplace \( (\sigma < 0) \) and Fourier transforms exist. (b) Laplace transform exists \( (\sigma > 0) \). There is no Fourier transform. (c) Laplace transform exists with \( \sigma = 0 \). There is, in general, no Fourier transform.
of \( t \) and has a Fourier transform also has a Laplace transform. In fact, if the integral of \( f(t) \exp(-j\omega t) \) converges, the integral \( f(t) \exp(-st) \) will surely converge if \( \text{Re} \ s > 0 \). Conversely, if a function \( f(t) \) has a Laplace transform

\[
F(s) = \int_0^\infty f(t)e^{-st} \, dt
\]

with \( \text{Re} \ s > \sigma > 0 \), it has in general no Fourier transform, since it is necessary for the integral to incorporate a decreasing exponential in order to obtain convergence. This is also often the case if \( \sigma = 0 \); thus the unit-step function \( u(t) \) has a Laplace transform, \( 1/s \) with \( \sigma = 0 \), but it has no Fourier transform, since the integral of \( u(t) \exp(-j\omega t) \) does not converge for \( t = \infty \).

In conclusion, the presence of an exponential with complex exponents (and not a pure imaginary exponent) in the Laplace integral renders the conditions of existence for the Laplace transform simpler and less severe than for the Fourier transform.

5. The Carson Transform. Several authors make use of the Carson transform

\[
c_{\alpha f}(t) = s \int_0^\infty e^{-sf(t)} \, dt = s\mathbb{L}[f(t)]
\]

The Carson transform has two advantages:

a. The Carson transform of a constant \( k \) is \( k \) (whereas the Laplace transform is \( k/s \), which, for absent-minded people, causes frequent errors).

b. If the time scale of a function \( f(t) \) with a Carson transform \( c(s) \) is changed, the new Carson transform is

\[
c_{\alpha f}(\alpha t) = c(\alpha s)
\]

whereas the new Laplace transform is \( \alpha F(\alpha s) \) (Sec. 4.4.1). This makes it possible to consider the Carson transform of a function \( f \) without having to specify what the independent variable is, as must be done when Laplace transforms are used.

On the other hand, the Laplace transform has the advantage of having the same form as the Fourier transform, when both transforms exist. At present, the Laplace transform is the one almost exclusively used in control-theory work.
CHAPTER 5

FIRST-ORDER SYSTEMS

Summary
1. Time constant. Transient response.
2. Harmonic response.
4. Examples of first-order systems.
5. Lags and time constants.

It has been shown in the preceding chapter how the Laplace transform enables one to study the behavior of linear systems, that is, to study their response under various input conditions. It will be shown later that this leads to characterizing a linear system by a function of the variable \( s \) that completely characterizes it and is called its transfer function. These applications will be introduced in a progressive manner, starting from simple systems (Chaps. 5 and 6), and progressing to the general terms outlined in Chaps. 7, 8, and 9.

5.1. TRANSIENT RESPONSE

5.1.1. Differential Equation, Time Constant. Consider a mercury thermometer, as an example. This device translates the outside temperature \( \theta \) into an output quantity \( r \), which is the temperature shown on the graduated scale and constitutes the output of the system (Fig. 5-1). If the instrument, which is assumed to be accurate and precise, is placed for a sufficient time in constant-temperature surroundings, the indicated temperature will, after a certain time, be equal to the ambient temperature. This is the steady state, \( r = \theta_0 \).

However, if \( \theta(t) \) changes rapidly, the expression for \( r(t) \) can be obtained only by writing the differential equation of the instrument. Since the rate of expansion of mercury \( dr/dt \) is proportional to the difference between its temperature and that of the surrounding medium, the differential equation of the system is

\[
\frac{dr}{dt} = K(\theta - r)
\] (5-1)

Under normal conditions, this assumption is approximately verified by experiment. Equation (5-1) can be written as
where the constant quantity $T$ has the dimension of time and is called the time constant of the system. It will be seen that the time constant $T$ completely characterizes the performance of the system.

5.1.2. Response to a Unit Step. 

The response to a unit step (Fig. 5-2) is obtained by integrating Eq. (5-1)\(^1\) and using the initial condition that $\theta = \theta_0$ when $r = 0$. The result is

$$r = \theta_0(1 - e^{t/T})$$

The response is an exponential curve asymptotic to the new steady state represented in Fig. 5-3.

**Fig. 5-3.** Unit-step response of first-order system with time constant $T$.

*Interpretation of the Time Constant.* It is to be noted that the time constant of the system characterizes the time scale for the transient. The time constant can be defined graphically in the two following ways:

1. The time constant $T$ can be expressed as $\tan \alpha$, where $\alpha$ is the slope of the response curve at the origin (Fig. 5-4).

2. The time constant is the time required for the instrument to have achieved $1 - 1/e = 0.633 \approx 2/3$ of its steady-state response (Fig. 5-5).

If the response is plotted vs. nondimensional time $t/T$—that is, if the

\(^1\) First-order systems, because of their simplicity, are the only case for which the Laplace transform is of little advantage in obtaining the solution of the differential equation.
time constant $T$ is taken as the unit of time—Fig. 5-6 applies to all first-order systems.

Response Time. Mathematically, the steady state is reached only after an infinite time, since the transient component is an exponential function of time. In practice, the duration of the transient is usually

\[ r(t) = \begin{cases} 
1 - e^{-\frac{t}{T}} & \text{for } 1 - \frac{t}{T} < 0 \\
1 & \text{for } 1 - \frac{t}{T} \geq 0 
\end{cases} \]

Fig. 5-4. First graphical interpretation of the time constant $T$.

Fig. 5-5. Second graphical interpretation of the time constant $T$.

characterized by the response time to within 5 per cent. In other words, the transient is considered to persist as long as the system has not attained 95 per cent of its steady-state condition and stayed there. Since 5 per cent is approximately $1/e^2$, the response time is three times the time constant.

A response time to within 2 per cent, sometimes called the indication time, may also be considered. Since 2 per cent is approximately $1/e^4$, the indication time is four times the time constant.

5.1.3. Impulse Response. The response of the system to a unit impulse $\theta = u(t)$ is found to be (Fig. 5-6)

\[ r(t) = \frac{1}{T} e^{-\frac{t}{T}} \]
The peak value attained is inversely proportional to the time constant \( T \); physical reasons render this fact rather obvious. The duration of the transient is characterized by the time constant \( T \), with a response time to within 5 per cent equal to \( 3T \).

5.1.4. Ramp Response. The response of the thermometer to an input of the form \( \theta = atu(t) \), that is, when temperature increases at a constant rate, can be easily found.

**Steady State.** It is found that

\[
r = a \cdot (t - T)
\]

that is

\[
r(t) = \theta(t - T)
\]

At time \( t \) the scale indicates the temperature that existed at time \( (t - T) \), that is, the response lags the input by \( T \) sec.

Fig. 5-7. Ramp response of first-order system, showing velocity error.

Fig. 5-8. Ramp response of first-order system, showing transient.

Another way to express the same idea (Fig. 5-7) is to say that there is a steady-state error equal to \( aT \):

\[
r(t) = \theta(t) - aT
\]

This error is the velocity error, defined in Sec. 3.3.2. It would be equal to zero if the time constant were infinitely short \((3T = 0)\). Note that this error exists regardless of the properties of the steady-state characteristics.

**Transient.** The transient component of the response is an exponential which links the initial steady state

\[
r = 0 \quad (t < 0)
\]

with the final steady state

\[
r(t) = \theta(t - T)
\]

At time \( t = T \), its value is

\[
\frac{1}{e} aT \cong \frac{aT}{3}
\]

which permits, in practice, the plotting of the response curve with sufficient accuracy (Fig. 5-8).
5.2. HARMONIC RESPONSE

5.2.1. General Considerations. For a sinusoidal input

\[ \theta = \theta_0 \sin \omega t u(t) \]

the forced response is

\[ r = r_0 \sin (\omega t + \Phi) \]

with

\[ r_0 = \frac{\theta_0}{(1 + T^2 \omega^2)^{1/4}} \quad \Phi = - \arctan \omega T \]

The output is not an exact reproduction of the input since there is an error characterized by a change in amplitude

\[ \frac{r_0}{\theta_0} = \frac{1}{(1 + T^2 \omega^2)^{1/4}} \lesssim 1 \]

and by a phase shift

\[ \arctan \omega T \neq 0 \]

As a result, the instrument gives an almost exact indication only for slowly varying phenomena, i.e., when \( \omega T \ll 1 \).

![Fig. 5-9. Frequency response of first-order system.](image)

The phase and amplitude response is represented as a function of angular frequency in Fig. 5-9, where the ordinate is \( r_0/\theta_0 \), nondimensional amplitude or amplitude ratio, and \( \Phi \), the phase of \( r \) with respect to \( \theta \). The phase angle of \( r \) with respect to \( \theta \) is negative: i.e., the response lags the input by the positive quantity \( \arctan \omega T \). Note that the time constant \( T \) is the slope of the phase-response curve (radians vs. radians per second) at low frequencies, since in the neighborhood of \( \omega = 0 \)

\[ \Phi = - \arctan \omega T \simeq - T \omega \]
5.2.2. Use of Nondimensional Angular Frequency. Occasionally the nondimensional angular frequency in a first-order system of time constant \( T \) is defined as \( u = \omega T \). Figure 5-10 shows the frequency response of first-order systems as a function of this nondimensional angular frequency.

5.2.3. Transfer Locus, or Nyquist Locus. The information given by Fig. 5-10 can be condensed by plotting the locus of the points defined in polar coordinates by: (a) radius vector = nondimensional amplitude \( A(\omega) \) and (b) polar angle = phase \( \Phi(\omega) \). The locus thus obtained, scaled in units of \( \omega \), is termed the frequency-response locus, transfer locus, or Nyquist locus of the system. It is a semicircle graduated in nondimensional angular frequency (Fig. 5-11) and is the locus of end points of \( 1/(1 + jT \omega) \) for the variable \( \omega, j^2 \) being equal to \(-1\).

5.2.4. Use of Logarithmic Coordinates. 1. General Considerations. For reasons which will be explained later (Chap. 8) it is often useful to scale the nondimensional angular frequency and the nondimensional amplitude axis in logarithmic coordinates. Or, if the nondimensional amplitude is expressed in decibels (abbreviated db), the axis is scaled in a linear manner (semilogarithmic coordinates), the value \( A \), in decibels, being defined as

\[
A \text{ db} = 20 \log A
\]

Figures 5-12, 6-9a and b, and Charts 1 and 2 at the back of the book are constructed in this way. Because of their usefulness in subsequent chapters (see especially Chap. 7), it is important that the reader be familiar with these curves. Because of this, their structure must be known quantitatively, and it will be explained, therefore, in great detail.
2. Amplitude-response Curve. This curve has two asymptotes and these intersect at the point \((T\omega = 1, A = 1, \text{ that is, } 0 \text{ db})\). For this reason the angular frequency \(\omega = 1/T\) is sometimes termed the break frequency of the system and the point \((0 \text{ db, } T\omega = 1)\) the break point.

The horizontal asymptote is the straight line \(A = 1\), that is, 0 db, and is approached by the curve at low frequencies. The oblique asymptote passes through the break point. It is the curve \(A = 1/T\omega\), which gives a straight line in logarithmic coordinates. To find its slope, two frequencies \(\omega_1\) and \(\omega_2\) are so taken that

\[
\omega_2 = 2\omega_1
\]

and the corresponding values \(A_1\) and \(A_2\) of \(A\) are compared. It follows that

\[
A_1 \text{ db } = -20 \log T\omega_1 \\
A_2 \text{ db } = -20 \log T\omega_2 = -20 \log 2 - 20 \log T\omega_1
\]

from which

\[
(A_1 - A_2) \text{ db } = 20 \log 2 = 6 \text{ db}
\]

When the frequency is doubled, the ordinate is lowered by 6 db, and, consequently, the slope is 6 db per octave.
For practical purposes, the position of the curve with respect to the asymptote is sufficiently well specified by the following properties:

a. For the break frequency \( \omega = 1/T \), the curve is 3 db below the break point.

b. For twice this frequency \((2/T)\), it is 1 db below the oblique asymptote.

c. For half the break frequency \((1/2T)\), it is 1 db below the horizontal asymptote.

\[ A = \frac{1}{\sqrt{1 + u^2}} \]

\[ \Phi = \arctan u \]

Fig. 5-12. Frequency response of first-order system using logarithmic coordinates (Bode plot).

3. Phase-response Curve. At very low frequencies \((\omega \ll 0)\), \(\Phi = 0\). At very high frequencies \((\omega \to \infty)\), \(\Phi = -90^\circ\). The phase angle is always negative and decreasing. This implies that the phase lag of the output with respect to the input increases from 0 to 90° as the frequency varies from 0 to \(\infty\). The curve has the form of an arctangent. It has an inflection point which is also a point of symmetry. Its coordinates are:

\[ \omega = \frac{1}{T} \text{ (break frequency)} \quad \Phi = -45^\circ \]

Note the following values:

a. At one octave above the break frequency

\[ \omega = \frac{2}{T} \quad \Phi = -63.5^\circ \]
Fig. 5-13. Alignment chart for $Ae^{j\phi} = 1/(1 + jT\omega)$. Sample computation: $T = 0.1$ sec, $\omega = 5$ rad/sec yields $A = 0.89$, $\phi = 27^\circ$.

b. At one octave below the break frequency

$$\omega = \frac{1}{2T} \quad \Phi = -26.5^\circ$$

5.2.5. Nichols Locus, or Amplitude-Phase Plot. An equivalent representation consists in plotting, as the abscissa, $\Phi$ (in degrees) and, as the ordinate, $A$ (in decibels) and in graduating the locus in units of $\omega$ (or $u$). The locus thus obtained is shown as a dotted curve in Fig. 6-9c. It is called the Nichols locus of the system, or its amplitude-phase plot.
5.2.6. Alignment Chart. Finally, the alignment chart of Fig. 5-13 gives, for specified values of \( T \) and \( \omega \), the amplitude ratio and the phase angle:

\[
A = \frac{1}{(1 + T^2 \omega^2)^{1/2}} \quad \Phi = -\arctan \omega T
\]

5.3. RESPONSE TO AN ARBITRARY INPUT, TRANSFER FUNCTION

5.3.1. Rudiments of Filtering. It is easily understood that, when a random input—for example, the sample of a random function of time—is fed into a first-order system with a very short time constant, the output will reproduce the input quite faithfully because the short response time enables the system to follow the variations of the input without appreciable lag (Fig. 5-14). On the contrary, if the same input is fed into a first-order system with a long time constant, the slowness of the system

![Fig. 5-14.](image)

![Fig. 5-15.](image)

prevents its output from reproducing one input variation before the next has arrived. This smoothing of the small variations on the input is termed a filtering effect (Fig. 5-15).

The important factor is not the absolute value of the time constant, but rather its value referred to the period of the predominant harmonic in the input signal.\(^1\) These qualitative considerations will later be generalized (Chaps. 6 and 8).

5.3.2. Concept of Transfer Function. If the input of a system has a Laplace transform \( \Theta(s) \), the Laplace transform of the output is given by

\[
T[sR(s) - r(0)] + R(s) = \Theta(s)
\]

For the particular case in which \( r(0) = 0 \) the equation can be written as

\[
(Ts + 1)R(s) = \Theta(s)
\]

and thus

\[
R(s) = \frac{1}{1 + Ts} \Theta(s)
\]

The Laplace transform of the output is equal to the product of the Laplace transform of the input by \( 1/(1 + Ts) \). In other words, the system is in terms of Laplace transforms, an operator corresponding to

\(^1\) It should be remembered that a rapidly varying function of time is rich in higher-order harmonics.
multiplication by \(1/(1 + Ts)\). The factor \(1/(1 + Ts)\) is termed the transfer function of the system. Note that the frequency-response locus defined in Sec. 5.2.3 is the locus of the tip of the vector representing the transfer function \(H(s)\) when \(s\) is taken as equal to \(j\omega\). These important results will be generalized later (Chaps. 6 to 8).

5.3.3. Problem. Given a physical system, is it possible to determine if it is of first order, i.e., if it can be represented correctly by a differential equation of the type

\[ T \frac{dr}{dt} + r = e \quad ? \]

If this is true, how can its time constant be determined?

**Solution 1.** A unit step is applied to the system, and the response is recorded (Fig. 5-16). If \(r_\infty\) is the limit of the output at \(t = \infty\), plot the curve \(
\log|r(t) - r(\infty)|\) as a function of time, for example, by plotting \(r - r_\infty\) on semilog paper. If this curve is a straight line, the system is of the first order. The time constant is then read from the graph as the time required by the system to reach approximately \(\frac{3}{4}\) (more exactly, \(1 - 1/e\)) of its steady-state value

\[ \frac{r(T) - r(\infty)}{r(0) - r(\infty)} = \frac{1}{e} = 0.367 \]

**Solution 2.** The system is subjected to harmonic tests at different frequencies, and the amplitude and phase of the forced response are recorded at each frequency. If the system is a first-order system, then:

a. At very high frequencies the phase shift should approach \(-\pi/2\).

b. At high frequencies the amplitude should decrease as \(1/\omega\), that is, at a rate of 6 db (20 log 2) per octave (\(\omega_1/\omega_1 = 2\)).

If these conditions are fulfilled, it is possible to find a first-order system whose time constant is the inverse of the angular frequency for which the phase lag is \(45^\circ\). Let this time constant be \(T\). To ensure that the system really is of first order, the results of the trials are compared with the theoretical frequency-response curves of a first-order system with the same time constant. This can be done by computing a few points, or by superimposing the curves. The latter process is generally adequate in practice.

5.4. EXAMPLES OF FIRST-ORDER SYSTEMS

All systems whose equations can be written in the form

\[ T \frac{dr}{dt} + r = e \]
where \( e \) is the input, \( r \) the response, and \( T \) the time constant of the system, are of first order and possess a transfer function \( 1/(1 + Ts) \).

By nondimensionalizing time, \( \tau = t/T \) (\( \tau \) = dimensionless time) one obtains

\[
\frac{dr}{d\tau} + r = e
\]

All first-order systems are analogous if their equations are written in terms of nondimensional time.

1. **RC Circuit.** The elementary circuit for a battery charging a capacitor through a resistance (Fig. 5-17) is a first-order network with a time constant

\[
T = RC \quad \text{seconds, ohms, farads}
\]

For an input

\[
e = Eu(t)
\]

obtained by closing the switch \( K \), the voltage across the capacitor results as

\[
V(t) = E(1 - e^{-t/RC})
\]

![Fig. 5-17.](image)

![Fig. 5-18. Relaxation oscillator.](image)

This system is important because it comprises the fundamental element of relaxation oscillators and saw-tooth generators used in various sweep circuits. The relaxation oscillator consists of two first-order networks, with different time constants as shown in Fig. 5-18, where \( N \) represents a neon tube. The oscillations are produced alternately:

a. By charging the capacitor \( C \) through \( R_1 \). The resulting time constant is \( R_1C \), and the neon tube behaves as an infinite resistance.

b. By discharging the capacitor through \( R_2 \). The resulting time constant \( (R_2 + \rho)C \) is much smaller, and the tube \( N \) behaves like a small resistance when it is conducting, permitting a flow of current (Fig. 5-19). The result is a saw-tooth wave made up of small segments of exponential curves. These segments can be considered as essentially straight if \( V_2 - V_1 \) is small compared to \( V_1 \).

![Fig. 5-19.](image)

2. **LR Network.** The elementary circuit of Fig. 5-20 is of first order if the current \( I \) is considered as the output; the input being the voltage \( E \) applied to the circuit:
\[
\frac{L}{R} \frac{dI}{dt} + I = \frac{E}{R}
\]

Its time constant, stemming from the inductance, is

\[T = \frac{L}{R}\] seconds, henrys, ohms

In the absence of \(L\) one would have for instantaneous values \(I = E/R\). This is the reason why it is often said that inductances introduce unwanted time constants.

An example of this system is afforded by a field-controlled d-c motor (Chap. 31). The inductance of the control coil, plus the inductance of the rest of the control circuit, produces a lag between the controlling voltage and the field. The corresponding time constant is often very important, for example, \(\frac{1}{60}\) to \(\frac{1}{6}\) sec, which is not negligible in servo systems. This is one of the reasons why d-c servomotors are often armature-controlled. In fact, the inductance of the armature is much smaller than that of the field coil.

Note. The formula \(T = L/R\) is often interpreted as follows: The time constant of the system can be considered as the ratio of twice the energy stored in the magnetic field of the system, \(\frac{1}{2}LI^2\), to the rate of energy dissipation \(RI^2\) (Chap. 31).

3. Rate-of-climb Indicator, Altimeter. The rate-of-climb indicator and the altimeter are instruments which measure the rate of climb and the altitude, respectively, as functions of ambient pressure. The rate-of-climb indicator is a very nearly a first-order system, characterized by a time constant which is usually of the order of 2 or 3 sec. The altimeter, which measures the ambient pressure by the deflection of a needle corresponding to the movement of an aneroid capsule, can only be assimilated into a first-order system by means of gross approximations.

Indeed, if \(e\) is the ambient pressure and \(r\) the pressure inside the capsule (Fig. 5-21), it can be proved, assuming that the temperature is constant, neglecting the inertia of the air column, and using the perfect-gas law and Poiseuille's law, that

\[e = r + \frac{\lambda}{r} \frac{dr}{dt} \quad \lambda = 32 \frac{\mu l^3}{D^2}\]

In these equations \(\mu\) is the viscosity coefficient of air and \(l\) and \(D\) are the length and diameter of the pipes that connect the ambient air with the capsule. Thus, although the equation is of first order, its coefficients are not constant. It is, therefore, impossible to speak of a first-order system having \(\lambda/r\) as a time "constant," unless only the regions where the variations of \(e\) and \(r\) are very small are considered.

4. Mechanical System. If positions are taken as variables, a mechanical system cannot be represented by an equation of the first order, because the law of dynamics \(F = ma\) leads to an equation that is of second order. The very exceptional case of a mechanical system being considered of the first order is due to the inertia being neglected, this force being small compared to the other forces acting on the system. This is usually the case for overdamped systems, as long as the input does not vary too rapidly with time (see Sec. 6.4.3, par. 3).

5.5. LAGS AND TIME CONSTANTS

If the time constant \(T\) is small compared to the time scale considered and if the input is a step or ramp, it is reasonable to consider, as a first
approximation, that a first-order system responds with a time lag of the order of $T$ (Sec. 5.1). This approximation is, however, good only when the input varies slowly with respect to $T$. There is an essential difference between a first-order system and a real time lag or delay (Fig. 5-22). For a first-order system a transfer function $1/(1 + Ts)$ has been found.

![Graphs showing time constant and pure lag](image)

**Fig. 5-22.** (a) Time constant. (b) Pure lag.

![Frequency response and Nyquist locus](image)

**Fig. 5-23.** Frequency response for a pure lag $e^{-sT}$.

**Fig. 5-24.** Frequency-response locus (transfer locus) for a pure lag.

The effect of a system characterized by a real lag is to shift the input by a time $T$:

$$r(t) = e(t - T)$$

whence, by means of the lag theorem

$$R(s) = e^{-sT}E(s)$$

which shows that the transfer function of the system is $e^{-sT}$. The frequency-response curves are shown in Fig. 5-23:

$$A(\omega) = 1 \quad \Phi(\omega) = -T\omega$$

The corresponding Nyquist locus $e^{-j\omega T}$ is the circumference of a circle centered at the origin (Fig. 5-24).

At low frequencies, the similarity between a time constant and a lag can be explained as follows:

a. Geometrically, the two loci $1/(1 + j\omega T)$ and $e^{-j\omega T}$ are tangent at the point $\omega = 0$ (Fig. 5-25).

b. Analytically, the first terms of the series expansion for small values
of $s$ are $1 - e^{-sT}$ both for $e^{-sT}$ and for $1/(1 + sT)$. This expresses the fact that, when $T$ is small,

$$f(t + T) \simeq f(t) + T \frac{df}{dt}$$

$c$. Using the frequency-response curve, a real time lag as well as a time constant results in a phase lag at low frequencies of

$$\arctan \omega T \simeq \omega T$$

It should be borne in mind that these approximations are valid only when $T$ is small with respect to the periods of the inputs under consideration (see Sec. 11.2.1, Example 4).

**Example.** The acceleration of a motor from rest can be represented by the superposition of a pure time lag $T$ and a time constant $T_2$ (Fig. 5-26), i.e., by a transfer function

$$e^{-sT_1} \frac{1}{1 + sT_2}$$

For electric motors, $T_1$ is mainly due to coulomb friction and backlash and $T_2$ is due to the inductance of the field coil. Typical orders of magnitude for servo motors are $T_1 = 3$ msec, $T_2 = 6$ msec.
CHAPTER 6
SECOND-ORDER SYSTEMS

Summary

1. Frequency-response curves and loci.
2. Transfer functions and loci.
3. Transient response.
4. Performance of second-order systems.
5. Poles of the transfer function in the complex plane.
6. Examples of second-order systems. Their importance.

The purpose of the present chapter is twofold: first, to study second-order systems and their performance; second, to introduce to the reader, using the example of second-order systems, the fundamental concepts of linear theory that will be generalized later and utilized throughout Parts 1 and 2.

6.1. FREQUENCY-RESPONSE CURVES AND LOCI

6.1.1. Differential Equation of a Second-order System. As an example, a mechanical system (a simplified accelerometer) will be considered. It consists of a mass \( m \) in horizontal rectilinear motion (in the \( x \) direction) with respect to a support to which it is attached by means of a spring that develops a force proportional to its elongation. It will be assumed, in addition, that viscous friction develops a force proportional to the velocity of the mass relative to its support (Fig. 6-1). Furthermore, static, or coulomb, friction, which produces a force of approximately constant amplitude having a direction opposite to that of the velocity of \( m \), will be considered to be negligible.

Hence, if the support has an acceleration \( \gamma \) with respect to the earth, and is directed along \( OX \), the motion of the mass relative to its support is expressed by the equation

\[
m \frac{d^2x}{dt^2} = -kx - f \frac{dx}{dt} + m\gamma
\]  

(6-1)

where \( x \) designates the position of the mass along \( OX \) relative to its support. Specifically, for the static case, \( dx/dt \) and \( d^2x/dt^2 \) are both zero, whence
\[ x = \frac{m}{k} \gamma \]

The displacement of the mass with respect to its support thus furnishes a measure of the acceleration to which the latter is subjected (Fig. 6-2). Equation (6-1) is linear, it has constant coefficients, and it is of the second order. The system which it represents will be called a second-order system.

6.1.2. Harmonic Test, Frequency-response Curves. If the support is subjected to an acceleration that varies sinusoidally with time in the manner

\[ \gamma(t) = \gamma_0 \sin \omega t \]

the mass \( m \), after a few transient oscillations, reaches its steady-state condition. This is a sinusoidal oscillation with a frequency of \( \omega/2\pi \), an amplitude of \( x_0 \neq (m/k)\gamma_0 \), and with a phase that differs from \( \gamma(t) \) by an angle \( \Phi \). \( x_0 \) and \( \Phi \) can be obtained by substituting

\[ x = x_0 \sin (\omega t + \Phi) \]

into the differential equation. This gives, after reduction,

\[ x_0 = \frac{m\gamma_0}{[(k - m\omega^2)^2 + f^2\omega^2]^{1/2}} \]
\[ \Phi = \arctan \frac{-f\omega}{k - m\omega^2} \]

The quantities \( A = x_0/\gamma_0 \) and \( \Phi \) are respectively called the amplitude ratio and the phase angle. They are represented graphically in the two curves of Fig. 6-3 as functions of the angular frequency \( \omega \), which is related to the frequency \( F \), expressed in cps, by \( \omega = 2\pi F \).

It can be seen that these curves are, approximately, of equations

\[ x_0 = \frac{m}{k} \gamma_0 \quad \Phi = 0 \]

as long as the frequency of \( \gamma(t) \) is sufficiently low (static condition), that is, as long as \( \omega \ll (k/m)^{1/2} \). For higher frequencies, \( \omega \), of the order of \( (k/m)^{1/2} \), there is a resonance effect. This is characterized by a rapid change in \( \Phi \); and, if \( f < 0.7 \times 2(km)^{1/4} \), by the existence of a maximum \( A \) greater than \( m/k \). This resonance effect is sharper for smaller values of \( f \) and is characteristic of a poorly damped system.

Finally, for very high frequencies, \( \omega \gg (k/m)^{1/2} \), there is a filtering effect that is essentially due to the presence of the inertia \( m \). This is characterized by large negative values of \( \Phi \); and, especially, by a very weak amplitude response, with \( x_0 \) being proportional to \( 1/\omega^2 \) (filtering out of high frequencies).
6.1.3. Significance of These Curves. Given an accelerometer, and knowing the curves corresponding to it, it is possible to predict its behavior under all circumstances. An intuitive explanation of this important fact can be formulated as follows (a rigorous explanation will be given in Sec. 6.2). Assume any excitation $\gamma(t)$. This expression $\gamma(t)$ can be expressed by a summation of sinusoidal functions of different frequencies by a Fourier series if $\gamma(t)$ is periodic, or by a Fourier integral if $\gamma(t)$ is of finite duration. Each sinusoidal function gives a displacement $x$ of the same frequency as the function, but of magnitude and phase that are evaluated from the curves of Fig. 6-3. Since these displacements constitute the Fourier analysis of the response $x(t)$ of the accelerometer, the latter response can be obtained by a summation.

6.1.4. Frequency-response Locus. It is possible to present these curves in a more condensed form. Thus, if the accelerometer is excited
by various sinusoidal accelerations of the same amplitude $\gamma_0$, but of different frequencies, it is possible to find at each frequency a corresponding point in the plane whose polar coordinates are exactly the values of the functions represented by the preceding curves.

![Diagram of frequency-response locus of an accelerometer.]

Fig. 6-4. Frequency-response locus of an accelerometer.

The radius vector is

$$\frac{x_0}{\gamma_0} = \frac{m}{(k - m\omega^2)^2 + f^2\omega^2}^{\frac{1}{2}}$$

and the phase angle is

$$\Phi = \arctan \frac{-f\omega}{k - m\omega^2}$$

A geometric locus may therefore be defined in polar coordinates scaled in values of the angular frequency $\omega$, as in Fig. 6-4. The locus contains all the information given by the curves of Fig. 6-3, and it therefore
completely defines the accelerometer. It is termed the frequency-response locus, Nyquist locus, or transfer locus of the system.

6.1.5. Alternate Method. The frequency-response curves and the frequency-response locus can also be found by a somewhat different method. This consists in integrating the equation

\[ m \frac{d^2x}{dt^2} + f \frac{dx}{dt} + kx = m\gamma_0 \sin \omega t \]

by searching for the forced solution

\[ x = x_0 \sin (\omega t + \Phi) \]

and recalling that

\[ j \sin \omega t = \text{Im} e^{j\omega t} \]

and

\[ j \sin (\omega t + \Phi) = \text{Im} e^{j(\omega t + \Phi)} \]

Proceeding in this manner it is found that

\[ m(j\omega)^2x_0e^{j(\omega t + \Phi)} + f(j\omega)x_0e^{j(\omega t + \Phi)} + kx_0e^{j(\omega t + \Phi)} = m\gamma_0e^{j\omega t} \]

Canceling \( e^{j\omega t} \) reduces this equation to

\[ [m(j\omega)^2 + f(j\omega) + k]e^{j\Phi}x_0 = m\gamma_0 \]

from which it is found that

\[ \frac{x_0}{\gamma_0} e^{j\Phi} = \frac{m}{m(j\omega)^2 + f(j\omega) + k} \]

or

\[ A e^{j\Phi} = \frac{1}{(j\omega)^2 + \frac{f}{m}(j\omega) + \frac{k}{m}} \]

The amplitude ratio at the angular frequency \( \omega \) is the modulus of the complex number

\[ \frac{1}{(j\omega)^2 + \frac{f}{m}(j\omega) + \frac{k}{m}} \]

The phase angle \( \Phi \) is its argument, as can easily be verified from the expressions found in Sec. 6.1.2. In other words, the transfer locus can be considered to be the locus of the end point of the vector

\[ H(j\omega) = \frac{1}{(j\omega)^2 + \frac{f}{m}(j\omega) + \frac{k}{m}} = \frac{m}{k} \frac{1}{1 + \frac{f}{k}(j\omega) + \frac{m}{k}(j\omega)^2} \]

as \( \omega \) varies from 0 to \( \infty \).

6.2. TRANSFER FUNCTIONS AND LOCI

6.2.1. Fundamental Problem. If it is desired to determine the response \( x(t) \) of the accelerometer to any excitation \( \gamma(t) \), it is necessary to integrate the differential equation
\[ \frac{d^2x}{dt^2} + \frac{f}{m} \frac{dx}{dt} + \frac{k}{m} x = \gamma(t) \]

For the initial conditions
\[ x(0) = x_0 \quad \left( \frac{dx}{dt} \right)_0 = v_0 \]
it is found, using Laplace transforms, that
\[ s^2X(s) - sx_0 - v_0 + \frac{f}{m} [sX(s) - x_0] + \frac{k}{m} X(s) = \Gamma(s) \]
Hence
\[ X(s) = \frac{\Gamma(s)}{s^2 + (f/m)s + k/m} + \frac{(s + f/m)x_0 + v_0}{s^2 + (f/m)s + k/m} \]
For the very special case of the initial conditions being zero, that is
\[ x_0 = 0 \quad v_0 = 0 \]
the above equation reduces to
\[ X(s) = \frac{\Gamma(s)}{s^2 + (f/m)s + k/m} \]
The function
\[ H(s) = \frac{1}{s^2 + (f/m)s + k/m} = \frac{m}{k} \frac{1}{1 + (f/k)s + (m/k)s^2} \]
characterizes the accelerometer, and \( X(s) \) can be written in the form
\[ X(s) = H(s)\Gamma(s) \]  \hspace{1cm} (6-2)

Hence, one arrives at the fundamental conclusion that the Laplace transform of the response of the accelerometer to a particular type of excitation is obtained by multiplying the Laplace transform of the excitation by \( H(s) \). In terms of the Laplace transform, the accelerometer is an operator that multiplies by \( H(s) \).

This function \( H(s) \) is called the transfer function of the accelerometer. To show that an accelerometer transforms an acceleration \( \gamma(t) \) into a displacement \( x(t) \) (Fig. 6-5), the representation shown in Fig. 6-6, which expresses Eq. (6-2), is used.

6.2.2. Transfer Locus (Nyquist Locus). The function \( H(s) \) is an analytic function of the complex variable \( s \). It is completely characterized by its modulus and its argument for imaginary values \( s = j\omega \), that is, by the locus of the end point of the vector \( H(j\omega) \), as scaled in units of
The locus consists of two curves \( L \) and \( L' \). Both are symmetrical about the real axis with \( L \) corresponding to the positive and \( L' \) to the negative values of \( \omega \).

The locus \( L \), hereafter referred to as the transfer locus, or Nyquist locus, of the accelerometer, is nothing other than the frequency-response locus described in Secs. 6.1.4 and 6.1.5. The precise reason why it characterizes the entire system is now obvious. The approximate reasoning of Sec. 6.1.3 is but an intuitive way of expressing the fundamental relationship of Eq. (6-2). (See in Sec. 4.5.3 the relationship between the Fourier and Laplace transforms.)

### 6.2.3. Use of Dimensionless Variables

It is possible to give, by the use of dimensionless variables, a more general significance to the expression \( H(s) \) and to the locus \( L \). First, it is useful to base the ordinate scale of the amplitude-response curve not on \( x_0/\gamma_0 \) but on \( (kx_0/m\gamma_0) \), which is a dimensionless number, and thus is independent of the units chosen. Second, if one defines

\[
\omega_n = \left( \frac{k}{m} \right)^{\frac{1}{2}} \quad \xi = \frac{f}{2(km)^{\frac{1}{2}}}
\]

then \( H(j\omega) \) may be written in the form

\[
H(j\omega) = \frac{1}{\omega_n^2} \frac{1}{1 + \frac{2\xi}{\omega_n}(j\omega) + (j\omega)^2}
\]

This is an expression in which \( \omega \) is introduced only by the ratio \( \omega/\omega_n \), and this leads, consequently, to writing \( \omega/\omega_n = u \), a nondimensional frequency.

Hence, using nondimensional variables, it is found that

\[
H(ju) = \frac{1}{1 + 2\xi(ju) + (ju)^2}
\]

which is an expression that is independent of \( \omega_n \).

The quantity \( \omega_n \) is called the undamped natural frequency of the system, and \( \xi \) its damping ratio. The physical significance of these quantities will be given in Sec. 6.4.

If then, the locus \( L \) and the abscissas of the curves of Figs. 6-3 and 6-4 are graduated in nondimensional frequencies, the new curves obtained (for all values of \( \xi \)) are applicable to all second-order systems (Figs. 6-7 and 6-8).

### 6.2.4. Logarithmic Coordinates

For convenience (Sec. 8.2.1), logarithmic coordinates are often used. Specifically, on a logarithmic scale the abscissas represent frequencies, while the ordinates represent:

- a. The phase on a linear scale.
- b. The amplitude on a logarithmic scale; or, which amounts to the same thing, decibels on a linear scale. \( A \) is expressed in decibels by the equation

\[
A \text{ db} = 20 \log A
\]
Fig. 6-7. (a) Amplitude response of second-order systems. (b) Phase response of second-order systems.

The curves obtained in this way are shown in Fig. 6-9a and b. However, in view of their great importance in the subsequent work, these curves have been redrawn to a doubled scale (1 cm for $10^\circ$ and 2 db) and are displayed on the Charts 1 and 2 at the back of the book.

Particular attention should be paid to the following:

1. **Amplitude-response Curves.** There are two asymptotes, and they intersect at the point whose ordinate is $A = 1$ (that is, 0 db) and whose
absissa is \( u = 1 \). One asymptote is horizontal, the other is oblique and has a slope which shows that, at high frequencies, the filtering is proportional to \( 1/\omega^2 \). The latter asymptote is said to have a slope of 12 db per octave. In fact, if

\[
\omega_1 = 2\omega_2 \text{ (difference of one octave)}
\]

then

\[
A(\omega_2) - A(\omega_1) = 20 \log \frac{A(\omega_2)}{A(\omega_1)} = 40 \log \frac{\omega_1}{\omega_2} = 40 \log 2 = 12 \text{ db}
\]

2. Phase-response Curves. The phase is negative and decreases from 0 to \(-180^\circ\) (phase lag increasing from 0 to \(180^\circ\)) with a sharp decrease occurring around \( \omega = 1 \) which increases in sharpness as \( \zeta \) decreases.

![Fig. 6-8. Transfer loci of second-order systems (Nyquist plots).](image)

The angular frequency \( \omega = 1/T \), that is \( u = 1 \), is sometimes called the break frequency, because it is at this frequency that the asymptotes intersect and the phase varies most rapidly. The point given by the coordinates \( u = 1, \Phi = -90^\circ \) is, for the curve, an inflection point and a center of symmetry.

6.2.5. Nichols Locus, or Amplitude-Phase Plot. The transfer locus can be drawn in a system of logarithmic coordinates to form the Nichols locus, or amplitude-phase plot, in which are represented abscissas \( \Phi \) and
Fig. 6-7. (a) Amplitude response of second-order systems. (b) Phase response of second-order systems.

The curves obtained in this way are shown in Fig. 6-9a and b. However, in view of their great importance in the subsequent work, these curves have been redrawn to a doubled scale (1 cm for 10° and 2 db) and are displayed on the Charts 1 and 2 at the back of the book.

Particular attention should be paid to the following:

1. Amplitude-response Curves. There are two asymptotes, and they intersect at the point whose ordinate is \( A = 1 \) (that is, 0 db) and whose
SECONDO-ORDER SYSTEMS

abscissa is \( u = 1 \). One asymptote is horizontal, the other is oblique and has a slope which shows that, at high frequencies, the filtering is proportional to \( 1/\omega^2 \). The latter asymptote is said to have a slope of 12 db per octave. In fact, if

\[
\omega_1 = 2\omega_2 \text{ (difference of one octave)}
\]

then

\[
A(\omega_2) - A(\omega_1) = 20 \log \frac{A(\omega_2)}{A(\omega_1)} = 40 \log \frac{\omega_1}{\omega_2} = 40 \log 2 = 12 \text{ db}
\]

2. Phase-response Curves. The phase is negative and decreases from 0 to \(-180^\circ\) (phase lag increasing from 0 to \(180^\circ\)) with a sharp decrease occurring around \( \omega = 1 \) which increases in sharpness as \( \zeta \) decreases.

![Fig. 6-8. Transfer loci of second-order systems (Nyquist plots).](image)

The angular frequency \( \omega = 1/T \), that is \( u = 1 \), is sometimes called the break frequency, because it is at this frequency that the asymptotes intersect and the phase varies most rapidly. The point given by the coordinates \( u = 1, \Phi = -90^\circ \) is, for the curve, an inflection point and a center of symmetry.

6.2.5. Nichols Locus, or Amplitude-Phase Plot. The transfer locus can be drawn in a system of logarithmic coordinates to form the Nichols locus, or amplitude-phase plot, in which are represented abscissas \( \Phi \) and
Amplitude vs. frequency for first and second-order systems

1st order (in discontinuous line): \( A = \frac{1}{\sqrt{1+u^2}} \)

2nd order: \( A = \frac{1}{\sqrt{(1-u^2)^2 + (2u)^2}} \)

Fig. 6-9a. Magnitude response of second-order systems in logarithmic coordinates (twofold reduction of Chart 1 at the back of the book).
Phase vs. frequency response
of first and second-order systems
1st order (discontinuous): \( \Phi = -\tan^{-1} u \)
2nd order: \( \Phi = -\tan^{-1} \frac{2u}{1-u^2} \)

Fig. 6-9b. Phase response of second-order systems in logarithmic coordinates (twofold reduction of Chart 2 at the back of the book).
Fig. 6-9c. Nichols loci for \(1/[1 + 2j\alpha + (j\omega)^2]\). In dotted line locus for \(1/(1 + j\alpha)\). Scale: abscissas 1 cm for 20°; ordinates 1 cm for 4 db. Loci are scaled in \(u\).

ordinates \(A\), in decibels. This is equivalent to drawing the locus of the endpoint of \(20N \ln H(j\omega)\), where \(N = 0.4343\) is the modulus of the Naperian logarithms and \(\ln\) designates the complex logarithmic function.

† Strictly speaking, it is actually the complex logarithm of the conjugate of \(H(j\omega)\) that is taken, because the positive direction is usually being chosen for the \(\Phi\) axis.
These loci are shown in Fig. 6-9c, where 1 cm represents either 20° or 4 db.

6.2.6. Inverse Transfer Locus. It is sometimes convenient, especially when dealing with feedback systems, to consider not the transfer locus \( L \), which is the locus of the end point of \( H(j\omega) \), but the inverse transfer locus, which is the locus of the end point of \( 1/H(j\omega) \).

This inverse locus can be obtained from the transfer locus \( L \) by a direct inversion about the origin followed by mirror reflection with respect to the real axis. The inverse transfer loci of second-order systems are shown, graduated in units of the nondimensional frequency \( u \), in Fig. 6-10. These loci are parabolas in which the point at infinity corresponds to zero frequency. An inverse Nichols locus can be considered in a similar manner.

6.2.7. Conclusion. It is important to understand fully and to remember that a second-order system is completely specified by any one of the following conditions:

a. The quantities \( \zeta \) and \( \omega_n \) (and knowledge of the fact that a second-order system is involved)

b. The frequency-response curves: \( A \) and \( \Phi \) as functions of \( \omega \)

c. The transfer locus, as either a Nyquist plot or a Nichols plot

6.3. TRANSIENT RESPONSE

6.3.1. Computing \( x(t) \) for a Step Input. The response of an accelerometer to a step function of acceleration which is written, for reasons of homogeneity, as
\[ \gamma(t) = \omega_n^2 u(t) \]

is obtained by integrating

\[ \frac{d^2 x}{dt^2} + \frac{f}{m} \frac{dx}{dt} + \frac{k}{m} x(t) = \omega_n^2 u(t) \]

that is,

\[ \frac{d^2 x}{dt^2} + 2\zeta \omega_n \frac{dx}{dt} + \omega_n^2 x(t) = \omega_n^2 u(t) \]

Assuming that the accelerometer is at rest at the time zero, it is found that

\[ X(s) = \frac{1}{s} \frac{\omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2} \]

or

\[ X(s) = \frac{1}{s} \frac{1}{(\frac{s}{\omega_n})^2 + 2\zeta \left(\frac{s}{\omega_n}\right) + 1} \]

By expanding in partial fractions and taking inverse Laplace transforms, \( x(t) \) can be derived from the above equation.

**6.3.2. Remarks.** Application of the final-value theorem (Sec. 4.2.3) to \( x(t) \) immediately shows that \( x(+\infty) = 1 \), while application of the initial-value theorem to \( x(t) \) shows that \( (dx/dt)(0) = 0 \).

The reason for the latter effect is the presence of the \( s^2 \) term in the denominator of the equation. This term is actually due to the inertia of the system and follows from the definition of inertia itself.

**6.3.3. Discussion.** 1. The transient behavior, depicted in Fig. 6-11, is essentially dependent on the damping. That this is true is apparent from the expression \( X(s) \); for the mathematical nature of \( x(t) \) depends on whether the poles of \( X(s) \) are real or imaginary. The discussion which follows is classic, and it is presented here only as a review.

a. If \( \zeta < 1 \), then

\[ x = 1 + Ae^{-\xi \omega_n t} \sin \left[ \omega_n (1 - \zeta)^{1/2} t + \varphi \right] \]

\[ \tan \varphi = \frac{(1 - \zeta^2)^{1/2}}{\zeta} \quad \cos \varphi = -\zeta \quad A = \frac{-1}{\sin \varphi} = (1 - \zeta^2)^{-1/2} \]

This is an oscillatory response and consists of oscillations that have a frequency \( \omega_n (1 - \zeta^2)^{1/2} \) and amplitude that is damped by the exponential term \( e^{-\xi \omega_n t} \). From Fig. 6-11 it is seen that the oscillations are barely perceptible when \( \zeta > 0.7 \).

b. If \( \zeta > 1 \), then

\[ x = 1 + Ae^{-at} + Be^{-bt} \]

where \( -a \) and \( -b \) are the roots of \( s^2 + 2\zeta \omega_n s + \omega_n^2 = 0 \) and

\[ A = \frac{b}{a - b} \quad B = \frac{-a}{a - b} \quad a, b > 0 \]
This is a nonperiodic response and is the sum of two exponentials. The quantities \(1/a\) and \(1/b\) are sometimes called the two time constants of the system and are generally written as \(T_1\) and \(T_2\).

When \(\xi\) is appreciably greater than 1, one of the two exponentials decreases much more quickly than the other. This exponential characterizes the initial acceleration of the system. Once it has disappeared, the response is similar to that of a first-order system.

\[
x(t) = 1 - e^{-\omega_n t}(1 + \omega_n t)
\]

2. The behavior of these transients is represented by a function of nondimensional time in Fig. 6-11. It will be referred to in some detail later on. At this time, it need only be noted that two systems of the same undamped natural frequency \(\omega_n\) may exhibit different transient behavior and that it is only the damping ratio \(\xi\) which characterizes the general form of the transient response.

The duration of a transient depends essentially only on the natural frequency \(\omega_n\), which thus characterizes the speed of the response. The time \(t\) is actually introduced through the product \(\omega_n t\). Hence, two accelerometers with the same damping, but of natural frequencies \(\omega_{n_1}\) and \(\omega_{n_2}\), will have transients of the same form but of durations proportional to \(2\pi/\omega_{n_1}\) and \(2\pi/\omega_{n_2}\).
6.3.4. Response to a Ramp Input. The response to a ramp input $\gamma(t) = atu(t)$, as given in terms of its Laplace transform, is

$$X(s) = \frac{a}{s^2 s^2 + 2\xi \omega_n s + \omega_n^2}$$

For the case in which $\omega_n$ is taken as unity (second-order system with unity gain, see Sec. 6.4.2), the presence of a steady-state error (velocity error) inversely proportional to $2\xi/\omega_n$ is to be noted. The transient exhibits oscillations under the same conditions as outlined for the step response.

6.4. PERFORMANCE OF SECOND-ORDER SYSTEMS

6.4.1. General. Second-order systems are characterized by a differential equation of the form

$$B_2 \frac{d^2 r}{dt^2} + B_1 \frac{dr}{dt} + B_0 r(t) = A_0 e(t)$$

where $e(t)$ is the input and $r(t)$ the response.\(^1\)

In the case of an accelerometer (Sec. 6.1.1) we have

$$e(t) = \gamma \quad r(t) = x$$

$$B_0 = k \quad B_1 = f \quad B_2 = A_0 = m$$

In the case of the RLC circuit shown in Fig. 2-11a, with equation

$$L_1 \frac{d}{dt} I + R_1 I + \frac{1}{C_1} \int I \, dt = E$$

we have

$$e(t) = E \quad r(t) = \int I \, dt$$

$$B_0 = \frac{1}{C_1} \quad B_1 = R_1 \quad B_2 = L_1 \quad A_0 = 1$$

The undamped natural frequency is

$$\omega_n = \left(\frac{B_2}{B_1}\right)^{\frac{1}{2}} \text{ sec}^{-1}$$

that is, $(k/m)^{\frac{1}{2}}$ for the accelerometer and $(1/L_1 C_1)^{\frac{1}{2}}$ for the electric circuit.

\(^1\) Systems governed by an equation of the form

$$B_2 \frac{d^2 r}{dt^2} + B_1 \frac{dr}{dt} + B_0 r(t) = A_0 e + A_1 \frac{de}{dt}$$

are also second-order systems. They can be grouped with those governed by the former equation by putting

$$e(t) + \frac{A_1}{A_0} \frac{de}{dt} = e_1(t)$$

6.4.2. Gain. The output/input ratio for the static condition \((\omega = 0)\) is termed the static gain or, more briefly, the gain of the system. The gain \(K\) is given by

\[
K = \frac{A_0}{\bar{v}}
\]

Thus, for an accelerometer the gain is \(m/k\) (it happens to be the inverse of the squared undamped natural frequency). For the electric circuit with \(\int I\, dt\), that is, the charge of the capacitor considered as output, the gain is \(C\).

6.4.3. Damping. 1. Damping Ratio. The damping is defined by the dimensionless coefficient

\[
\zeta = \frac{B_1}{2(B_eB_2)^{\frac{1}{2}}}
\]

which is usually called the damping ratio.

For the case of the accelerometer the damping ratio is \(f/2(km)^{\frac{1}{2}}\). For the case of the electric circuit it is \(R_1\).

For \(\zeta = 0\) the system is undamped, or just oscillatory (a purely theoretical case), and

\[
B_2 \frac{d^2r}{dt^2} + B_0 r(t) = e(t)
\]

For \(\zeta = 1\), the system is critically damped
For \(\zeta \gg 1\), the system is heavily damped

Typical values of damping ratio for an accelerometer are from 0.5 to 0.1. The value 0.7 (often referred to as 70 per cent of critical damping) is usually considered the optimum value, for reasons which will become clearer later on.

2. Damping and Frequency Response. Figures 6-8 and 6-9c show that the form of the frequency-response locus depends\(^1\) on the damping coefficient \(\zeta\). For a heavily damped system \((\zeta > 2\) or 3) the transfer locus approaches a semicircle (Fig. 6-12) and the characteristics of the second-order system approach those of a first-order system. That the system is actually of second order becomes apparent only from the fact that, at very high frequencies, the portion of the transfer locus that is nearest to the origin is always tangent to the negative real axis.

On the other hand, when \(\zeta < 0.7\) there is a resonant effect; that is, for a certain range of frequencies the amplitude of the response is greater than in the static case. This effect becomes more pronounced for smaller values of \(\zeta\) and is reflected (Fig. 6-13) in:

\(^1\) Its graduation, however, depends on the natural frequency of the instrument.
a. The Nyquist plot by the existence of a point whose distance from the origin is greater than that at the zero-frequency point
b. The Nichols plot by the existence of a peak value
c. The frequency-response curves by an amplitude ratio greater than 1

The maximum value of the amplitude ratio is termed the *resonance ratio* $Q$, or sometimes the peak value of magnification $M_p$. The frequency at which this occurs is termed the resonant frequency $\omega_R/2\pi$ (Fig. 6-13). The resonant frequency is given in terms of the natural frequency $\omega_n = (k/m)^{1/2}$ by $\omega_R = \omega_n(1 - 2\zeta^2)^{1/2}$. The resonance ratio is related to $\zeta$ by the function represented graphically in Fig. 6-14. With regard to orders of magnitude, it is to be noted that a $Q$ of 1.3 corresponds to a damping ratio between 0.4 and 0.5.

3. *Damping and Transient*. The value of the damping coefficient characterizes the transient behavior.\(^1\) It has been shown that this may or may not be oscillatory; depending on whether or not $\zeta$ is smaller or greater than the critical value $\zeta = 1$.

In a highly damped system ($\zeta > 2$ or 3), the response to a step function is approximately that of a first-order system. The only difference is in the neighborhood of $t = 0$ (Sec. 6.3.2), where the presence of inertia causes the velocity to be zero, thus creating an essential difference between this system and a first-order system. In a lightly damped system the response is oscillatory. There is an overshoot which becomes more pronounced as $\zeta$ becomes smaller (Fig. 6-15). Figure 6-16 displays the overshoot as a function of $\zeta$.

Regarding orders of magnitude, it should be noted that, for $0.7 < \zeta < 1$,

\(^1\) The value of the natural frequency characterizes its time scale.
the oscillations are hardly perceptible, even though the response is oscillatory from the mathematical point of view.

6.4.4. Resonant Frequency and Natural Frequency. 1. Resonant Frequency. The resonant frequency of an oscillatory second-order system is the frequency \( F_R = \frac{\omega_R}{2\pi} \) at which the modulus of the transfer function is a maximum. For this condition

\[
\omega_R = \omega_n (1 - 2\zeta^2)^{1/4}
\]

a relationship in which \( \omega_n = (k/m)^{1/2} \) is the natural undamped frequency of the system.

The frequency \( \omega_R \) is read directly from the transfer locus of the system; it is the frequency of the point situated farthest from the origin (Figs. 6-8 and 6-13a). In Nichols coordinates it is the point with the greatest ordinate (Figs. 6-9c and 6-13b).

2. Natural Frequency. In an oscillatory second-order system the frequency \( F_p = \frac{\omega_p}{2\pi} \) of the oscillations of the transient response is called the natural frequency of the system. For this condition

\[
\omega_p = \omega_n (1 - \zeta^2)^{1/2}
\]

where \( \omega_n \) is the undamped natural frequency of the system.
3. Undamped Natural Frequency. This is the frequency \( F_n = \omega_n/2\pi \) for which \( \omega_n = (k/m)^{1/2} \) and the phase shift is 90° (Fig. 6-7b or 6-9b). It is, therefore, the frequency at which the transfer locus intersects the imaginary axis (Fig. 6-8).

If \( \zeta = 0 \), then \( \omega_R = \omega_p = \omega_n \). It is justifiable, therefore, to call \( \omega_n \) the undamped natural frequency.

4. Relationship between These Three Frequencies. The equations

\[ \omega_R = \omega_n(1 - 2\zeta^2)^{1/2} \]

and

\[ \omega_p = \omega_n(1 - \zeta^2)^{1/2} \]

are represented graphically in Fig. 6-17.

It is apparent from this figure that, as long as \( \zeta \) is small, these frequencies are nearly equal. Thus, for example, when \( \zeta = 0.4 \), \( \omega_R \) differs from \( \omega_n \) by less than 20 per cent.

5. Note on Terminology. The terminology of this subject is not consistent. Many writers make no distinction between the frequencies \( \omega_p/2\pi \), \( \omega_R/2\pi \), and \( \omega_n/2\pi \), which they call the natural frequency, the free frequency, the fundamental frequency, etc. This inconsistency is permissible only when the damping is small.
6. Extension to Systems of Order Greater than Two. It will be seen in the next chapter that the concept of a resonant frequency can be extended to linear systems which exhibit resonance and which are not of the second order. The resonant frequency of the latter systems is the frequency at which the modulus of the transfer function is a maximum. The transient oscillations have a frequency that is approximately the same, especially when the resonance is accentuated. For this case the natural frequency and the resonant frequency can be considered to be essentially identical.

6.4.5. Cutoff Frequency. Bandwidth. 1. General. In general, the frequency response of a second-order system can be summarized as follows: The low-frequency signals are transmitted with an amplitude ratio that is approximately unity, whereas the high-frequency signals are filtered out. In other words, a second-order system can be approximately compared to a low-pass filter, the ideal frequency response of which would be (Fig. 6-18)

\[ A = 1, \quad \phi = -k\omega \quad \text{for } \omega < \omega_c \]
\[ A = 0 \quad \text{for } \omega > \omega_c \]

where \( \omega_c \) is the cutoff frequency of the filter and the frequency range 0 to \( \omega_c \) its passband. A system with this type of characteristic does not distort a signal which contains no components with frequencies higher than \( \omega_c/2\pi \), but it does produce a time shift that is equal to the slope \( k \) of the phase-response curve (see Fig. 5-23).

2. Definition of the Cutoff Frequency. Should it be decided to represent a second-order system by a low-pass filter (which is, of course, only a very crude approximation), a question that arises is what value is to be chosen for \( \omega_c \). Most authors consider the cutoff frequency of a second-order system to be that frequency at which the amplitude of the output is 6 db below its zero-frequency value and define it as

\[ F_c = \frac{\omega_c}{2\pi} \]

Then for all values of \( \omega \) greater than \( \omega_c \) one has (Fig. 6-19)

\[ A(\omega) < A(0) - 6 \text{ db} \]

The frequency \( F_c \) is usually termed the 6-db cutoff frequency of the system. Some authors, however, use a different notation. Some term the frequency corresponding to the 3-db point the cutoff frequency, while yet others term the undamped natural frequency \((\omega_n/2\pi)\) for which the phase shift is 90° the cutoff frequency. To prevent confusion, it is neces-
sary to specify whether the cutoff frequency is that corresponding to the 6- or 3-db point or to the 90° phase-shift point.

The 6- and the 3-db cutoff frequencies are shown in Fig. 6-20, as functions of the damping ratio $\zeta$.

3. Interest. The importance of the concept of a cutoff frequency for a second-order system is due to the fact that this is a very useful way of expressing the rapidity of the system response (see below). In addition, there is the advantage that this definition is valid for an oscillatory as well as for a nonoscillatory system.

4. Generalization. The concept of cutoff frequency may be extended without difficulty to linear systems of order higher than the second.

5. Bandwidth. The cutoff frequency $F_c$ is a very important characteristic of a second-order system because it determines, to a large degree, the time scale of the system. Its physical significance can be ascertained from the following two complementary facts:

a. It characterizes the system's speed of response. Thus, for example, if two accelerometers have the same $\zeta$, but natural frequencies of 10 and 20 cps, their transient responses will be of the same shape, but the transient of the first will last twice as long as that of the second. It is, therefore, of great advantage to have the natural frequency as high as possible.
b. It characterizes the system's bandwidth. This follows from the fact that a second-order system does not transmit equally the various frequency components of a signal. It passes the signal components whose frequencies are less than its cutoff frequency and filters out those whose frequencies are greater. This system constitutes, therefore, a low-pass filter whose bandwidth is zero to $F_c$.

![Graph showing second-order system response](image)

**Fig. 6-21.** Response time $T$ of second-order systems to within ±5 per cent. Abscissas: damping ratio $\zeta$. Ordinates, nondimensional response time to 5 per cent: $T\omega_n/2\pi$. (After C. S. Draper, W. McKay, and S. Lees, "Instrument Engineering," vol. 2, p. 265, McGraw-Hill, New York, 1953.)
It is now apparent that, since any input applied to a system may be resolved into sinusoidal components of various frequencies, the system will transmit only those components of the input whose frequencies are smaller than its cutoff frequency. Therefore, if the input to the system is of such a nature that it contains few components whose frequencies are greater than the cutoff frequency, the system will very accurately reproduce the input. If, on the contrary, the input to the system is of such a nature that it contains many components whose frequencies are greater than the cutoff frequency of the system, the latter will reproduce the input only in a very distorted manner. This is due to the attenuation by the system of all those components of the input whose frequencies are greater than \( F_c \).

It may now be said that the natural frequency of the system limits the input speeds that can be accurately reproduced. Therefore, if it is desired that the system accurately follow rapidly varying inputs, it is necessary that the system have a large bandwidth, or correspondingly, a high natural frequency.

Finally, it is important to note that it is impossible to overemphasize the fact that \textit{the speed of response and the bandwidth of a system are fundamentally identical}. In particular, it is absolutely necessary to understand that a study of the frequency response of a system is just another way of evaluating its speed of response.

6.4.6. Response Time. In general, the response time of a system is defined as the time after which the response of the system to a step input remains between 95 and 105 per cent of its final steady-state value.

It is obvious that the response time, which determines the duration of the transient, is essentially a function of the undamped natural frequency to which it is inversely proportional. For a given undamped natural frequency \( \omega_n/2\pi \), the response time is a function of the damping of the system. Thus, for very lightly damped systems the transient persists for a long time owing to the very slow decrease in the amplitude of oscillation, while for a very heavily damped system it persists for a long time owing to the very sluggish start of the system. A plot is given in Fig. 6-21 of the response time as a function of damping, from which it is seen that the response time is a minimum for a value of damping ratio of approximately 0.7. It should also be noted that sometimes a response time to an accuracy of 2 or 1 per cent is considered.

6.4.7. Orders of Magnitude. In general, for accelerometers, it is preferable to use a value of damping that is near the so-called optimum value of \( \zeta = 0.7 \). There are two reasons for this. First, the corresponding response time is small (Fig. 6-21).
Second, the amplitude ratio is comparable to its static value for a relatively wide range of frequencies (mathematically, the amplitude-vs.-frequency curve is oscillatory to the horizontal line that corresponds to its static value).

So far as choice of the natural frequency is concerned, it depends essentially on the time scale of the phenomena to be studied. It is desirable that the amplitude and phase distortion be small. This usually leads to choosing a natural frequency 3 or 4 octaves above the dominant frequencies of the input (Fig. 6-22): about 15 to 25 cps for mechanical phenomena, about 150 to 200 cps for vibration measurements.

6.4.8. Problem. Determine the transfer function of a second-order system from its step response.

Solution.
1. The static gain is obtained by inspection of the steady state.
2. The undamped natural frequency and the damping ratio, or the two time constants, are obtained from the relations between the frequency response and the transient response.
3. The coefficients of $s$ in the transfer function are then calculated without difficulty.

Fig. 6-23. Second-order systems: determining the transfer function from the step response; input is $1 - u(t)$, initial condition is $dr/dt = 0$. (After C. S. Draper, W. McKay, and S. Lees, "Instrument Engineering," vol. 2, p. 265, McGraw-Hill, New York, 1953.)
In practice:

a. If the response exhibits oscillations, the damping ratio $\zeta$ is readily obtained from the overshoot by means of Fig. 6-16. The undamped natural frequency is then obtained from the period of the oscillations $T$ by

$$\omega_n = \frac{2\pi}{T(1 - \zeta^2)^{1/2}}$$

b. Whether or not the response is oscillatory, the method outlined by Draper, McKay, and Lees\(^1\) can be applied step by step (Figs. 6-23 to 6-25). In all cases it is seen that a single test, the response to a step input, enables complete characterization of the system.

6.5. THE POLES OF THE TRANSFER FUNCTION, AS LOCATED IN THE COMPLEX PLANE

6.5.1. General. The expression for the transfer function of a second-order system was found to be

\[ H(s) = \frac{K}{1 + (2\zeta/\omega_n)s + s^2/\omega_n^2} = \frac{k}{s^2 + 2\zeta\omega_n s + \omega_n^2} \]

where \( K \) is the static gain (\( K = 1/\omega_n^2 \) for an accelerometer) and \( k \) is a quantity that arises when \( H(s) \) is expressed in the alternative form (\( k = 1 \) for an accelerometer). For more detail the reader is referred to Sec. 7-15.

It was also shown that the transient response of the system is primarily dependent on whether the zeros of the denominator, i.e., the poles of \( H(s) \), are real (\( \zeta > 1 \)) or complex (\( \zeta < 1 \)). This problem will now be examined in detail.

6.5.2. Case 1. \( \zeta > 1 \). The function \( H(s) \) has the two real poles

\[ p_1 = -\zeta\omega_n - \omega_n(\zeta^2 - 1)^{1/2} \quad \text{and} \quad p_2 = -\zeta\omega_n + \omega_n(\zeta^2 - 1)^{1/2} \]

In the Laplace-plane, or Laplace domain (the complex plane of the variable \( s \)), these roots lie on the negative real axis, as shown in Fig. 6-26. The transfer function can be written in terms of \( p_1 \) and \( p_2 \) as follows:

\[ H(s) = \frac{k}{(s - p_1)(s - p_2)} = \frac{k}{p_1 - p_2} \left( \frac{1}{s - p_1} - \frac{1}{s - p_2} \right) \]
The expression for the unit-impulse response of the system is

\[ h(t) = \mathcal{L}^{-1} H(s) = \frac{k}{p_1 - p_2} (e^{p_1 t} - e^{p_2 t}) \]

and the expression for the unit-step response is

\[ q(t) = \mathcal{L}^{-1} \frac{H(s)}{s} = \frac{k}{p_1 p_2} + \frac{k}{p_1(p_1 - p_2)} e^{p_1 t} + \frac{k}{p_2(p_2 - p_1)} e^{p_2 t} \]

These expressions contain the two exponentials \( e^{-t/T_1} \) and \( e^{-t/T_2} \), whose time constants are respectively

\[ T_1 = -\frac{1}{p_1}, \quad T_2 = -\frac{1}{p_2} \]

Thus, the location of the poles \( p_1 \) and \( p_2 \) in the Laplace plane immediately gives information regarding the transient response of the system.

For example, if \( T_2 > T_1 \), that is, if the pole \( p_2 \) lies nearer the origin than \( p_1 \), it is possible to state that after \( t = 3T_2 \) the system will be within 5 per cent of its steady-state value. This consideration enables an approximate estimation to be made of the duration of the transient.

Furthermore, if \( T_2 \) is substantially greater than \( T_1 \), the term \( e^{-t/T_1} \) quickly vanishes. As soon as the term has become negligible (say, for \( t > 3T_1 \)), the system behaves like a first-order system with a time constant \( T_2 \). This occurs with systems whose damping ratio is very high (\( \zeta > 1 \)).

**6.5.3. Case 2. \( \zeta < 1 \).** The function \( H(s) \) has the two complex poles

\[ p_1 = -\zeta \omega_n + j\omega_n(1 - \zeta^2)^{1/2} \quad \text{and} \quad p_2 = -\zeta \omega_n - j\omega_n(1 - \zeta^2)^{1/2} \]

that are complex conjugates. The corresponding points are symmetric with respect to the real axis and are located in the left half plane; that is, the real parts are negative \( -\zeta \omega_n \). Their location is shown in Fig. 6-27,
which provides a geometrical interpretation of $\zeta$ and $\omega_n$. In fact if $p_1$ is written as $p_1 = \alpha + j\beta$, the magnitudes of $\alpha$ and $\beta$ are given by

$$|\alpha| = \zeta \omega_n \quad \text{and} \quad |\beta| = \omega_n (1 - \zeta^2)^{1/2} = \omega_p$$

The resulting magnitude of the vector $OP_1$ is

$$OP_1 = (\alpha^2 + \beta^2)^{1/2} = \omega_n$$

and the angle that this vector makes with the imaginary axis is given by

$$\sin \psi = \frac{\alpha}{(\alpha^2 + \beta^2)^{1/2}} = \zeta$$

The relations

$$OP_1 = OP_2 = \omega_n \quad \sin \psi = \zeta$$

enable one to find the fundamental characteristics of the system (that is, $\zeta$ and $\omega_n$) directly from its pole configuration. Note that $\beta$ is the natural frequency $\omega_p$ defined in Sec. 6.4.4. The pair of poles $p_1$ and $p_2$ is termed an oscillatory mode of the system, characterized by the corresponding values of $\alpha$ and $\beta$ (or of $\zeta$ and $\omega_n$).

The transient response of the system will be of the form

$$A e^{-\alpha t} \sin (\beta t + \phi) = A e^{-\omega_p t} \sin [\omega_n (1 - \zeta^2)^{1/2} t + \phi]$$

which is a sine wave with a period $2\pi/\beta$ and an amplitude that is damped by an exponential having a time constant $1/|\alpha|$. Thus, the poles lying close to the imaginary axis (Fig. 6-28a) correspond to an oscillatory transient ($\zeta \ll 1, |\alpha| \ll |\beta|$).

The value $\psi = 45^\circ$ (that is, $|\alpha| = |\beta|$) corresponds to $\zeta = 0.7$. In this case (Fig. 6-28b) the exponential has decayed to 2 per cent of its initial value, after a time $t = 4/|\alpha|$, that is, after two-thirds of the period of the sinusoid. This is the reason for the lack of physical oscillations in practical systems.

When $\psi > 45^\circ$, the points $P_1$ and $P_2$ move toward the real axis. For this case, $\zeta > 0.7$, there are again no oscillations in the transient response of a practical system.

6.5.4. Case 3. $\zeta = 1$. This is the "critical case," in which the expression for the transient solution involves the term $te^{-\omega_p t}$. Physically this is nearly equivalent to the case in which $p_1$ and $p_2$ are very close
together, being either real or imaginary conjugates. Owing to the fact
that the coefficients of the equations for physical systems are never
exactly known, the case of two equal poles is purely mathematical.

6.6. EXAMPLES OF SECOND-ORDER SYSTEMS. THEIR
IMPORTANCE

6.6.1. Examples. Examples of second-order systems are innumerable.
It is impossible to quote all the systems for which the equations are of the
form

\[ m \frac{d^2 r}{dt^2} + f \frac{dr}{dt} + kr = e(t) \]

The behavior of all such systems is characterized by the following
quantities or their analogs:

1. An inertia or moment of inertia
2. A viscous friction developing a force (or torque) proportional to
   velocity (or angular velocity)
3. A proportional restraint developing a force (or torque) proportional
to \( r \)

Some examples are:

a. Angular accelerometers, with oil or electromagnetic damping.
b. The classical RLC circuits used in electrical engineering.
c. Recording oscillographs (their natural frequency is usually high, up to 250 cps).
d. Rate gyros, for measuring or stabilizing purposes; they may be spring or electrically
   restrained. Their natural frequency ranges from 5 cps (airplanes) to 12 cps
   (missiles); the damping ratio is usually 0.4 or 0.5.
e. Rotating amplifiers (hydraulic or Ward-Leonard) if the angular velocity of the
   output shaft is considered as the output.

6.6.2. Characteristic Properties. All second-order systems have the
following properties:

1. For specific values of their parameters they can exhibit oscillations.
   This is a characteristic that is new compared to first-order systems and is
   essentially due to the presence of inertia. Oscillations correspond to
   exchange of energy between the potential energy (\( \frac{1}{2} k x^2 \)) and kinetic
   energy (\( \frac{1}{2} M v^2 \) or \( \frac{1}{2} I \omega^2 \)) or their analogs.
2. The oscillations are always damped because damping is essentially
   positive; that is, energy can only be dissipated. Therefore, a second-
   order system is essentially stable.

6.6.3. Immediate Interest. The study of second-order systems is
important because a very large number of physical systems can be
described in satisfactory manner by second-order differential equations.
However, any physical systems that are liable to display instability
cannot be described in terms of second-order systems. Examples of
this are servomechanisms, as will be shown in Part 2.

6.6.4. General Interest. Second-order systems are, moreover, impor-
tant because concepts concerning them are often extended to higher-order
systems. In particular:
1. The resonant frequency of a higher-order system is the frequency at which the transfer function is a maximum and is a generalization of $\omega_n(1 - 2\zeta^2)^{1/4}$. The natural frequency is also sometimes considered. It is the frequency of transient oscillations and generalizes $\omega_n(1 - \zeta^2)^{1/4}$. If the resonance effect is marked, both frequencies are approximately equal.

2. For a second-order system, the resonance ratio $Q$ is directly related to its damping ratio. For higher-order systems, the equivalent of a damping ratio is defined by considering the flatness of the resonance peak characterized by its resonance ratio $Q$.

Obviously, such generalizations are only possible for systems that are approximately "similar" to second-order systems. Systems, for instance, that exhibit two marked resonance peaks, cannot be approximated as second-order systems.¹

¹ If the frequencies at which the resonances occur are sufficiently separated, it is usually possible to approximate the system by a "product" of two second-order systems. A noteworthy example is the case of airplanes in longitudinal motion. In this case, the lower resonant frequency corresponds to the phugoid oscillations of the center of gravity; the higher resonance frequency corresponds to oscillations of the aircraft around its center of gravity.
CHAPTER 7
TRANSFER FUNCTIONS

Summary
1. The concept of transfer function.
2. Application to steady state.
3. Application to transient.
4. Philosophy of transfer functions. The two practical approaches.
5. Determination of transfer functions.

The purpose of this chapter is to present the concept of transfer function for linear systems in general. This concept, which is the core of the whole first part of the book, consists of characterizing a linear system by an analytical function of the Laplace variable $s$. It will be shown in this chapter what this function is and how it can be determined for a given system. Later, two techniques that will enable practical manipulation of the transfer-function concept with a special view to synthesis will be outlined. These two techniques are the harmonic approach (Chap. 8) and the pole-zero-configuration approach (Chap. 9).

7.1. THE CONCEPT OF TRANSFER FUNCTION

7.1.1. Linear Systems. Let $S$ be any system (mechanical, electrical, hydraulic, etc.). It is influenced by a certain number of inputs, and

![Fig. 7-1. General diagram.](image)

yields an output, or response, which is a function of the inputs and of the system. Often in the operation of the system, one of the inputs is of predominant importance. If so, it is called the input (or the main input) to the system, and the other inputs are referred to as secondary inputs. There may be two main inputs, e.g., a command, or control input, and a disturbance (Sec. 1.2.4). The system $S$ can be represented by the schematic diagram shown in Fig. 7-1. The only assumption concerning this system is that the input and the response are related by a linear differential equation with constant coefficients which can be written in the form\(^1\)

\[ A_{\dot{\phi}}(t) = A_{\phi}(t) + A_{\cdot\cdot\cdot} - B_{\cdot\cdot\cdot}. \]

\(^1\) It is always possible to eliminate eventual constant terms $A_{\cdot\cdot\cdot}$ and $B_{\cdot\cdot\cdot}$ by setting, for instance, $A_{\cdot\cdot\cdot}(t) = A_{\cdot\cdot\cdot}(t) + A_{\cdot\cdot\cdot} - B_{\cdot\cdot\cdot}$. 

114
\[ A_m \frac{d^me}{dt^m} + \cdots + A_1 \frac{de}{dt} + A_0 e(t) = B_n \frac{d^n r}{dt^n} + \cdots + B_1 \frac{dr}{dt} + B_0 r(t) \]

(7-1)

The system \( S \) is said to be linear. For many systems, useful results can be obtained by assuming linear operation. Furthermore, nonlinear systems can often be linearized, for example, by restricting the range of operation. The errors arising from the use of a linear analysis will be dealt with later (see Chaps. 11 and 22).

7.1.2. The Transfer Function.

Given a linear system described by Eq. (7-1), its response to an input \( e(t) \) may be determined as follows: Let it be assumed that the pertinent initial conditions are:

\[
\begin{align*}
    r(0) &= r_0 & e(0) &= e_0 \\
    \frac{dr}{dt}(0) &= r'_0 & \frac{de}{dt}(0) &= e'_0 \\
    \frac{d^{n-1}r}{dt^{n-1}}(0) &= r_0^{n-1} & \frac{d^{m-1}e}{dt^{m-1}}(0) &= e_0^{m-1}
\end{align*}
\]

Transforming (Chap. 4) yields the Laplace transform of the output (Fig. 7-2):

\[
R(s) = \frac{A_m s^m + \cdots + A_1 s + A_0}{B_n s^n + \cdots + B_1 s + B_0} E(s)
\]

(7-2)

\[
(B_n s^{n-1} + \cdots + B_0) r_0 + \cdots + B_n r_0^{n-1}
\]

\[
- (A_m s^{m-1} + \cdots + A_0) e_0 \cdots A_m e_0^{m-1}
\]

\[
\frac{B_n s^n + \cdots + B_1 s + B_0}{B_n s^n + \cdots + B_1 s + B_0}
\]

For the particular case of zero initial conditions,

\[
r_0 = 0 \quad r'_0 = 0 \quad \cdots \quad r_0^{n-1} = 0 \quad e_0 = 0 \quad \cdots \quad e_0^{m-1} = 0
\]

this reduces to

\[
R(s) = \frac{A_m s^m + \cdots + A_1 s + A_0}{B_n s^n + \cdots + B_1 s + B_0} E(s)
\]

The function

\[
\frac{A_m s^m + \cdots + A_1 s + A_0}{B_n s^n + \cdots + B_1 s + B_0} = H(s)
\]

completely characterizes the system and is called its transfer function.

The Laplace transform of the response for zero initial conditions is thus the product of the transfer function of the system and the Laplace transform of the input. The influence of the system \( S \) is treated as a multiplication operator \( H(s) \):

\[
E(s) H(s) = R(s)
\]

(7-3)
Note 1. If more than one system is under consideration, subscripts may be added to the symbol \( H \) in order to identify the system, with further subscripts to identify inputs and outputs if necessary. For example, the transfer function of the accelerometer considered in Chap. 6 may be written as \( H_{\text{act}}(\gamma, z) \).

Note 2. When a linear system has two independent inputs, there are, of course, two transfer functions to be considered (Fig. 7-3).

Note 3. There are many other accepted expressions for the term transfer function. Many electrical engineers call it the transmittance, while many aeronautical engineers, after C. S. Draper, call it the performance operator. However, the expression transfer function is, at present, most commonly used.

### 7.1.3. The Characteristic Equation

The function \( H(s) \) is, in general, the ratio of two polynomials in \( s \): \( H(s) = \frac{P(s)}{Q(s)} \). It may also include lag factors, thus

\[
H(s) = \frac{P(s)}{Q(s)} e^{-sT_r} e^{-sT_p}
\]

The equation \( Q(s) = 0 \) is called the characteristic equation of the system. The roots of the characteristic equation are the zeros of \( Q(s) \), which are also the poles of \( H(s) \). It will be seen later that they characterize the stability of the system.

### 7.1.4. Integrations

When the transfer function of a system \( S \) has a single pole equal to zero, it can be written in the form

\[
H(s) = \frac{P(s)}{sQ_1(s)}
\]

where \( Q_1(0) \neq 0 \). The factor \( 1/s \) is an integrating factor (Sec. 4.4.3). Correspondingly, it is said that the transfer function \( H(s) \), thus the system itself, encompasses an integration.

If the zero pole for the transfer function is a double pole, then

\[
H(s) = \frac{P(s)}{s^2Q_1(s)}
\]

where again \( Q_1(0) \neq 0 \). In this case the system is said to have double integration; and so on for higher-order poles.

The presence of integrations in the transfer function of a system is of the greatest importance in the study of its steady-state behavior.

### 7.1.5. Gain

Transfer functions are usually written in the form \( H(s) = KF(s) \). The constant \( K \), which is termed the gain of the system, plays an important part in characterizing the static behavior of the system. The quantity \( F(s) \) is termed the frequency-dependent part of the transfer function \( H(s) \) and characterizes the dynamic behavior of the system. Two definitions are usually given for the gain of a system:

1. \( F(s) \) may be written in such a way that the constant terms (or, when integration is implied, the coefficients of the lower powers of \( s \)) in its
numerator and denominator will be equal to unity. For example,

\[
H(s) = \frac{7 + 14s}{3 + 8s + 12s^2} = \frac{7}{3} \frac{1 + 2s}{1 + \frac{8}{3}s + 4s^2}
\]

\[
H_1(s) = \frac{7 + 14s}{s(3 + 8s + 12s^2)} = \frac{7}{3} \frac{1 + 2s}{s + \frac{8}{3}s + 4s^2}
\]

The gain of these two systems is \( K = \frac{7}{3} \).

2. \( F(s) \) may also be written in such a way that the coefficients of the highest powers of \( s \) in its numerator and denominator will be unity. For the above transfer functions:

\[
H(s) = \frac{7 + 14s}{3 + 8s + 12s^2} = \frac{14}{12} \frac{s + \frac{1}{2}}{s^2 + \frac{8}{12}s + \frac{9}{12}}
\]

\[
H_1(s) = \frac{7 + 14s}{s(3 + 8s + 12s^2)} = \frac{14}{12} \frac{s + \frac{1}{2}}{s^3 + \frac{8}{12}s^2 + \frac{9}{12}s}
\]

The gain of the two systems is now \( k = \frac{14}{12} = \frac{7}{6} \).

Both definitions of gain are used in linear-system analysis, one or the other being more convenient according to the method used. In order to avoid confusion, we will always denote by \( K \) the definition given in (1) and by \( k \) the definition given in (2). To be specific, when no integration is involved:

\[
H(s) = K \frac{1 + a_1s + a_2s^2 + \cdots}{1 + b_1s + b_2s^2 + \cdots} = K \frac{(1 + \alpha_1s)(1 + \alpha_2s) \cdots}{(1 + \beta_1s)(1 + \beta_2s) \cdots}
\]

\[
H(s) = k \frac{s^m + a_{m-1}s^{m-1} + \cdots}{s^n + a_{n-1}s^{n-1} + \cdots} = k \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots}
\]

where the quantities \( z_i = -1/\alpha_i \) and \( p_i = -1/\beta_i \) are the zeros and poles of \( H \), respectively. The quantity \( K \) is termed the gain of the system, or its static gain, since \( H(0) = K \), which gives a physical interpretation for \( K \). The quantity \( k \) has no physical meaning and should preferably not be called gain.

Similarly, when integration is implied:

\[
H(s) = \frac{K 1 + a_1s + a_2s^2 + \cdots}{s 1 + b_1s + b_2s^2 + \cdots} = \frac{K (1 + \alpha_1s)(1 + \alpha_2s) \cdots}{s (1 + \beta_1s)(1 + \beta_2s) \cdots}
\]

\[
H(s) = \frac{k s^m + a_{m-1}s^{m-1} + \cdots}{s^n + a_{n-1}s^{n-1} + \cdots} = \frac{k (s - z_1)(s - z_2) \cdots}{s (s - p_1)(s - p_2) \cdots}
\]

Both \( K \) and \( k \) are called the gain of the system, but \( K \) is generally termed the velocity gain (see Sec. 7.2.1 for explanation of the term). For small values of \( s \), \( H(s) \approx K/s \).

Similarly, systems involving two integrations have an acceleration gain \( K \), and for small values of \( s \), \( H(s) \approx K/s^2 \).
More generally, if the system involves $i$ integrations, the gain $K$ is
\[ \lim_{s \to 0} s^i H(s). \]

In all cases, the quantities $k$ and $K$ are related by
\[ K = k \frac{z_1 z_2 \cdots z_m}{p_1 p_2 \cdots p_{n-i}}. \]

In particular, if all the zeros and poles have negative real parts, $k$ has
the same sign as $K$ (usually $+$), and
\[ K = k \frac{z_1 z_2 \cdots z_m}{p_1 p_2 \cdots p_{n-i}} \]

(7-4)

7.2. APPLICATION TO STEADY STATE

7.2.1. The Effect of Integrations (Fig. 7.4). Let a linear system $S$ be
subjected to a step input. The steady state is the particular integral of

\[ \frac{1}{s} \]

(a)

\[ \frac{1}{s^2} \]

(b)

\[ \frac{1}{s^3} \]

(c)

Fig. 7-4. Effect of integrations on steady state. (a) No integration. (b) One inte-
gration. (c) Two integrations.

the differential equation (7-1) when $e(t) = e_0 u(t)$. If no integration is
present, the steady state is given by $r(t) = (A_0 / B_0) e_0 = Ke_0$.

If an integration is present, $B_0$ is zero and the steady state is
\[ \frac{dr}{dt} = \left( \frac{A_0}{B_1} \right) e_0 = Ke_0 \]

\[ 1 \] This is the case for stable and minimum-phase-shift systems (see Secs. 8.5.3 and
9.1.2).
that is, the steady state is not a constant-\( r \) state, but a constant-\( dr/dt \) state. If the system is a positional system, that is, if \( e(t) \) and \( r(t) \) are linear positions or angles, the steady-state response is a constant velocity. This is why \( K \) is called the *velocity gain* of the system. This terminology has been extended to systems which are not positional systems.

If two integrations are present, \( B_0 \) and \( B_1 \) are zero and the steady state \( d^2r/dt^2 = (A_0/B_2)e_0 = K_0 \) is a constant acceleration. \( K \) is termed the *acceleration gain* of the system.

### 7.2.2. Position Error

Let a positional-control system be subjected to a step input \( e(t) = e_0u(t) \). Let it be assumed that the system is stable and does not have integration. After the transient has died away, the steady state may or may not involve an error. It is said that there is, or is not, a steady-state error for a step position input, or more briefly a position error (Fig. 3-12). The position error can be found in terms of the transfer function:

\[
\lim_{t \to \infty} r(t) = \lim_{s \to 0} \frac{1}{s} H(s) = H(0) = K
\]

It is seen that the *condition for zero position error* is that the static gain of the system be equal to unity.

It should be noticed that the expression "position error" is extended by usage to control systems which are not positional systems.

### 7.2.3. Velocity Error, Velocity Constant

Let a stable positional-control system be subjected to a ramp input \( e(t) = atu(t) \). Let it also be assumed that the system is stable, that it has no integration and that its static gain is unity. The steady state may involve an error

\[
\varepsilon(\infty) = e(\infty) - r(\infty)
\]

which is called the *steady-state error for a ramp* (or velocity-step) input, or more briefly, the *velocity error* (Fig. 7-5a, b, and c). Just as for position error, it is to be pointed out that the terminology *velocity error* is extended by usage to control systems which are not positional servomechanisms. The term *velocity error* thus expresses the steady-state error for a ramp input.

---

\(^1\) The application of the final-value theorem is allowed because of the assumption made that the system is stable (Secs. 4.2.3 and 9.1.2), thus \( H(s) \) has no poles in the right-hand half of the \( s \) plane.
It should be emphasized, in order to avoid frequent misunderstanding, that the velocity error is an error and therefore has the same dimension as the system error defined as \( e = \varepsilon - r \). For example, in a positional servo the velocity error is a deviation in position. The expression for the velocity error can be obtained from the transfer function:

\[
\lim_{t \to \infty} [e(t) - r(t)] = \lim_{s \to 0} s \left[ \frac{a}{s^2} - \frac{a}{s^2} H(s) \right] = \lim_{s \to 0} \frac{a}{s} [1 - H(s)]
\]

The velocity error is proportional to the ramp coefficient \( a \) and inversely proportional to the quantity \( C_v \), defined by

\[
\frac{1}{C_v} = \lim_{s \to 0} \frac{1 - H(s)}{s}
\]

which is called the velocity constant of the system.

7.2.4. **Generalization.** Similarly, for a stable system with unit gain and infinite velocity constant, one may consider an acceleration error, i.e., the steady-state error for an input \( e(t) = \frac{1}{2} at^2 u(t) \):

\[
\lim_{t \to \infty} [e(t) - r(t)] = \lim_{s \to 0} \frac{a}{s^2} [1 - H(s)] = \frac{a}{C_a}
\]

where \( C_a \) is the acceleration constant of the system (Fig. 3-14).

More generally, the error/input transfer function for the system

\[
\frac{E(s)}{E(s)} = \frac{E(s) - R(s)}{E(s)} = 1 - H(s)
\]

can be expanded in a series in \( s \). The coefficients \( K, C_v, \) and \( C_a \) appear in the series as follows:\(^1\)

\[
1 - H(s) = 1 - K + \frac{s}{C_v} + \frac{s^2}{C_a} + \cdots
\]

For example, in a second-order system with unit static gain

\[
H(s) = \frac{1}{1 + (2\zeta/\omega_n)s + s^2/\omega_n^2}
\]

one has

\[
\frac{1}{C_v} = \frac{2\zeta}{\omega_n}, \quad \frac{1}{C_a} = \frac{1 - 4\zeta^2}{\omega_n^2}
\]

7.2.5. **Obtaining the Velocity Constant Directly.** The inverse velocity constant \( 1/C_v \) is the coefficient of the first-degree term in the series expansion of \( 1 - H(s) \). Thus, \( 1/C_v \) can be readily found as the time constant of the first-order system that is equivalent to \( H(s) \) at very low frequencies. For a first-order system (Sec. 5.1.4),

\[ H(s) = \frac{1}{1 + Ts} \]

\[
\frac{1}{C_v} = T
\]

For a second-order system,

\[
H(s) = \frac{1}{1 + (2\xi/\omega_n)s + s^2/\omega_n^2} \approx \frac{1}{1 + (2\xi/\omega_n)s}
\]

\[
\frac{1}{C_v} = \frac{2\xi}{\omega_n}
\]

For a third-order system,

\[
H(s) = \frac{1}{1 + T_1s + T_2s^2 + T_3s^3} \approx \frac{1}{1 + (T_1 + T_3)s}
\]

\[
\frac{1}{C_v} = T_1 + T_3
\]

For a second-order system with a real zero in the numerator,

\[
H(s) = \frac{1 + T_1s}{1 + (2\xi/\omega_n)s + s^2/\omega_n^2} \approx \frac{1}{1 + (2\xi/\omega_n - T_1)s}
\]

\[
\frac{1}{C_v} = \frac{2\xi}{\omega_n - T_1}
\]

and so forth for more complicated systems.

A graphical interpretation of \( C_v \) from the frequency-response plots will be given later (Sec. 8.3.2).

7.2.6. The Importance of These Constants. When the system \( S \) is a control system, it is generally desired that the output \( r(t) \) follow the command \( e(t) \) as closely as possible. In particular, it is generally desired that the steady-state response be equal to the input; that is, that there be no steady-state error. Such a condition is ordinarily prescribed for typical inputs: step, ramp, or acceleration. Ideally, it would be desirable that a control system have (Sec. 3.4.4) a static gain equal to unity, an infinite velocity constant, and an infinite acceleration constant. In practice, it is desired that:

a. The static gain be unity, or as near to unity as possible
b. The velocity constant be as high as possible
c. The acceleration constant be high

In other words, the static gain, the velocity constant, and the acceleration constant are figures of merit for the steady-state performance of the system.

7.2.7. Relation between Transfer Function and Harmonic Response. Consider now the steady-state of the system when it is subjected to a sinusoidal or harmonic input \( e(t) = e_0 \sin \omega t \). After the transient has died away, there will result at the output a forced response which is a sinusoid of angular frequency \( \omega \): \( r(t) = Ae_0 \sin (\omega t + \Phi) \), where \( Ae_0 \) is its magnitude (the ratio \( r_0/e_0 = A \) is often called the amplitude ratio) and \( \Phi \) its phase with reference to the input.

By substituting the latter expression for \( r(t) \) into the differential equation (7-1) and by taking into account that \( j \sin k = \text{Im} e^{jk} \), one obtains

\[
A_0e_0(j\omega)^m e^{j\omega} + \cdots + A_0e_0e^{j\omega} = B_0e_0(j\omega)^n A e^{i(\omega t + \Phi)} + \cdots + B_0e_0 A e^{i(\omega t + \Phi)}
\]

\footnote{If these three conditions happen to conflict, the third is generally considered less important than the first two.}
Dividing both sides by $e^{j\omega t}$:

$$A_m(j\omega)^m + \cdots + A_0 = B_n(j\omega)^n e^{j\phi} + \cdots + B_0 A e^{j\phi}$$

$$A e^{j\phi} = \frac{A_m(j\omega)^m + \cdots + A_0}{B_n(j\omega)^n + \cdots + B_0}$$

that is $$A e^{j\phi} = H(j\omega)$$

Thus, when the system $S$ is subjected to a sinusoidal input, the forced response is characterized as follows: Its magnitude is that of the input multiplied by the magnitude of $H(j\omega)$, and its phase leads that of the input by the argument of $H(j\omega)$. The importance of this result lies in the fact that, given a system $S$, it is possible to characterize it completely by subjecting it to sinusoidal tests at various $\omega$ and plotting $A$ and $\Phi$ as functions of $\omega$. This enables one to find the transfer function of a system without having to consider the differential equation. This question will be dealt with in more detail later (Sec. 8.4.1).

7.3. APPLICATION TO TRANSIENTS

7.3.1. Transfer Function and Unit-impulse (or Unit-step) Response.

1. If the system $S$ is subjected to a unit-impulse input $e(t) = u_1(t)$, that is, $E(s) = 1$, the Laplace transform of the response will be $R(s) = H(s)$. Hence, the transfer function of a system is the Laplace transform of its response to a unit impulse.

2. If the system is now subjected to a unit-step input $e(t) = u(t)$, that is, $E(s) = 1/s$, the Laplace transform of the response will be

$$R(s) = \left(\frac{1}{s}\right) H(s)$$

Hence, the transfer function of a system is the Laplace transform of the time derivative of its unit-step response.

3. It can readily be seen from the above relation that the system's response to a unit-impulse input is the time derivative of its response to a unit step. Similarly, the response to a unit-ramp input $e(t) = tu(t)$, that is, $E(s) = 1/s^2$, is the integral of the response to a unit step. These properties are sometimes referred to as the Ludbrook theorems.

7.3.2. More on the Unit-step Response. The Laplace transform of the unit-step response is $(1/s)H(s)$. When $H(s)$ is the ratio of two polynomials $H(s) = P(s)/Q(s)$, and if $p_1, p_2, \ldots, p_n$ are the poles of $H(s)$, that is, the zeros of $Q$, then the partial-fraction expansion of $(1/s)H(s)$ yields, assuming that all poles of $H(s)$ are simple,

$$\frac{1}{s} H(s) = \frac{C_0'}{s^2} + \frac{C_0}{s} + \sum_{i=1}^{i=n} \frac{C_i}{s - p_i}$$

Hence the unit-step response is

$$h_1(t) = C_0' t + C_0 + \sum_{i=1}^{i=n} C_i e^{p_i t}$$

1 If this argument is negative, there will, of course, result a phase lag of the response with respect to the input.
The coefficients $C_i$ are given by $C_i = P(q_i)/p_iQ'(q_i)$ where $Q'(q) = dQ/ds$. As for $C_0$ and $C'_0$:

a. If $H(s)$ has no pole equal to zero,

$$ C_0 = \frac{P(0)}{Q(0)} = H(0) \quad C'_0 = 0 $$

b. If $H(s)$ has a single pole equal to zero,

$$ C_0 = \frac{d}{ds} \left( \frac{P(s)}{Q(s)} \right)_{s \to 0} \quad C'_0 = \frac{P(0)}{Q'(0)} $$

These formulas often find useful application. Finally, it is to be noted that, when $H(s)$ represents the product of such a function and a lag factor $e^{-st}$, the step response is the same, but delayed by a time $T$.

### 7.3.3. Applications of the Unit-step Response

It is often very convenient to characterize a system by its unit-step response $q(t)$. The unit-step response can be used in particular to determine the response of the system to an input which is a sum of step functions.

**Example 1.** The function $f(t)$ shown in Fig. 7-6 can be considered as the sum of a step occurring at $t = 0$ and a step of equal magnitude but opposite sign occurring at $t = T$: $f(t) = au(t) - au(t - T)$.

If $f(t)$ is the input to a system characterized by its unit-step response $q(t)$, then the response of the system to the $f(t)$ input will be $aq(t) - aq(t - T)$, or, more correctly, $aq(t)u(t) - aq(t - T)u(t - T)$.

**Example 2.** The function $g(t)$ shown in Fig. 7-7 can be considered as

$$ au(t) - 2au(t - T_1) + 2au(t - T_2) - 2au(t - T_3) + \cdots $$

When such an input actuates a system $S$ characterized by its unit-step response $q(t)$, the output will be

$$ r(t) = aq(t) - 2aq(t - T_1) + 2aq(t - T_2) - 2aq(t - T_3) + \cdots $$

where, for the sake of brevity, $q(t - T_n)$ is written instead of $q(t - T_n)u(t - T_n)$.

This formula is useful whenever a linear system is actuated by an on-off element. In particular, if $g(t)$ is periodic ($T_2 = 2T_1$, $T_3 = 3T_1$, etc.), the formula becomes

$$ r(t) = aq(t) - 2aq(t - T_1) + 2aq(t - 2T_1) - 2aq(t - 3T_1) + \cdots $$

and will be used in Part 3 in finding the periodic solutions for on-off control systems.
Dividing both sides by $e^{j\omega t}$:

$$A_m(j\omega)^m + \cdots + A_0 = B_n(j\omega)^n e^{j\omega t} + \cdots + B_0 e^{j\omega t}$$

$$A e^{j\omega t} = \frac{A_m(j\omega)^m + \cdots + A_0}{B_n(j\omega)^n + \cdots + B_0} \quad \text{that is} \quad A e^{j\omega t} = H(j\omega)$$

Thus, when the system $S$ is subjected to a sinusoidal input, the forced response is characterized as follows: Its magnitude is that of the input multiplied by the magnitude of $H(j\omega)$, and its phase leads that of the input by the argument of $H(j\omega)$. The importance of this result lies in the fact that, given a system $S$, it is possible to characterize it completely by subjecting it to sinusoidal tests at various $\omega$ and plotting $A$ and $\Phi$ as functions of $\omega$. This enables one to find the transfer function of a system without having to consider the differential equation. This question will be dealt with in more detail later (Sec. 8.4.1).

7.3. APPLICATION TO TRANSIENTS

7.3.1. Transfer Function and Unit-impulse (or Unit-step) Response.

1. If the system $S$ is subjected to a unit-impulse input $e(t) = u_1(t)$, that is, $E(s) = 1$, the Laplace transform of the response will be $R(s) = H(s)$. Hence, the transfer function of a system is the Laplace transform of its response to a unit impulse.

2. If the system is now subjected to a unit-step input $e(t) = u(t)$, that is, $E(s) = 1/s$, the Laplace transform of the response will be

$$R(s) = \left(\frac{1}{s}\right) H(s)$$

Hence, the transfer function of a system is the Laplace transform of the time derivative of its unit-step response.

3. It can readily be seen from the above relation that the system’s response to a unit-impulse input is the time derivative of its response to a unit step. Similarly, the response to a unit-ramp input $e(t) = tu(t)$, that is, $E(s) = 1/s^2$, is the integral of the response to a unit step. These properties are sometimes referred to as the Ludbrook theorems.

7.3.2. More on the Unit-step Response. The Laplace transform of the unit-step response is $(1/s)H(s)$. When $H(s)$ is the ratio of two polynomials $H(s) = P(s)/Q(s)$, and if $p_1, p_2, \ldots, p_n$ are the poles of $H(s)$, that is, the zeros of $Q$, then the partial-fraction expansion of $(1/s)H(s)$ yields, assuming that all poles of $H(s)$ are simple,

$$\frac{1}{s} H(s) = \frac{C_0'}{s^2} + \frac{C_0}{s} + \sum_{i=1}^{n} \frac{C_i}{s - p_i}$$

Hence the unit-step response is

$$h_1(t) = C'_0 t + C_0 + \sum_{i=1}^{n} C_i e^{p_i t}$$

1 If this argument is negative, there will, of course, result a phase lag of the response with respect to the input.
The coefficients \( C_i \) are given by
\[
C_i = \frac{P(p_i)}{Q'(p_i)} \quad \text{where} \quad Q'(s) = \frac{dQ}{ds}. \]  

As for \( C_0 \) and \( C'_0 \):

a. If \( H(s) \) has no pole equal to zero,
\[
C_0 = \frac{P(0)}{Q(0)} = H(0) \quad \text{and} \quad C'_0 = 0
\]

b. If \( H(s) \) has a single pole equal to zero,
\[
C_0 = \frac{d}{ds} \left( \frac{P(s)}{Q} \right)_{s \to 0} \quad \text{and} \quad C'_0 = \frac{P(0)}{Q'(0)}
\]

These formulas often find useful application. Finally, it is to be noted that, when \( H(s) \) represents the product of such a function and a lag factor \( e^{-sT} \), the step response is the same, but delayed by a time \( T \).

### 7.3.3. Applications of the Unit-step Response.

It is often very convenient to characterize a system by its unit-step response \( q(t) \). The unit-step response can be used in particular to determine the response of the system to an input which is a sum of step functions.

**Example 1.** The function \( f(t) \) shown in Fig. 7-6 can be considered as the sum of a step occurring at \( t = 0 \) and a step of equal magnitude but opposite sign occurring at \( t = T \):
\[
f(t) = au(t) - au(t - T).
\]

If \( f(t) \) is the input to a system characterized by its unit-step response \( q(t) \), then the response of the system to the \( f(t) \) input will be \( aq(t) - aq(t - T) \), or, more correctly, \( aq(t)u(t) - aq(t - T)u(t - T) \).

**Example 2.** The function \( g(t) \) shown in Fig. 7-7 can be considered as
\[
au(t) - 2au(t - T_1) + 2au(t - T_2) - 2au(t - T_3) + \cdots
\]

When such an input actuates a system \( S \) characterized by its unit-step response \( q(t) \), the output will be
\[
r(t) = aq(t) - 2aq(t - T_1) + 2aq(t - T_2) - 2aq(t - T_3) + \cdots
\]

where, for the sake of brevity, \( q(t - T_n) \) is written instead of \( q(t - T_n)u(t - T_n) \).

This formula is useful whenever a linear system is actuated by an on-off element. In particular, if \( g(t) \) is periodic \( (T_2 = 2T_1, T_3 = 3T_1, \text{etc.}) \), the formula becomes
\[
r(t) = aq(t) - 2aq(t - T_1) + 2aq(t - 2T_1) - 2aq(t - 3T_1) + \cdots
\]

and will be used in Part 3 in finding the periodic solutions for on-off control systems.
Example 3. Any input \( f(t) \) may be approximated (Fig. 7-8) by a sum of step functions which follow one another at small intervals of time

\[
e(0)u(t) \\
[e(\Delta t) - e(0)]u(t - \Delta t) \\
[e(2\Delta t) - e(\Delta t)]u(t - 2\Delta t)
\]

Because of the principle of superposition, the response is

\[
r(t) = e(0)q(t) + [e(\Delta t) - e(0)]q(t - \Delta t) + \cdots = \sum_{0}^{n} \frac{e(n \Delta t) - e[(n - 1) \Delta t]}{\Delta t} q(t - n \Delta t) \Delta t
\]

If now \( n \) is made to approach infinity, \( n \Delta t \) tending toward \( \tau \), the expression of \( r(t) \) becomes

\[
r(t) = \int_{0}^{\tau} \frac{de(r)}{dr} q(t - r) \, dr
\]

Hence,

\[
R(s) = \frac{H(s)}{s} sE(s) = \mathcal{L} \int_{0}^{t} \frac{de(r)}{dr} q(t - r) \, dr
\]

which is a proof of Duhamel's (or Borel's) theorem (Sec. 4.4.5) with

\[
f_{1}(t) = q(t) \quad f_{2}(t) = \frac{de}{dt}
\]

Note 1. Figure 7-8 can also be interpreted in terms of impulse functions by writing the approximate equation

\[
e(t) = \sum_{0}^{n} e(n \Delta t)u_{1}(t - n \Delta t) \Delta t
\]

whence, as \( \Delta t \) approaches zero, another expression for the convolution integral is

\[
H(s)E(s) = \mathcal{L} \int_{0}^{t} e(r)h(t - r) \, dr
\]

\( h(t) \) being the impulse response \( \mathcal{L}^{-1}H(s) \) of the system.

Note 2. The ramp-function response of a system can be used in a similar way to determine the response of the system to an input when the latter is considered to be the superposition of step and/or ramp inputs (e.g., succession of triangular pulses, etc.).

7.3.4. Important Consideration. If a system is at rest before \( t = 0 \) and is then subjected to an input \( e(t) \), the fundamental equation (7-3)

\[
R(s) = H(s)E(s) \quad (7-3)
\]

yields the Laplace transform of the response directly without involving initial conditions.

In other words, Eq. (7-3) yields both the steady-state response and that part of the transient response which takes place when the system starts from rest. If there are nonzero initial conditions, the response is given by

\[
R(s) = H(s)E(s) + g(s)
\]
where \( s(s) \) is a function of the initial condition [see Eq. (7-2)]. The term \( H(s)E(s) \) accounts for both the steady-state response and the transient response that corresponds to starting from rest, while the term \( s(s) \) accounts for the additional transient component due to nonzero initial conditions.

It should be pointed out that the term \( s(s) \) has the same denominator as the transfer function \( H(s) \). As a result, its poles are the zeros of the characteristic equation of the system. The results obtained in Sec. 4.6 may thus be determined by a shorter and simpler method, which will always be the case when the system starts from rest. Conversely, if the system does not start from rest and if discontinuities appear, it is necessary to specify the initial conditions at \( t = 0^+ \) by analyzing the function near \( t = 0^+ \). It may also be possible to introduce a change of variable, so that for all \( t < 0 \) the system will be considered to be at rest.

7.4. PHILOSOPHY OF THE TRANSFER FUNCTION: THE TWO PRACTICAL APPROACHES

7.4.1. Differential Equations and Transfer Functions. When a system is said to be linear, it is assumed that the differential equation contains complete information concerning the system. The transfer function of the system is an alternative way of expressing this information. By means of the Laplace transform, a differential equation may be replaced by an algebraic fraction involving the complex quantity \( s \). Mathematically, the differential equation and the transfer function are equivalent; in practice, however, the transfer function, based on use of general complex-variable theory, is easier to handle.

7.4.2. Introduction to the Theory of Analytic Functions. The general theory of functions of a complex variable is due to the French mathematician Cauchy who elaborated the theory of fonctions analytiques. Transfer functions, which are practically always polynomial fractions (sometimes multiplied by exponentials), represent a specialized application of analytic-function theory.

Broadly speaking, two theorems are fundamentally important to the theory of analytic functions. They are Cauchy's theorem and Liouville's theorem. Both are rather obvious when the analytic function under consideration is a polynomial fraction.

a. Cauchy's theorem may be formulated as follows: Any analytic function \( H(s) \) of the complex variable \( s \) is completely defined, i.e., its value is known for any location of the \( s \) point in the Laplace plane, when its values are known for all locations of the \( s \) point on a closed curve in the Laplace plane. In other words, it is possible to ascertain the value at any point \( s \) in the whole \( s \) plane from the values on any closed curve in the \( s \) plane. For example, let a function \( H(s) \) be known for all the values of \( s \) located on the imaginary axis,\(^1\) that is, for all \( s = j\omega \), whatever may be the real quantity \( \omega \). According to the theorem, the function \( H(s) \) has been completely defined. In physical terms this implies that the frequency response of a linear system gives complete information concerning its transfer function, i.e., enables its behavior under any circumstance to be specified.

b. Briefly, Liouville's theorem states that any analytic function is completely characterized by its singularities and the corresponding residues. So far as the discussion

\(^1\) If completed by a circle of "infinitely" great radius centered at the origin, the imaginary axis can be considered as a closed curve. For more detail, see Fig. 16-7.
is concerned here, it is equivalent to saying that the function is characterized by its poles and zeros. For example, if one knows the values of s which are the zeros of \( H(s) \), that is, the zeros of its numerator, and those which are its poles, that is, zeros of the characteristic equation of the system, then one has complete information concerning the system.

7.4.3. The Two Practical Approaches to Transfer Function. The two practical approaches for studying or synthesizing transfer functions are merely applications of the above two fundamental mathematical theorems concerning analytic functions. One can study the \( H(j\omega) \) function, i.e., the behavior of \( H(s) \) on the imaginary axis of the Laplace plane. This is the harmonic (or frequency-response or transfer-locus) approach. It is an implicit application of Cauchy’s theorem. Conversely, one can study the singularities of \( H(s) \), that is, study the location of its poles and zeros in the Laplace plane. This is the pole-zero-configuration approach. It is an implicit application of Liouville’s theorem.

Both approaches are, theoretically, perfectly equivalent. Technically speaking, the harmonic approach is more general in its application. When both approaches are possible, one of them may be preferable according to the problems under consideration, or according to the personal experience and preference of the engineer. It is suggested that the reader study and practice both approaches in order to master them equally well. He will then be able to choose one or the other; or, possibly, he will prefer to use them simultaneously, whenever possible, as they often complement one another. In the present book these two approaches will be presented and applied to feedback control systems. For linear systems in general, Chap. 8 will deal with the harmonic approach and Chap. 9 with the pole-zero-configuration approach. In Part 2, both approaches will be applied to linear feedback control system. The harmonic approach will lead to the Hall or Nichols chart (Chap. 13) and the Nyquist criterion (Chap. 16). The pole-zero-configuration approach will lead to Evans’ root-locus technique (Chap. 14).

Before a detailed study of these two approaches is undertaken, the present chapter will be concluded by giving some more information on how transfer functions may be obtained in practice.

Note. A third approach to transfer-function problems might consist in applying the concept of conformal mapping of analytic functions. This approach, however, is seldom used by feedback-control engineers. In the course of the present book, it will be only occasionally alluded to and related to the harmonic or to the pole-zero-configuration approach. (See Leonhard’s criterion in Sec. 16.4.4; also, see the left-hand criterion in Sec. 16.3 and Prob. 37.)

7.5. DETERMINATION OF TRANSFER FUNCTIONS

7.5.1. General. The determination of the transfer function of a linear system is equivalent to the determination of the differential equation relating the input and output of the system. The determination of a system’s transfer function must be carried out as follows:
1. The input and output variables are chosen from those referring to
the system.
2. The differential equation relating the input and output of the system
is written.
3. Neglecting the initial conditions, this differential equation is Laplace-
transformed and the ratio of the transformed input and output is obtained.
4. The corresponding locus is sketched. The Bode diagram, the
Nyquist plot, or the Nichols locus is used, depending on the problem
under consideration.

Remark. It is important that stages 2 and 3 be well separated. The differential
equation of the system must be written in terms of time functions before the Laplace
transform is used. The reasons for this are as follows:
a. If the functions of \( t \) and \( s \) are mixed by writing, for example, \( sf(t) \), an error will
almost surely be made, through confusion of \( sF(s) \) with

\[
L \frac{df}{dt} = sF(s) - f(0)
\]

b. If, subsequently, the time behavior of the system must be studied, one has only
to refer to the differential equation written in stage 2 and to associate the initial condi-
tions of the problem with it. The new operational equation is then inferred. If,
conversely, the equation had been directly written in terms of the Laplace trans-
form, in order to study the response of the system without neglecting the initial con-
ditions, it would be necessary to evaluate again stages 2 and 3; that is, to start again
from the beginning.

7.5.2. Examples. 1. Aircraft in Symmetric Flight. Let it be assumed that it is
necessary to study the performance of an aircraft that is in approximately horizontal
and symmetric flight:

1. It is logical to choose the angle of deflection \( \delta \) of the rudder as the controlling
input variable. The output variable may be the pitch angle \( \theta \) of the airplane, i.e.,
the angle the airplane makes with the horizontal plane; the pitch angular velocity
d\( \theta/dt \); or the angle of attack, i.e., the angle made by the airplane with its velocity
vector.

Let the pitch angle be chosen as output. As a simplification, the increments of the
horizontal speed \( V \) will be neglected. Furthermore, since the performance of the
aircraft in the neighborhood of horizontal flight is being studied, the increments of
the corresponding variables with respect to their values for steady horizontal flight
will be considered as the controlling inputs.

2. Let us represent the aircraft and the forces acting on it to be as shown in Fig. 7-9
(also see Figs. 1-13 and 1-14). To a first approximation, the vertical components of
the aerodynamic forces are equal to \( iK_1 \) (lift force caused by the incidence) and
\(-\delta K_3 \) (force caused by the rudder deflection). The quantities \( K_1 \) and \( K_3 \) are constant coefficients, and \( iK_1 \) and \( \delta K_3 \) are posi-
tive for a horizontal flight, as shown in the figure.

If it is assumed that \( \theta \) and \( \delta \) are small, the equations of the vertical forces and of
the torques with respect to the pitch axis are

\[\delta L = \delta F \tan \theta \]

\[\tau = iK_1 \delta + \delta K_3 \tan \theta \]

\[\theta = \theta \]

\[\delta = \delta \]

Fig. 7-9. Longitudinal motion of an airplane.
where \( \omega \) is the vertical velocity of the aircraft (positive upwards) and the angle of attack \( i \) is \( \theta - \omega / V \). For the horizontal flight (\( \omega = 0 \)), this may be written as

\[
0 = i_0 K_i - \delta_0 K_i l - P \\
0 = i_0 K_i h - \delta_0 K_i l \\
i_0 = \theta_0
\]

Hence, by subtraction,

\[
m \frac{d\omega}{dt} = i K_i - \delta K_i l \\
J \frac{d^2 \theta}{dt^2} + M \frac{d\theta}{dt} = -i K_i h + \delta K_i l
\]

\[
i = \theta - \frac{\omega}{V}
\]

where \( i \) and \( \theta \) are the increments with respect to steady flight conditions.

3. Assuming that the initial conditions are

\[
\frac{d\omega}{dt} = 0 \quad \theta = 0 \quad \frac{d\theta}{dt} = 0
\]

the Laplace-transformed equations are

\[
m s^2 W(s) = K_i I(s) - K_i l (s) \\
J s \Theta(s) + M s \Theta(s) = -l(s) K_i h + \Delta(s) K_i l
\]

\[
I(s) = \Theta(s) - \frac{W(s)}{V}
\]

Eliminating \( I \) and \( W(s) \) yields the transfer function

\[
H(s) = \frac{\Theta(s)}{\Delta(s)} = \frac{i K_i m V s + K_i K_i (l - h)}{J V m s^2 + (J K_i + M V m) s + K_i (h V m + M) s}
\]

This analysis is important, and therefore the loci representing its results, even though the hypotheses are oversimplified. In fact, a more detailed analysis would only yield similar equations, the coefficients of which would be more complicated functions of the mechanical and aerodynamic characteristics of the aircraft.

2. Normalized Dimensionless Variables. It is better to write the transfer function using the least parameters possible in such a manner that a minimum of parameters is used, for this enables a more general expression to be obtained. Thus, two problems which at first may appear to be quite different may be shown to be similar. These two problems are then said to be analogous. If one sets

\[
K = \frac{K_i (l - h)}{h V m + M} \\
r = \frac{i m V}{(l - h) K_i} \\
\omega^2_n = \frac{1}{J} \left( 1 + \frac{M}{h V m} \right) K_i = \frac{2 \omega}{\omega_n} = \frac{M V m + J K_i}{K_i (h V m + M)}
\]

the transfer function obtained in the preceding example becomes

\[
\Theta(s) = \frac{K}{s} \frac{1 + r s}{s^2/\omega^2_n + (2 s/\omega_n) s + 1} = \frac{K}{\omega_n} \frac{1 + r \omega_n u}{u^2 + 2 \omega u + 1}
\]

where \( s/\omega_n = u \) is a dimensionless (or normalized) ratio.

\( \dagger \) \( M \, d\theta / dt \) is an aerodynamic damping torque.
The transfer function depends only on the two parameters, $z$ and $\omega_n$, and on the quantity $K/\omega_n$. The value of $z$ usually varies between 0.05 and 0.2. It is thus possible to represent all aircraft in symmetrical flight, with the simplifying assumptions which have been made, by a set of loci sketched for the values $z = 0.05, 0.075, 0.1, 0.2,$ and $0.3$, the loci being graduated in $\omega/\omega_n$. They are shown in Figs. 8-1 to 8-6.

3. **Remark.** The denominator of the transfer function is of third order, whereas most analyses yield one that is of fourth order. This latter condition is necessary in order that the existence of the two natural oscillations, the high-frequency one and the phugoid one, be explained. This will not be discussed in detail because it lies beyond the scope of this book; it will be mentioned, however, that the paradox is caused by the assumption of constant speed. The assumption causes a neglect of the phugoid oscillation, which is principally related to the speed of the aircraft and is not related to the pitch angle, by having $s$ replace one of the two second-degree factors of the denominator.

The result of this assumption is that integration occurs and that a finite deflection of the rudder causes a pitching rate, rather than an incremental increase of the pitch angle. This result is not in accordance with experimental measurements, and is due to the assumption of constant speed. According to this, an aircraft could not be able to attain a new steady-flight condition upon rudder deflection.\(^1\)

7.5.3. **Limitations of This Method.** If the equations are of high order, the transfer function of a complex system is often difficult to obtain and to use. To obtain the transfer function or the transfer locus of the system, one must either study the components separately or have recourse to experimental tests. These methods will be studied in the following sections.

7.5.4. Determination of the Transfer Functions of More Complex Systems. **General.** To determine the transfer function of a complex system from the transfer function of the components, one eliminates the intermediate input and output variables to obtain a relation between the input and the output of the whole system. This can often be done by mere inspection of the block diagram of the system, which is a functional representation of the system, each box representing one of the components of which the transfer function is known.

An algebraic relation of the following type is thus obtained,

$$H = f(H_1, H_2, H_3, \ldots) \quad (7-6)$$

Thus one may either compute $H$ for each frequency by means of this equation or deduce the transfer locus $L$ of the whole system from the transfer loci of the components by means of a graphical interpretation of Eq. (7-6).

In general, the first thing that must be done in the study of a complex system is to determine its block diagram. Following this, the transfer functions of the components are computed and arranged to obtain the transfer function of the whole system.

7.5.5. Cascade Systems. The transfer function of a cascade system, that is, a system such that the output of one component is the input to another (Fig. 7-10), is easily obtained from the transfer functions of the components. Indeed,

\[ E_1(s) = H_1(s)E_0(s) \]
\[ E_2(s) = H_2(s)E_1(s) \]
\[ E_3(s) = H_3(s)E_2(s) \]

Therefore,

\[ E_3(s) = H_1H_2H_3E_0(s) \]  \hspace{1cm} (7-7)

The transfer function of the whole system is the product of the transfer functions of the components. The locus is obtained as a Nichols locus by vectorial addition at each frequency of the loci of the components. This is faster than eliminating \( e_1(t) \) and \( e_3(t) \) in the differential equations of the three components. It is also more intuitive in that, if a sinusoidal signal is used as the input to a cascade system, the magnitudes multiply and the phase shifts add.

*Example.* If it is desired to study the angle of attack of an aircraft in symmetrical flight with the rudder deflected, the block diagram of Fig. 7-11a is subdivided as shown in Fig. 7-11b. Hence, to deduce the transfer function of the aircraft \( H_{\text{aircraft}}(\delta,i) \), from that obtained from the aircraft and meter by means of flight tests

\[ H_{(\text{aircraft+meter})}(\delta,i_m) \]

one has to divide it by the transfer function of the meter \( H_{\text{meter}(i,i_m)} \), which therefore must be known.

7.5.6. Feedback Systems. Feedback systems, as defined in Chap. 1, form a very important class which includes in particular all servo systems. Consider a linear servo which slaves the motion of a system to a control input \( e(t) \): a sensing device compares the input \( e(t) \) with the actual motion \( r(t) \) (Fig. 7-12). This deviation, or error

\[ \varepsilon(t) = e(t) - r(t) \]

is the input to the forward path.
If $KG(s)$ is the transfer function of the forward path (open-loop transfer function), one may write

$$R(s) = KG(s)E(s)$$

These two equations yield the transfer function $H$ of the whole systems,

$$H(s) = \frac{R(s)}{E(s)} = \frac{KG(s)}{1 + KG(s)} \quad (7-8)$$

This equation will be more thoroughly studied and generalized in the second part of this book. In fact, this equation is the very basis of this latter part of the book.

![Fig. 7-12. Feedback (or closed-loop) system.](image)

If the feedback system incorporates superposed loops as shown in Fig. 7-13, which may, for example, be the case of an autopiloted aircraft (see Sec. 1.3.5), the over-all transfer function is obtained from step-by-step graphical inspection of the block diagram. First, Eq. (7-8) yields for the internal loops:

$$\frac{\Delta}{J} = \frac{B}{1 + B}$$

Then Eq. (7-7) yields for the over-all output/error transfer function

$$\frac{\Theta}{E} = A \frac{B}{1 + B} C$$

Hence finally, by means of Eq. (7-8),

$$\frac{\Theta}{\Theta_0} = \frac{A \frac{B}{1 + B} C}{1 + A \frac{B}{1 + B} C}$$

![Fig. 7-13. Superposed loops (e.g., piloted aircraft: see Fig. 1-45).](image)
CHAPTER 8
THE HARMONIC APPROACH: TRANSFER LOCI

Summary
1. Nyquist loci and equivalent representations.
2. Nichols loci. Practical construction of transfer loci (Bode plots).
3. Relation with the time response.
5. Relation between amplitude and phase response.

It has been shown in the preceding chapter that a linear system can be characterized by its transfer function $H(s)$, and that the practical study of $H(s)$ can be performed in the following two ways:

a. By studying the function $H(j\omega)$, where $\omega$ is real, which is the harmonic approach.

b. By studying the singularities of $H(s)$ in the complex plane, which is the pole-zero-configuration approach.

The purpose of the present chapter is to outline approach a.

8.1. NYQUIST LOCI AND EQUIVALENT REPRESENTATIONS

8.1.1. The Direct Transfer, or Nyquist, Locus. General. Consider a linear system $S$ with a transfer function

$$H(s) = \frac{A_m s^m + \cdots + A_0}{B_n s^n + \cdots + B_0} \quad (8-1)$$

If $s$ is given purely imaginary values $s = j\omega$, the complex quantity

$$H(j\omega) = \frac{A_m (j\omega)^m + \cdots + A_0}{B_n (j\omega)^n + \cdots + B_0} \quad (8-2)$$

is a function of $\omega$. Let $A(\omega)$ be its magnitude and $\Phi(\omega)$ its phase.

The transfer locus of the system $S$ is the geometrical locus $L$ in the complex plane (Nyquist plane) of the points $M$ represented by the complex number $H(j\omega)$, for all positive values of $\omega$. As $H(j\omega)$ is a rational fraction with real coefficients, the locus of the points which correspond to $\omega < 0$ is symmetric with $L$ about the real axis.

Figures 8-1 to 8-6 show the transfer loci of airplanes in symmetrical motion, the transfer functions of which have been derived in Sec. 7.5.2.

The transfer locus defines all the characteristics and the behavior of the system, but a certain amount of experience is needed before one can interpret a transfer locus by mere inspection. The following considera-
tions are intended to help the reader become acquainted with this technique.

2. Graduation. It must not be forgotten that the frequency graduation of a transfer locus is as important as its shape.

3. Resonance (Fig. 8-7). That point of the locus L whose distance from the origin is a maximum corresponds to resonance, the corresponding frequency being the resonant frequency. The ratio of the magnitude of the radius vector for this frequency to the radius vector corresponding to \( \omega = 0 \), when the latter is finite, characterizes the peak value of magnification: it is called the peak-value-of-magnification coefficient, or the resonance ratio of the system for the frequency under consideration, and is generally designated by \( M_r \) or \( Q \).

4. High Frequencies. That portion of the transfer locus which lies near the origin nearly always corresponds to the high frequencies. This
Fig. 8-2. Transfer locus of airplane \( z = 0.05 \).

\[
\theta = \frac{K}{\omega_n} \times \frac{1 + r \omega_n (ju)}{(ju)^2 + 2z (ju) + 1}
\]

\( r \omega_n = 15, 20, 30 \)

Fig. 8-3. Transfer locus of airplane \( z = 0.075 \).

\[
\theta = \frac{K}{\omega_n} \times \frac{1 + r \omega_n (ju)}{(ju)^2 + 2z (ju) + 1}
\]

\( r \omega_n = 13, 20 \)
Fig. 8-4. Transfer locus of airplane ($z = 0.10$).

Fig. 8-5. Transfer locus of airplane ($z = 0.15$).
must always be true for positional systems, because the ratio of the output to the input is zero at high frequencies owing to the inertia of the components. Hence, for such a system the degree of the denominator is necessarily higher than that of the numerator. Some transfer functions may not satisfy this condition because they have been simplified by omission of some of the inertias, or analog quantities, which may lead to contradictions.

The tangent to the transfer locus at the origin is a function of the difference between the degree of the denominator and the degree of the numerator. An example of this is as follows:

\[ H(j\omega) \] [see Eq. (8-2)] is equivalent to

\[ \frac{A_m}{B_n (j\omega)^n - m} \]

When \( A_m \) and \( B_n \) are positive,\(^1\) the following results are obtained:

\(^1\) Because of this restriction, these results should be applied only to systems which are stable (Sec. 9.1) and have minimum phase (Sec. 8.5). The transfer function of
a. If \( n - m = 1 \), \( H(j\omega) \) is equivalent to \( K/j\omega \), and the locus is tangent to the negative \( j \) axis (Fig. 8-8). Typical examples of this are the first-order systems studied in Chap. 5.

![Fig. 8-8.](image)

![Fig. 8-9.](image)

b. If \( n - m = 2 \), then \( H(j\omega) \) is equivalent to \( K/(j\omega)^2 \) and the locus is tangent to the negative real axis (Fig. 8-9). Typical examples of this are the second-order systems studied in Chap. 6.

c. If \( n - m = 3 \), \( H(j\omega) \) is equivalent to \( K/(j\omega)^3 \) and the locus is tangent to the positive \( j \) axis (Fig. 8-10).

![Fig. 8-10.](image)

d. Finally, it is apparent from the above that a phase shift of \( \pi/2 \) occurs each time the \( n - m \) difference is increased by 1.

5. Low Frequencies. On the transfer locus, the zero-frequency point corresponds to the steady-state behavior of the system.

a. The zero-\( \omega \) point is at finite distance if the system has no integration. Its distance from the origin is equal to the static gain of the system. Note that the transfer locus generally starts perpendicularly from the real axis in the low-frequency region, because the curve consisting of the transfer locus \( L \) plus its symmetric curve \( L' \) is an algebraic curve that can have no discontinuity in slope. As a consequence, the tangent at the point \( \omega = 0 \) is vertical; that is, \( (dA/d\omega)_{\omega=0} \) is zero.

b. On the contrary, if the system incorporates integration, the \( \omega = 0 \) point lies at infinity. A first example is a field-controlled d-c motor, the output being the angular position of the rotor shaft (Fig. 8-11). A constant field current \( i \) does not produce a constant position \( \theta \) of the output shaft, but results in a constant angular velocity \( d\theta/dt \) (under given load conditions): the system has integration. At low frequencies, the asymptotic line of the transfer locus is the negative \( j \) axis (Fig. 8-12).

A second example is a system with inertia and viscous friction, the input to this system being a torque (Fig. 8-13). The equations are

---

an unstable system has at least one positive pole (or pole with positive real part), and that of a nonminimum-phase system has at least one positive zero. Hence, if the system considered is known to be stable and of minimum phase, \( A_m \) and \( B_n \) are sure to be positive; otherwise, no statement can be made regarding the signs of \( A_m \) and \( B_n \).
\[ \dot{\theta} = J \frac{d^2\theta}{dt^2} + f \frac{d\theta}{dt} \]
\[ \frac{\Theta}{C} = \frac{1}{Js^2 + f_0} \] (8-3)

The transfer locus tends to become parallel to the negative \( j \) axis (Fig. 8-11). A third example has been studied in Sec. 7.5.2.

c. In the same way, a system may have a second-order pole at the origin, i.e., two integrations. This is the case for the second of the examples just quoted when the viscous friction is equal to zero, since the transfer function (8-3) becomes, for \( f = 0 \), \( \Theta/C = 1/Js^2 \). The locus then has a parabolic branch parallel to the negative real axis (Fig. 8-14).

8.1.2. The Inverse Transfer Locus. The inverse transfer function \( 1/H(s) \) and the inverse transfer locus may also be considered. The latter is deduced from the direct locus by taking its inverse with the origin as pole and then taking the symmetrical locus with respect to the real axis. An example has been given in Sec. 6.2.6. Many systems, especially feedback systems with nonunity feedback, can be studied more easily by means of these loci. Unfortunately, these loci are seldom used in practice.

8.1.3. First Equivalent Representation. Instead of plotting the transfer locus \( L \), one may also separately represent the magnitude and the phase shift as functions of the frequency by drawing the frequency-response curves or the Bode diagram of the system. One obtains the magnitude-vs.-frequency response and the phase shift-vs.-frequency response. This representation is practical for a qualitative analysis or interpretation of \( H(j\omega) \), and especially useful for the quantitative study of a particular system.

A resonance corresponds to a maximum of the magnitude. The angular frequency \( \omega_R \) at which it occurs is called the resonance frequency. Generally, the slope \( |d\Phi/d\omega| \) of the phase-shift curve is a maximum in the neighborhood of \( \omega_R \). This is a consequence of Bode's equation (Sec. 8.5), and can be intuitively conceived by a consideration of the transfer locus provided that the frequency graduation is sufficiently continuous.\(^1\)

8.1.4. The Use of Logarithmic Coordinates. It is often easier to work with logarithmic coordinates. For example, for the magnitude curve

Abscissa: \( \omega \) octave = \( \log_2 \omega \)
Ordinates: \( A \) db = \( 20 \log_{10} A \)

\(^1\) Also see Sec. 9.3.2.
the slope of the curve is measured in \textit{decibels per octave}.\footnote{\omega \ is \ sometimes \ measured \ in \ \textit{decades}} \ At high frequencies, this slope characterizes the difference \( n - m \) between the degree of the numerator and the degree of the denominator of the expression for the transfer function.

Six decibels per octave, if \( n - m = 1(A \cong K/\omega \) for first-order systems)

Twelve decibels per octave, if \( n - m = 2(A \cong K/\omega^2 \) for second-order systems)

Six times \( q \) decibels per octave, if \( n - m = q(A \cong K/\omega^q) \)

\textbf{8.1.5. Other Equivalent Representations.} Other representations are also possible. They are theoretically equivalent, but may be more advantageous for some applications. For example, the real and complex parts of the transfer function,

\[ A \cos \Phi \quad \text{and} \quad A \sin \Phi \]

may be represented separately as functions of \( \omega \).

\textbf{8.2. NICHOLS LOCI. PRACTICAL CONSTRUCTION OF TRANSFER LOCI (BODE PLOTS)}

\textbf{8.2.1. The Use of Logarithmic Coordinates.} In practice, the transfer locus is often drawn in a Nichols plane with coordinates:

\begin{align*}
\text{Abscissa:} & \quad \Phi \quad \text{degrees} \\
\text{Ordinates:} & \quad A \text{ db } = 20 \log A
\end{align*}

It is then termed the "Nichols locus" or the "magnitude-phase locus." This is equivalent to representing \( H(j\omega) \) in the \( j \log H(s) \) plane, or, more precisely, in the \( j \log^* H(s) \) plane. The asterisk indicates the conjugate complex quantity because of the positive sense chosen for \( \Phi \) on the usual diagrams. These are shown symmetrical with respect to the \( \Phi = 0 \) axis in Figs. 8-15a and b. The ordinates are graduated either in decibels, as in the figure, or in numerical values with a logarithmic scale. Figure 8-16 shows the Nichols locus of a second-order-plus-integration system.

That the Nichols coordinates are of interest is due to the following two facts which are of fundamental importance in the design of feedback control systems:

\( a. \) A variation of system gain corresponds to a translation parallel to the ordinate axis.

\footnote{\omega \ is \ sometimes \ measured \ in \ \textit{decades}}

\[ \omega \ \text{decade} = \log_{10} \omega \]

and the magnitude in \textit{decilogs} (abbreviated \( \text{dg} \))

\[ A \ \text{dg} = 10 \log_{10} A \]

At high frequencies the slope of the magnitude curve is then equal to \(-10(n - m)\) dg/decade.
Fig. 8-15. Usual origin and positive directions for Nichols plane.

Fig. 8-16. Nichols loci for $\frac{1}{j\omega [1 + 2\pi j\omega + (j\omega)^2]}$. 
b. The multiplication of numerous transfer functions is performed by a vectorial addition. Thus,

\[
\log A = \log A_1 + \log A_2 + \cdots \\
\Phi = \Phi_1 + \Phi_2 + \cdots
\]

8.2.2. Interpreting a Nichols Locus.  

1. In the Nichols plane (Fig. 8-17) a resonance corresponds to a peak in the transfer locus, i.e., to a point the ordinate of which is a maximum. The resonant angular frequency is given by the graduation of the locus. The peak value of magnification \( Q \) db is equal to the vertical distance between points \( \omega = 0 \) and \( \omega = \omega_R \). It must be remembered that \( Q = 1.3 \) corresponds to 2.3 db.

2. At high frequencies the asymptotic line of the locus is generally parallel to the negative \( \log A \) axis. For stable and minimum-phase systems (see footnote of Sec. 8.1.1, par. 4):

a. If \( n - m = 1 \), it becomes the \( \Phi = -90^\circ \) line (example of first-order systems, see Fig. 8-18).

b. If \( n - m = 2 \), it becomes the \( \Phi = -180^\circ \) line (example of second-order systems, see Fig. 8-18).

c. If \( n - m = 3 \), it becomes the \( \Phi = -270^\circ \) line (example of third-order systems).

3. At very low frequencies, in the neighborhood of the steady state, the locus tends to the static gain, which is the ordinate of the point \( \omega = 0 \). If there is one integration, the locus tends toward infinity for \( \omega = 0 \) along the \( \Phi = -90^\circ \) line. If there are two integrations, the locus tends toward infinity along the \( \Phi = -180^\circ \) line, etc.

Example. The locus shown in Fig. 8-19 represents the Nichols locus of the longitudinal motion of a guided missile flying at constant speed; the input is the rudder deflection, and the output is the pitch angle. Inspection of this locus shows that:

a. The system has one integration; that is, a finite deflection of the control surface results in a pitching rate (see Sec. 7.5.2, par. 3).

b. The resonance frequency is \( \omega_R = 3 \) rad/sec = 0.5 cps.

c. The difference between the degree of the denominator and that of the numerator is equal to 2.
If it be known that the degree of the denominator is equal to 3, which is a classical result in airplane dynamics, it may be asserted that the transfer function is
\[
\frac{As + B}{s(s^2 + 2\omega_n s + \omega_n^2)}
\]
where \( \omega_n \) and \( \omega_R \) usually do not greatly differ.

8.2.3. The Practical Drawing of Nichols Loci. Transfer loci are practically drawn as Nichols loci by use of the following technique, called the Bode plot technique.

1. Steps of the Bode Plot Technique. a. By factoring of the numerator and denominator of the transfer function, one obtains a product of factors of the form:
\[
\frac{1}{ju} \quad \frac{1}{1 + ju} \quad \frac{1}{1 + 2\zeta (ju) + (ju)^2}
\]
where \( u \) is the normalized frequency (Secs. 5.2.2. and 6.2.3.)

b. Following this, one draws the magnitude responses of the different factors in logarithmic coordinates and the phase-vs.-frequency responses in semilogarithmic coordinates. The technique for drawing these curves will be explained below. Graphical additions then yield the magnitude and the phase of the transfer function.

c. The Nichols locus is then plotted, with frequency graduation.

2. Magnitude and Phase Response for Elementary Factors. The method of drawing the magnitude and the phase-vs.-frequency responses for the elementary factors will now be studied.

a. For the factor \( 1/ju \) the magnitude response is a straight line with negative slope of 6 db/octave (or 10 dg/decade), which passes through the point with coordinates \((u = 1, A = 0 \text{ db})\). The phase response is the line \( \Phi = -90^\circ \).

The responses of the \( ju \) factor are easily deduced from these results (opposite decibel magnitude and phase).

b. The magnitude response of the factor \( 1/(1 + ju) \) has been studied in detail in Sec. 5.2.4 (see dotted curves in Charts 1 and 2, at the back of the book, and in Fig. 6-9). It has two asymptotes which meet at the break point \((u = 1, A = 0 \text{ db})\). One is horizontal, \( A = 0 \text{ db} \); the other is oblique, with a negative slope equal to 6 db/octave (or 10 dg/decade).

The position of the magnitude-response curve with respect to its asymptotes is, in practice, specified with sufficient accuracy by its ordinates at the break frequency and at one octave above and one octave below this frequency. At the break frequency \((u = 1), A = -3 \text{ db} \); that is, the curve lies 3 db under the asymptotes. At one octave above \((u = 2)\), the curve lies 1 db under the oblique asymptote. At 1 octave
below \((u = 0.5)\), \(A = 1\) db; that is, the curve lies 1 db under the horizontal asymptote.

The phase-response curve (Sec. 5.2.4), which starts at \(\Phi = 0\) at \(u = 0\) and decreases to \(\Phi = -90^\circ\) at infinite frequencies, has a point of inflection and a symmetry center at the break frequency \((u = 1, \Phi = -45^\circ)\). It is graphically defined by the points:

\[
\begin{align*}
u & = 0.5 & u & = 2 \\
\Phi & = -26.5^\circ & \Phi & = -63.5^\circ
\end{align*}
\]

The responses of the following factors are easily deduced:

Same magnitude, opposite phase: \(\frac{1}{1 - ju}\)

Opposite decibel magnitude and phase: \(1 + ju\)

c. The magnitude response of the factor \(1/[1 + 2zju + (ju)^2]\) has been studied in Sec. 6.2.4. Its asymptotes intersect at the break point \((u = 1, A = 0\) db). One is horizontal, \(A = 0\) db; the other is oblique with a negative slope equal to 12 db/octave (or 20 dg/decade). When the asymptotes are drawn, the curves may be traced from Chart 1 at the back of the book, in which 10 cm represents 20 db or 1 decade. This diagram is also shown in Fig. 6-9a with a one-half scale.\(^1\) The phase-response curve starts from \(\Phi = 0\) for \(u = 0\) and goes down to \(\Phi = -180^\circ\) for infinite frequencies. It has a point of inflection at the break frequency \((u = 1, \Phi = -90^\circ)\) and may be traced from Chart 2 at the back of the book, in which 10 cm represents 100° or 1 decade. This diagram is also shown in Fig. 6-9b, to one-half scale.

The response of \(1 + 2zju + (ju)^2\) is similarly obtained (opposite decibel magnitude and phase).

3. Introduction of the Gain. To draw a transfer locus \(KG(j\omega)\), the locus \(G(j\omega)\) is generally drawn by the above method. Then the gain \(K\) is introduced by vertically shifting either the locus \((K\) db upward if \(K > 1)\), or the scale \((K\) db downward if \(K > 1)\).

8.2.4. Important Remark. The Bode plots are the fundamental tool of the servo engineer for most applications. It is absolutely necessary that the reader train himself until he has perfect mastery of the corresponding technique. Use of the charts at the back of the book makes it possible to draw Nichols loci on ordinary millimetric tracing paper.

8.3. RELATION TO THE TIME RESPONSE

8.3.1. General. A linear system is completely defined by its transfer locus; therefore it is possible to draw from the frequency response complete information concerning the time response of the system. It is the purpose of the following sections to show how such information can be obtained in practice. It is suggested that the reader, whenever he meets

\(^1\) In these figures, the scale chosen for 20 db is equal to that chosen for 1 decade, the slope of the oblique asymptote being \(-1\); that is, 45°, for the first-order factor, and \(-2\) for the second-order factor.
a transfer function, acquire the habit of systematically translating it into terms of time response, at least in the form of a qualitative sketch.

8.3.2. Low-frequency Region of the Transfer Locus. The low-frequency region of the transfer locus \((\omega \to 0)\) corresponds to the static behavior of the system.

The presence or absence of integration immediately gives information concerning the steady state in the system response to a step, impulse, or velocity input (Sec. 7.2.1). The value of the gain completely specifies this steady state.

The velocity constant \(C_v\) of the system can be directly read from the phase-vs.-frequency response. It can be shown that, if the static gain is unity, \(1/C_v\) is the slope of the phase-vs.-frequency curve in the low-frequency region (as \(\omega\) approaches zero). This can be easily seen from Sec. 7.2.5, or it can be shown directly as follows. From Eq. (7-5):

\[
-\frac{1}{C_v} = \lim_{s \to 0} \frac{dH(s)}{ds} = \lim_{\alpha \to 0} \frac{dH(\alpha + j\omega)}{d(\alpha + j\omega)}
\]

Writing \(H(j\omega)\) as \(A(\omega) \exp[j\Phi(\omega)]\) gives

\[
-\frac{1}{C_v} = \lim_{\omega \to 0} \frac{d}{d\omega} [Ae^{j\Phi}] = \lim_{\omega \to 0} \frac{1}{j} \left[ \frac{dA}{d\omega} e^{j\Phi} + Aj \frac{d\Phi}{d\omega} e^{j\Phi} \right]
\]

As \(\omega\) approaches zero, \(A \exp(j\Phi)\) becomes \(H(0) = 1\), and \(dA/d\omega\) becomes zero (see Sec. 8.1.1, par. 5a), whence

\[
\frac{1}{C_v} = \left| \frac{d\Phi}{d\omega} \right|_{\omega=0}
\]

8.3.3. High-frequency Region of the Transfer Locus. By use of the initial-value theorem and the results obtained in Secs. 8.1.1 or 8.1.4, it is possible to specify the initial slope \(dr/dt\) of the response of the system to a step or impulse, provided the data available concerning the high-frequency region are sufficiently accurate.

8.3.4. Resonance. If the frequency response shows a marked resonance effect at a frequency \(\omega_R\), the time response exhibits oscillations at a frequency which is close to \(\omega_R\). This frequency is called the natural frequency of the system. The corresponding resonance ratio \(Q\) gives an idea of how much these oscillations are damped.

The sharpness of resonance can also be measured by the abruptness of the fall of the phase-vs.-frequency curve near \(\omega_R\). This procedure can also be applied to the case of oscillatory modes characterized by a damping ratio greater than 0.7 (\(\Psi\) greater than 45° in Fig. 6-27 or 9-11), that is, when no actual resonance peak occurs.\(^1\)

8.3.5. Duration of the Transient. The considerations of Sec. 6.4.5 concerning the relation between passband and speed of response can be

\(^1\)In this sense, \(\max \frac{d\Phi}{d\omega}\) is a more convenient parameter than is the resonance ratio \(Q\). The theory and technique of its manipulation have been developed by M. Saintillan, "Cours de servomécanismes linéaires," vol. 2, pp. 42–48, École Supérieure du Génie Maritime, Paris.
generalized to all systems which can be considered as low-pass filters: the wider the passband, the faster the transient.

This is sometimes\(^1\) stated in the following quantitative way:

\[ \omega_c T = \pi \]

where \( \omega_c \) is the cutoff frequency and \( T \) is the response time. This equation has the merit of dramatizing the inverse relationship between passband and response time. Unfortunately, the definitions it assumes for \( \omega_c \) and \( T \) are not the usual ones.

**8.3.6. Approximate Sketch of the Time Response.** The results just outlined enable one to perform a rough sketch of the step or impulse response of a system from its frequency response. This procedure can be systematized in many different ways. Semiempirical relations can give the important characteristics of the transient response with an accuracy of the order of 25 per cent for most practical systems.\(^2\)

![Fig. 8-20.](image)

**8.3.7. Obtaining the Time Response Itself. Proust's Method.** It is possible actually to plot the step response of a linear system characterized by its frequency-response curves. Several methods are available.\(^3\) Floyd's method\(^4\) is the best known in the United States. In what follows a purely graphical method is presented, one that is due to R. Proust.\(^5\) The method consists in obtaining the response of the system to the periodic input (Fig. 8-20)

\[ e_T(t) = M \left[ u(t) - 2u\left( t - \frac{T}{2} \right) + 2u(t - T) - 2u\left( t - \frac{3T}{2} \right) + \cdots \right] \]

by making use of the Fourier expansion of \( e_T(t) \). If the period \( T \) is chosen large enough, the response thus obtained can be considered as the step response of the system.

---

5. "Cours de servomécanismes linéaires," Centre d'Instruction des Transmissions, Pontoise, France.
The Fourier expansion of $e_T(t)$ is (Sec. 4.5.1)

$$e_T(t) = \frac{4M}{\pi} \left[ \sin \frac{2\pi}{T} t + \frac{1}{3} \sin \frac{2\pi}{T} t + \frac{1}{5} \sin \frac{2\pi}{T} t + \cdots \right]$$  \hspace{1cm} (8-4)

If the system is characterized by its frequency-response curves $A(\omega)$ and $\Phi(\omega)$, its response will be, on letting $\omega = 2\pi / T$ for the sake of brevity,

$$r_T(t) = \frac{4M}{\pi} \left[ A(\omega) \sin (\omega t + \Phi(\omega)) + \frac{1}{3} A(3\omega) \sin (3\omega t + \Phi(3\omega)) + \cdots \right]$$
Thus \( r_\pi(t) \) is obtained by graphical addition of the curves \( \sin \omega t \), \( \frac{1}{3} \sin 3\omega t \), \( \frac{1}{5} \sin 5\omega t \ldots \), each being multiplied by the corresponding \( A \) and shifted by the corresponding \( \Phi \).

The critical point in the procedure is the choice of \( T \). Proust suggests that \( T \) should be chosen in order that the 12-db cutoff frequency \( \omega_c \) of the system corresponds to the eleventh harmonic in the Fourier expansion (under such conditions, the amplitude of the thirteenth harmonic will not exceed 0.24/13, that is, less than 2 per cent of the fundamental). Hence

\[
\omega_c = 11 \frac{2\pi}{T} \quad \omega = \frac{\omega_c}{11} \quad T = \frac{69}{\omega_c}
\]  

(8-5)

Once \( T \) has been found, \( A(n\omega) \) in decibels and \( \Phi(n\omega) \) are read for each harmonic from the frequency-response curves. Universal \( A \sin n\omega t \) curves are shown in Fig. 8-21a and b, \( n = 1 \) to 11, and for different values of \( A \) (in decibels). For each harmonic, the \( A \sin n\omega t \) curve with \( A = A(n\omega) \) should be shifted by \( \Phi(n\omega) \) and drawn on tracing paper. The six curves obtained should then be added graphically, thus yielding the response of the system.
Example. The curves $A$ and $\phi$ are shown in Fig. 8-22. The 12-db cutoff frequency is $\omega_c = 2.1 \text{ rad/sec}$. Hence, $T = 69/\omega_c = 32.9 \text{ sec}$; and Table 8-1 gives the amplitude and phase of each harmonic.

<table>
<thead>
<tr>
<th>Harmonic</th>
<th>rad/sec</th>
<th>$A$ (db)</th>
<th>deg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.19</td>
<td>0</td>
<td>-11</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>1</td>
<td>-41</td>
</tr>
<tr>
<td>5</td>
<td>0.96</td>
<td>0.1</td>
<td>-86</td>
</tr>
<tr>
<td>7</td>
<td>1.34</td>
<td>-3.75</td>
<td>-120</td>
</tr>
<tr>
<td>9</td>
<td>1.72</td>
<td>-8.5</td>
<td>-140</td>
</tr>
<tr>
<td>11</td>
<td>2.1</td>
<td>-12</td>
<td>-145</td>
</tr>
</tbody>
</table>

The procedure is summarized in Fig. 8-23, in which the harmonics $A(n\omega)(1/n) \sin [n\omega t + \phi(n\omega)]$ are drawn and added graphically. The time scale is fixed by $T$, since $T/4$ is equal to 8.2 sec.

Note. Once the step response is obtained, the impulse or ramp response can easily be obtained by graphical differentiation or integration, by application of Ludbrook’s theorems (Sec. 7.3.1).

8.3.8. Remark Concerning the Forced Sinusoidal Response. When a linear system is subjected to a sinusoidal input $e_0 \sin \omega t$, the response is $Ae_0 \sin (\omega t + \phi)$, where $A$ and $\phi$ are the magnitude and phase of the transfer function at the corresponding $\omega$. In the case of a control system,
it is desired that the output reproduce the input as faithfully as possible. For this, \( A(0) = 1 \); that is, the static gain should be unity, an obvious condition. But it must not be forgotten that, when \( A(0) \) is unity, there may well be an error \( e(t) - r(t) \) between the instantaneous values of \( e(t) \) and \( r(t) \). This error is a result of the phase shift \( \Phi(\omega) \). Figure 8-24 clearly shows that the instantaneous error is a maximum when \( e(t) \) is zero, the maximum value referred to \( e_0 \) being equal to \( \Phi(\omega) \) radians.
8.3.9. Regular Systems. If the Floyd and Proust methods, which are always applicable, be excepted, the foregoing considerations are simple intuitive extensions, in approximate equation form, of properties pertinent to second-order systems. To what extent are they legitimate? It is very difficult to answer this question (furthermore, the exact meaning of the word legitimate would have to be specified). For practical purposes, however, it can be considered that such extensions are valid when the transfer locus of the system has a shape comparable to that of a second-order transfer locus, at least in the neighborhood of the resonance. Systems satisfying this condition can be called regular systems.

For example, second-order-plus-integration (Fig. 8-16) or second-order-plus-lag systems are regular systems. Conversely, a transfer locus with two pronounced resonances is not regular, unless the two corresponding frequencies are sufficiently separated (by at least two octaves) so that the system may be represented by the product of two regular transfer functions the resonance frequencies of which are located in two different frequency domains. This applies, in particular, to aircraft where separation is made of the fast oscillation, which principally concerns movement about the center of gravity, and the phugoid, which essentially concerns the movement of the center of gravity (see Sec. 7.5.2). Although the concept of regular system may appear somewhat vague and indefinite, it is of great practical use, and it will be frequently used in the following chapters.

8.3.10. Unstable Systems. It has been shown that second-order systems become less damped as $\zeta$ is decreased. If a negative damping were realized by means of an auxiliary device, an unstable system would result. Its response to an excitation would tend toward infinity with increasing oscillations. In practice, however, the magnitude would be limited by stops or by system breakdown, i.e., by conditions for which the system is no longer linear. More generally, unstable systems of high order may be considered. As will be seen later, servo systems may become unstable under certain conditions.

Transfer functions and transfer loci are defined for these systems as well as for others. The only difference is that one can no longer speak of frequency response, since the mathematical forced response is concealed by the diverging transient state.

8.4. PRINCIPLE OF EXPERIMENTAL TECHNIQUES FOR THE DETERMINATION OF TRANSFER LOCI

8.4.1. Response to Sinusoidal Inputs. The most obvious method for experimentally determining the transfer function of a given system is that of subjecting the system to a sinusoidal (frequency-response) test. The system is excited by means of a sinusoidal input and the magnitude and phase of the output are measured. When this has been done for many different frequencies, the transfer function may be evaluated.

This method, which is very general, was first applied to electrical

1 Similar to the negative resistances which are sometimes considered in electrical engineering.
systems. Since the end of the World War II, it has been also applied to mechanical systems; it was first applied to aeronautical problems in the United States by Milliken, later by Seamans. The performance of an aircraft was studied by producing sinusoidal movements of the control surfaces and recording the aircraft response. This method is now commonly used in Europe as well as in the United States. It is also applied to wind-tunnel models having one or more degrees of freedom and to guided missiles in which a cam controls the control-surface deflection.

The techniques used when applying the sinusoidal-test method differ, of course, according to the nature of the system to be tested. They generally make use of laboratory equipment specially designed to generate sinusoidal inputs and to record the response of the system. The majority of these types of apparatus give directly the magnitude $A$ and the phase $\Phi$ of the transfer function. Some give the real and the imaginary parts, $A \cos \Phi$ and $A \sin \Phi$, respectively.

This sinusoidal-test method gives the transfer function directly and intuitively. Its prime disadvantage is that it is lengthy and costly. For example, in aircraft flight tests for each frequency one must await the occurrence of the forced response. This is a long and expensive process.

It will be shown in Chap. 11 that the sinusoidal-test method very often necessitates the use of frequencies that are much higher than those which would be considered appropriate a priori. This is an important point which is too often overlooked.

Remark. When a system $S$ is linear, its response to a sinusoidal input $e_0 \sin \omega t$ is a sine wave of the same frequency; i.e., the higher harmonics are negligible. Moreover, $A$ and $\Phi$ depend on the forcing frequency $\omega$ but not on the input amplitude $e_0$. Conversely, given a physical system subjected to sinusoidal tests, the condition that $A$ and $\Phi$ be independent of the input amplitude is a condition for the linearity of the system. If $A$ and $\Phi$ depend on $e_0$, the system is not linear, but the experimental functions $A(e_0, \omega)$ and $\Phi(e_0, \omega)$ constitute a kind of generalized transfer function and can be used for characterizing the system (Chap. 24).

8.4.2. Response to Step Inputs. The transfer function of a linear system may be obtained by means of a step-input test. The input varies instantaneously from zero to a finite value. The input and the response are recorded. Mathematically, this can be expressed by the Fourier transformation, as follows: The input can be considered as the function shown in Fig. 8-20, when $T$ is very large. The Fourier expansion of this function is known [see Eq. (8-4)]. The response can also be approximated by rectangles; that is, the response is replaced by a superposition of functions of the $e_T(t)$ type, to which Eq. (8-4) is applied. Thus, the input and output are resolved into sinusoidal functions, the magnitude and the phase being compared at each frequency.


In aeronautical engineering, this method is used by applying a sudden deflection to the control surface, and the results are used to complement the results of sinusoidal tests. The method is applied to wind-tunnel models and guided missiles as well as to aircraft.

Instead of using a step input, an impulse may be used. The transfer function of the system is the Laplace or Fourier transform of the unit-impulse response.

The advantage of these two methods is obvious. The tests are not as elaborate as the sinusoidal method. They are short and generally easy to carry out. On the other hand, the ideal step or unit-impulse input is never obtained; consequently the interpretation of the results is relatively long.

Remark. As in the case of sinusoidal tests, the proportionality of output to input is a criterion of the linearity of the system. If \( q(t) \) is the response of a given system \( S \) to a unit-step input, its response to step inputs of different magnitudes \( a_1 u(t), a_2 u(t), \) etc., should be recorded. If these responses are \( a_1 q(t), a_2 q(t), \) etc., the system \( S \) is linear; if they are different, the system is not linear. The results thus obtained are, in general, less convenient than those obtained by applying the sinusoidal-test method with different input amplitudes.

8.4.3. Response to an Arbitrary Input. The reasoning in the foregoing paragraph is also valid when the input to the system is arbitrary. The input and the response are divided into sinusoidal functions; they define the behavior of the system at each angular frequency and hence its transfer function. Mathematically, if the input can be written as

\[
\frac{e(t)}{e_0} = 1 + e_1 e^{i(\omega t + \varphi_1)} + \cdots + e_n e^{i(n \omega t + \varphi_n)} + \cdots
\]

and the corresponding response \( r(t) \) of the system

\[
\frac{r(t)}{r_0} = 1 + r_1 e^{i(\omega t + \varphi_1)} + \cdots + r_n e^{i(n \omega t + \varphi_n)} + \cdots
\]

the transfer function of the system can be deduced from the amplitude ratio \( A \) and the phase \( \varphi_n - \psi_n \) for each angular frequency \( n\omega \)

\[
A = \frac{r_n}{e_n} \quad \Phi = \varphi_n - \psi_n
\]

In practice, the so-called "arbitrary" input is chosen—starting from a steady state and leading to a steady state. The results can be conveniently analyzed by means of a harmonic analyzer of the Coradi type.

In the field of aeronautical engineering, this method was first applied in Europe by J. F. Vernet to the determination of the transfer function of an aircraft, producing a rudder motion such as that shown in Fig. 8-25.\(^1\) The advantages of the method are obvious. It not only requires

tests that are as short as those of the step-input method, but also does not require sinusoidal or step inputs. Unfortunately the analysis of the results is a lengthy process.

8.4.4. Significance of These Methods. The method of interpreting the results of these tests will not be discussed here because it varies considerably with the particular case under consideration; furthermore, it is still in the process of development in some fields. Only the significance of the tests will be discussed. All the foregoing methods deduce \( H(s) \) from the fundamental equation

\[
H(s) = \frac{R(s)}{E(s)} \tag{8-6}
\]

They differ only in the choice of \( E(s) \):

- \( E(s) = K/s \) for the step input
- \( E(s) = K \) for the impulse input
- \( E(s) \) is "arbitrary" for Vernet's method

The harmonic analysis is equivalent to the evaluation of \( E(s) \) and \( R(s) \) from \( e(t) \) and \( r(t) \), because of the similarity between Laplace and Fourier transforms pointed out in Chap. 4 (Sec. 4.5 and appendix to Chap. 4, Sec. 3).

8.4.5. Interpretation of These Methods. The foregoing methods characterize a system by means of a small number of tests, instead of studying its behavior in normal functioning. Hence they are rather similar to the psychological tests used for characterizing a person by studying his reactions to artificially created situations. In both cases, a theory is required to interpret the results. Just as the layman may not see the relation between some tests and the aptitudes which are to be judged, so the response of the system must be studied at frequencies higher than those which will occur during normal operation.

The fundamental equation (8-6) shows that all the tests are at least theoretically equivalent and that the interpretation of only one test is sufficient to characterize the system completely. This very simple result is a consequence of the assumption of linear operation. The number of properties of a linear system is small enough for the system to be characterized by one single test.

It should be stressed, however, that the theoretical identity of these different methods is moderated by the range of frequencies that is of interest. For example, in the sinusoidal-test method it is obvious that, if one is particularly interested in the resonance region, a greater number of tests must be performed in the vicinity of the resonant frequency. A similar remark is applicable to the other methods, and the "arbitrary" Vernet input is chosen so that the frequencies it actually involves are
those in which one is interested when studying the transfer function. These methods are not, therefore, automatic; no technique dispenses with the process of thought.

8.5. RELATION BETWEEN AMplitude AND PHASE

8.5.1. Introduction. When one has acquired practical experience with frequency response, he is inclined to consider almost exclusively either the amplitude response (important parameter is the resonant peak) or the phase response (important parameter is the slope of the phase curve). We will now state this question more precisely and explain why it is necessary, in practice, in order to characterize a linear system, to consider at the same time its amplitude and phase response.

8.5.2. Bode's Problem. 1. The logarithmic amplitude and the phase of the transfer function $H(s)$ of a linear system may be considered as the real and imaginary parts of the analytic function $\log H(s)$. Hence, they are not independent; there exists between them a relationship (Laplace equation) which defines the form of one when the other is known. The question then arises whether or not it is really necessary, when characterizing a linear system, to know at the same time its amplitude and phase response, or if one of them is enough.

2. The problem was stated in the following manner by Bayard, and later by Bode: If the amplitude response $A(j\omega)$ of a linear system is known, is it possible to deduce its phase response $\Phi(j\omega)$?

3. The answer is as follows:

   a. There are an infinite number of systems $S_0, S_1, S_2, \ldots$ with transfer functions $H_0, H_1, H_2, \ldots$ that have the same amplitude response $A(j\omega)$. They differ only by factors of the form

   $$\frac{1' - as}{1 + as}, \quad a > 0$$

   Such factors are called all-pass or pure phase-shift elements, because their transfer function for $s = j\omega$ always has a modulus of 1 but a phase that varies from 0 to $-\pi$ as $\omega$ increases from zero to infinity.

   b. One of these systems, $S_0$, has no pure phase-shift factor in its transfer function, which consequently does not have zeros with positive real parts. It can be shown that this system is unique.

   From the properties of the pure phase-shift factor, it is evident that, for the same frequency $\omega$, the phase of $S_0$ is larger than those of systems $S_1, S_2, \ldots$; we may also say, if we call phase lag a negative phase shift, that its phase lag is smaller than those of systems $S_1, S_2, \ldots$.

4. It is now natural to introduce the following concepts:

   Minimum-phase-shift system. The transfer function of such a system does not have zeros with positive real parts.

   Nonminimum-phase-shift system. The transfer function of such a system has at least one zero with positive real part. The transfer function of such a system differs from that of a minimum-phase-shift system by the presence of one or more pure phase-shift factors.
5. The results from the studies of Bayard and Bode may then be expressed as follows: The amplitude response of a linear system defines without ambiguity a unique minimum-phase-shift system $S_0$, but all the nonminimum-phase-shift systems which differ from $S_0$ by a pure phase-shift element have the same amplitude response.

8.5.3. Minimum-phase-shift Systems, Bode's Equation. It follows from the above that the phase $\Phi_e$ of a minimum-phase-shift system at a frequency $\omega_e$ is perfectly determined by its amplitude response $A(j\omega)$. The equation that gives $\Phi_e$ as a function of $A(j\omega)$ is called Bode's equation.\(^1\) An expression of it is

$$\Phi_e = \frac{2\omega_e}{\pi} \int_{-\infty}^{+\infty} \frac{\ln A(\omega) - \ln A(\omega_e)}{\omega^2 - \omega_e^2} \, d\omega$$

where $\Phi_e$ is expressed in radians and $\ln$ refers to natural logarithms.

In spite of the interest of these questions, and the elegance of these methods, we will not discuss them here. It is not necessary; in our opinion, that the ordinary engineer working on servo systems be highly versed in these methods. Their prime importance lies in quite advanced work. The reader interested in this matter may refer to Bode's fundamental work.\(^2\) The only point on which we will insist is that Bode's equation, in practice, does not authorize characterizing a linear system by its amplitude response or phase response alone (except in rather special cases of electric networks). There are three reasons for this.

1. The accuracy of $\Phi(j\omega)$ as obtained by Bode's equation from $A(j\omega)$ will, in practice, decrease when the frequency becomes high.
2. Bode's equation applies only to minimum-phase-shift systems. Often, one is not sure that a given physical system fits this category.
3. It is very seldom, except for electric networks, that a system is perfectly linear within the limits of experimental accuracy. Linear analysis implicitly replaces the actual system by a linear model, which is to represent the system as well as possible on the basis of the available experimental data. Obviously, the linear model will approximate the actual system more closely, and subsequent linear analyses will be more valuable, if the experimental data are more complete. Thus, considering only the experimental amplitude (or phase) response and obtaining the other by Bode's equation amounts to deliberately rejecting half the information one possesses concerning the system; as a result, subsequent linear analyses of it are less realistic.

8.5.4. Nonminimum-phase-shift Systems. We shall deal here only with nonminimum-phase-shift systems the transfer functions of which have only one pure phase-shift factor. They are by far the most common. The behavior of such systems with nonminimum phase shift possesses one characteristic property: it is faulty at the outset. To understand what is meant by this expression, consider a system, the output of which must

\(^1\) In Europe, the Bayard-Bode law. First expression of it was given by Hilbert.
reproduce the input, subjected to a step input from 0 to 1. The output goes from 0 to 1 (to within a gain factor) through the medium of a more or less complicated transient state which may or may not be oscillatory (see Fig. 7-4). In the case of a minimum-phase-shift system, the output is positive after the abrupt variation of the input [specifically, \((dr/dt)(0^+) > 0\); or, if \((dr/dt)(0^+)\) is zero, \((d^2r/dt^2)(0^+) > 0\)], the first movement of the system being to respond in the direction of the command.

In the case of a nonminimum-phase-shift system, however, the output *first decreases* before it follows the direction of the input (Fig. 8-26). This property is easily shown from the mathematical form of the transfer function. The response to a step input, for a minimum-phase-shift system, is given by the following equation (when initial conditions are zero):

\[
R(s) = \frac{1}{s} \frac{A_m s^m + \cdot \cdot \cdot + A_1 s + A_0}{B_n s^n + \cdot \cdot \cdot + B_1 s + B_0}
\]

where \(B_n\) and \(A_m\) are positive, \(B_n\) because the system is assumed to be stable, \(A_m\) because it is a minimum-phase-shift system. The application of the initial-value theorem shows immediately that the first nonzero derivative of \(r(t)\) is positive.

On the other hand, the presence of a pure phase-shift factor will give

\[
R(s) = \frac{1}{s} \frac{A_m s^m + \cdot \cdot \cdot + A_0}{B_n s^n + \cdot \cdot \cdot + B_0} \frac{1 - as}{1 + as}
\]

with \(a > 0\) and assuming \(A_m\) and \(B_n > 0\). It is easily seen in the same manner that the first nonzero derivative is negative this time. This faulty behavior at the outset is expressed by a supplementary delay which can be considered, to a first approximation, to be of the order of \(T\) in Fig. 8-26.

**8.5.5. Examples.** 1. Some electrical networks are nonminimum-phase-shift systems. Those most frequently encountered in practice are bridged-T circuits; for example, the network shown in Fig. 8-27 has a zero transfer function for \(\omega = 1/SRC\).
2. Certain aeronautical examples are among the most instructive. For instance, an aircraft of the usual type undergoing small longitudinal movements is a non-minimum-phase-shift system if the vertical speed $w$ of the center of gravity is taken as output. The reason is that, when the elevator is moved in order to make the aircraft rise, its first movement is to sink, because the first effect of the control-surface deflection is to decrease the lift. This fact can be proved by eliminating $\theta(t)$ and $i(t)$ from the equations obtained in Sec. 7.52. In the transfer function $W(s)/\Delta(s)$ thus obtained, the numerator has a constant negative term and hence possesses a zero with a positive real part.

3. An aircraft in lateral motion, if one considers the roll angle produced by a rudder deflection, is also a nonminimum-phase-shift system, provided the rudder is above the center of gravity of the aircraft. The reason is that the first effect of a rudder deflection is to bank the aircraft in that direction [(a) in Fig. 8-29], opposite to the roll corresponding to the desired turn. On the other hand, the aircraft will behave as a minimum-phase-shift system if the rudder is below the center of gravity.

4. Conclusion. Because of their faulty behavior at outset, nonminimum-phase-shift systems introduce delays that are more difficult to compensate than delays caused by ordinary lags. Therefore, it is necessary that the presence of nonminimum-phase-shift elements be avoided at any cost in fast servo systems; for example, a longitudinal autopilot should detect the pitch angle or the angle of attack, but not the vertical velocity.

8.5.6. Practical Conclusions. 1. To characterize a linear system, it is necessary, in practice, to know at the same time both its amplitude and its phase frequency response.

2. One should distrust nonminimum-phase-shift systems, and as a general rule eliminate them from fast-response servo systems.

3. Bode's equation should never be applied before it has been established that the system is of the minimum-phase-shift type. In particular, properties of linear systems developed from this relation should not be presented as general.

---

8.6. SYNTHESIS: CONCLUSION ON THE HARMONIC APPROACH

8.6.1. A Survey of the Harmonic Approach. It is now possible to set out systematically the harmonic approaches which have been studied in the preceding chapters (Table 8-2).

<table>
<thead>
<tr>
<th>Table 8-2. Linear systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential equations</td>
</tr>
<tr>
<td>Laplace transform</td>
</tr>
<tr>
<td>Algebraic equations</td>
</tr>
<tr>
<td>Frequency response</td>
</tr>
<tr>
<td>Transfer function</td>
</tr>
<tr>
<td>Transfer locus</td>
</tr>
</tbody>
</table>

If a linear system is defined by its differential equation, use of the Laplace transform reduces solution to that of an algebraic problem. The concepts of frequency response and transfer function then enable use of graphic methods (transfer locus).

Conversely, if the equations of the system under consideration are unknown, the frequency-response technique enables one to check its linearity and translate its properties into terms of a transfer locus. Hence, equations and experiments tend to achieve the same result.

8.6.2. Conclusions. It has been shown that a linear system can be completely characterized by (a) its differential equation, (b) its transfer function, or (c) its transfer locus. The transfer locus (harmonic approach) is of special interest because it is possible to draw the transfer locus of a physical system whose equations are not known, as well as the transfer locus of a system which is only defined on paper by its equations. The same techniques apply to both cases; experimentally and mathematically, the form of expression is the same. This is rare in the field of engineering as a whole. The importance of this fact in the servo field is considerable.
CHAPTER 9
THE POLE-ZERO-CONFIGURATION APPROACH.
STABILITY

Summary
1. The pole configuration and stability.
2. Algebraic stability criteria.
3. The pole-zero configuration.
4. Conclusion on the pole-zero approach.

9.1. THE POLE CONFIGURATION AND STABILITY

9.1.1. The Concept of Stability. Roughly speaking, a system is said to be stable when it tends to return to its steady-state condition after a disturbance. It is said to be unstable if it tends to diverge from its steady-state condition after a disturbance. This definition is quite sufficient so far as linear systems are concerned. In such systems, a momentary disturbance produces, in stable cases, a damped transient response; in an unstable case, the response becomes infinite. (Actually, for the latter, the output goes beyond the limit within which the linear equations are valid.)

If more general systems are to be considered, the following more precise definition should be given for the concept of stability. First, consider a system with one input $e$ and one output $r$, and suppose that there exists a steady state (equilibrium position) which will be taken as $e = 0$, $r = 0$. If the system is disturbed from its steady-state position and left to itself at time $t = 0$ with the initial condition $e(0) = e_0$, a response $r(t)$ will result. The system is said to be stable if, given any positive number $\varepsilon$, it is possible to find a positive number $\eta$ such that, if $e_0$ complies with $e_0 < \eta$, one is sure that, for any $t > 0$, $|r(t)| < \varepsilon$. In the contrary case, that is, if there are positive numbers $\varepsilon$ such that it is impossible to find a corresponding $\eta$ and satisfy the above inequalities, the system is said to be unstable.

Note that stability does not demand that $r(t)$ approach zero as $t$ tends to infinity. If this additional condition is fulfilled, the system is said to be not only stable but asymptotically stable. Some systems can be stable without being asymptotically stable.

More generally, consider a system with $n$ inputs $e_1$, $e_2$, ..., $e_n$ and $m$ outputs $r_1$, $r_2$, ..., $r_m$ and suppose that $e_1 = e_2 = \cdots = e_n = 0$, $r_1 = r_2 = \cdots = r_m = 0$ is a steady-state position. This system is said to be stable if, given any positive number $\varepsilon$, it is possible to find a positive number $\eta$ such that, when the system is left to itself at $t = 0$ with initial conditions complying with

$$e_1^2(0) + e_2^2(0) + \cdots + e_n^2(0) < \eta$$

159
the outputs will satisfy for all positive values of $t$

$$r_1^2(t) + r_2^2(t) + \cdots + r_n^2(t) < \epsilon$$

If, moreover, $r_1(t), r_2(t), \ldots, r_m(t)$ approach zero for infinite $t$, the system is said to be **asymptotically stable**. If, for certain values of $\epsilon$, it is not possible to find a number $\eta$ such that the above two inequalities are satisfied, the system is said to be **unstable**.

**9.1.2. Fundamental Condition for Stability.** The response $r(t)$ to an input $e(t)$ of a **linear** system with a transfer function $H(s) = P(s)/Q(s)$ is given by

$$R(s) = \frac{P(s)}{Q(s)} \times E(s) + \frac{s(s)}{Q(s)}$$

where $s(s)$ is a function which depends on the initial conditions and becomes zero when the latter are zero (see, for example, Secs. 6.2.1 and 7.3.4).

Now, consider a system which is subjected to a zero input, that is, $E = 0$. The partial-fraction expansion of the rational fraction $s(s)/Q(s)$ gives:

a. Terms of the form $A/(s - c)$ which correspond to the real zeros $c$ of the polynomial $Q(s)$

b. Terms of the form

$$A/[(s - a)^2 + b^2]$$

which correspond to the complex zeros of $Q(s)$, of the form $a \pm jb$

It is apparent that the response $r(t)$ is the summation of functions of time which are of the form (a) $e^{at}$ for the real zeros and (b) $e^{at} \sin(bt + \varphi)$ for the complex zeros.

There are, consequently, two cases which must be considered:

**Case 1.** The exponentials $e^{at}$ and $e^{bt}$ decrease with time. The system is then stable. This is true only if all the zeros of $Q(s)$ have negative real parts.

**Case 2.** At least one of the exponentials, $e^{at}$ or $e^{bt}$, increases with time. This is true whenever one zero of $Q(s)$ has a positive real part.

By noting the fact that the zeros of $Q(s)$ are the poles of $H(s)$, the
following theorem can be stated: A linear system is stable provided that all the poles of its transfer function have negative real parts. This is a necessary and sufficient condition. This theorem enables the stability of a system to be determined from a simple inspection of the position in the complex plane of the poles of the transfer function of the system (Fig. 9-1).

The particular case in which \( H(s) \) has two purely imaginary poles corresponds to an undamped sinusoidal response, and the system is said to be just oscillatory. (It is stable but not asymptotically stable.)

### 9.1.3. Concept of Modes.

A real pole, or a pair of conjugate imaginary poles, of \( H(s) \) constitutes what is called a "mode" of the corresponding system, the real pole being called a nonoscillatory mode and the pair of complex conjugate poles an oscillatory mode.\(^1\) The transient response of the system—for example, its step or impulse response—consists of time functions related to the modes.

Consider, for instance, a system for which the poles of the transfer function are \(-0.8 \) and \(-2.5 \pm 2j\) (Fig. 9-2). This system possesses two stable modes, one being nonoscillatory and the other oscillatory.

The Laplace transform of the unit-impulse response of the system will be of the form

\[
H(s) = \frac{F(s)}{(s + 0.8)[(s + 2.5)^2 + 4]}
\]

where the degree of \( F(s) \) cannot exceed three (Sec. 8.1.1, par. 4). Thus, \( h(t) \) will be of the form

\[
h(t) = Ae^{-0.8t} + Be^{-2.5t} \sin(2t + \varphi)
\]

Similarly, the unit-step response \( q(t) \) will be given by

\[
\frac{1}{s} H(s) = \frac{F(s)}{s(s + 0.8)[(s + 2.5)^2 + 4]}
\]

\[
q(t) = A + Be^{-0.8t} + Ce^{-2.5t} \sin(2t + \varphi)
\]

The \( e^{-0.8t} \) term is a damped exponential with a time constant of 1.25 sec. After 5 sec it has decayed to 2 per cent of its initial amplitude. The sine term is an oscillation of period \( 2\pi/2 = 3.1 \) sec, multiplied by an exponential with a 0.4-sec time constant. It has decayed to 2 per cent of its initial value after 1.6 sec, that is, approximately after a half-period. Accordingly, practically no oscillation will be visible in the plot of the transient response.

Thus a mere glance at the pole configuration in the Laplace plane (or s plane) enables one, without having to perform any actual computation, not only to see that the system is stable, but to state some properties of the transient:

a. Practically, no oscillation is visible, in spite of the presence of an "oscillatory" mode.
b. The transient has died out (within 2 per cent) after 5 sec.

The relative importance of the two terms of the transient (comparison of the coefficients \( B \) and \( C \)) depends on the zeros of \( F(s) \) (see Sec. 9.3.5, example 2).

\(^1\) The word mode does not have the same meaning here that it has in the dynamics of mechanical systems with many degrees of freedom.
9.1.4. Stability Margin. In order to guarantee that a linear system will be stable, it is necessary that the poles of its transfer function should not lie in the right-hand half of the complex plane (Fig. 9-3). This restriction, however, does not necessarily ensure that the transient response of the system will be satisfactory; for a slow transient or a poor damping can occur if some poles lie too near the imaginary axis. From this fact has arisen the practice of restricting the operating region to a particular area of the complex plane to ensure good stability. This can, in fact, be done in two ways:

1. To prevent the occurrence of a slow transient, systems are so designed that no roots fall within a particular region of the complex plane that is limited by a parallel to the imaginary axis (Fig. 9-4). This

![Fig. 9-3. Guarantee of pure stability.]

![Fig. 9-4. Absolute stability margin: all exponential terms decay faster than $e^{-\alpha t}$.]

![Fig. 9-5. Relative stability margin: all quadratic factors are damped more than $\arcsin \psi$.]

is termed an imposition of absolute stability margin. Thus, if the region to the right of $\text{Re} (s) = -3$ is that which is considered prohibitive from the design standpoint, it is certain that all the exponentials will decrease with time faster than $e^{-3t}$ and that the transient will be completed to within 2 per cent of its final value after $4 \times 0.33 = 1.3$ sec.

2. To ensure a well-damped transient, that area of the complex plane that lies between the imaginary axis and the two straight lines passing through the origin, as shown in Fig. 9-5, is also restricted. This is sometimes termed imposition of relative stability margin. In fact, the damping factor of a damped sinusoid is equal to the sine of the angle $\psi$ shown in Fig. 9-5 (see also Sec. 6.5.3, case 2). In particular, if $\psi > 45^\circ$, oscillations do not appear. This was the case of the example in Sec. 9.1.3.

9.1.5. Remarks. In spite of the fact that knowledge of poles of the transfer function of a system gives information regarding the mathe-
matical nature, duration, and damping of each of the terms of the transient, it does not make possible the plotting of the transient. In fact, in the example of Sec. 9.1.3, the shape of the transient depends on the values of the coefficients of \( A, B, \) and \( C \) and on the value of \( \varphi \). Each of these values depends on the numerator of the Laplace transform of the response, i.e., on the zeros of the transfer function of the system.

Finally, a knowledge of the poles and zeros of the transfer function of a linear system completely determines its transient, the poles giving particular information regarding the stability of the system.

### 9.2. ALGEBRAIC STABILITY CRITERIA

#### 9.2.1. General

When the transfer function of a linear system is given, the general condition for stability obtained in Sec. 9.1.2 enables one to see whether or not the system is stable provided the poles of the transfer function have been found, that is, provided the characteristic equation has been solved. However, equations of the fourth or higher degree usually are not solved without considerable labor. Hence it is desirable to study the stability without actually solving the characteristic equation. This is possible by making use of stability criteria which enable one to settle the question of stability from the coefficients of the characteristic equation.

#### 9.2.2. A Necessary Condition for Stability

If the characteristic equation is written in the form

\[
A_n s^n + A_{n-1} s^{n-1} + \cdots + A_1 s + A_0 = 0
\]

it is necessary for all the coefficients \( A_i \) to have the same sign in order that all the roots lie in the left half plane. In particular, if the equation is written with \( A_n = 1 \), it is necessary that \( A_{n-1}, \ldots, A_1 \) and \( A_0 \) all be positive.

This condition is necessary. Therefore, if \( A_n = 1 \) and if one (or more) of the coefficients is negative, one is sure that the system is not stable. But this condition is not sufficient: if all the coefficients are positive, one is not sure that the system is stable.

**Example.** Consider a system with the characteristic equation

\[
s^2 + (\lambda + 1)s^2 + (\lambda + \mu - 1)s + \mu - 1 = 0
\]

which depends on two real parameters \( \lambda \) and \( \mu \). For the system to be stable, it is necessary that

\[
\lambda > -1 \quad \lambda + \mu > 1 \quad \mu > 1
\]

that is, the \((\lambda, \mu)\) point should not lie at the left of the \( \lambda = -1 \) line (Fig. 9-6a), nor below the \( \lambda + \mu = 1 \) line (Fig. 9-6b), nor below the \( \mu = 1 \) line (Fig. 9-6c). Therefore, if the \((\lambda, \mu)\) point lies in the shaded area shown in Fig. 9-6d, the system is unstable. If the \((\lambda, \mu)\) point lies in the nonshaded area, the system may be stable or unstable.

#### 9.2.3. Routh's Criterion

Sets of necessary and sufficient conditions for stability were stated by E. J. Routh and by A. Hurwitz.\(^1\) Routh's

and Hurwitz’s conditions furnish much the same information. Hurwitz’s criterion is more commonly used in Germany and in Russia, Routh’s in the United States, England, and France. Routh’s criterion can be stated as follows: Let the characteristic equation be written

\[ A_n s^n + A_{n-1} s^{n-1} + \cdots + A_1 s + A_0 = 0 \]

in which the \( A \)'s are real coefficients, with \( A_n \) positive (e.g., equal to unity) and \( A_0 \neq 0 \) (that is, any zero root is assumed to have been removed). Arrange the coefficients in two rows as

\[
\begin{array}{c|cccc}
  s^n & A_n & A_{n-2} & A_{n-4} & \cdots \\
  s^{n-1} & A_{n-1} & A_{n-3} & A_{n-5} & \cdots \\
\end{array}
\]

The coefficients of a third row are obtained by cross multiplication as follows:

\[
\begin{array}{c|cc}
  s^{n-2} & A_{n-1} A_{n-2} - A_n A_{n-3} & A_{n-1} A_{n-4} - A_n A_{n-5} & \cdots \\
  & A_{n-1} & A_{n-1} & \cdots \\
\end{array}
\]

The coefficients of a fourth row are obtained again by cross multiplication, using the second and the third rows and so forth until \( (n + 1) \) rows are obtained.\(^1\)

If all the terms in the first column are positive, the system is stable. If there are \( c \) changes of sign, the equation has \( c \) roots with positive real parts.

\(^1\) In this procedure it is permissible to multiply or divide the coefficients of any row by any positive number in order to simplify the numerical work. This does not alter the result.
Fig. 9-7. Routh conditions for stability. Shaded areas are unstable. Nonshaded area in (d) is stable.

**Example 1.** Let the characteristic equation be

\[ 2s^6 + 4s^5 + s^4 - 32s^3 + 51s^2 + 3s - 15 = 0 \]

Following the procedure presented above, the table of coefficients is

\[
\begin{array}{cccc}
s^6 & 2 & 1 & 51 & -15 \\
s^5 & 4 & -32 & 3 & \\
s^4 & 1 & 2.9 & -0.88 & \text{After dividing by 17} \\
s^3 & -1 & 0.15 & & \text{After dividing by 43.6} \\
s^2 & 1 & -0.29 & & \text{After dividing by 3.05} \\
s^1 & 0.44 & & & \\
s^0 & -0.29 & & & \\
\end{array}
\]

There are three unstable roots.

**Example 2.** Consider again the characteristic equation

\[ s^3 + (\lambda + 1)s^2 + (\lambda + \mu - 1)s + \mu - 1 = 0 \]

The table of coefficients is

\[
\begin{array}{c}
s^3 & 1 & \\
s^2 & \lambda + 1 & \mu - 1 \\
s^1 & \lambda(\lambda + \mu) & \\
s^0 & \lambda(\lambda + \mu)(\mu - 1) & \\
\end{array}
\]

The necessary and sufficient conditions for stability are

\[ \lambda + 1 > 0 \quad \lambda(\lambda + \mu) > 0 \quad (\mu - 1) > 0 \]

In the \((\lambda,\mu)\) plane these conditions are equivalent to excluding the shaded areas shown in Fig. 9-7a to c. Thus it is seen that the system is stable if (and only if) the two following conditions are met:

\[ \lambda > 0 \quad \mu > 1 \]

that is, if the \((\lambda,\mu)\) point lies in the nonshaded area of Fig. 9-7d.
9.2.4. Particular Cases. There are two exceptions to this general process:

1. If the first term in any row is zero but the remaining terms are not zero, the equation under study should be multiplied by a factor \((s + a)\) with a real and positive, which has the effect of restoring the missing power of \(s\).

2. If all the coefficients in a row are zero, this indicates that there are two equal and opposite roots, in particular, two conjugate roots which are purely imaginary; i.e., the system is just oscillatory. The process can be continued by writing in place of that row the coefficients of the derivative of an auxiliary polynomial the coefficients of which are the terms of the last nonzero row. The mentioned roots with zero real part are zeros of the auxiliary polynomial.

The latter case is important because it enables one to find the conditions for a linear system to be oscillatory. It will be applied later to constructing root loci of linear servo systems (Sec. 14.2.2, rule 4).

Example 1. Consider the characteristic equation

\[ s^4 + s^3 + 5s^2 + 4s + 4 = 0 \]

The table of coefficients is

\[
\begin{array}{llll}
s^4 & 1 & 5 & 4 \\
s^3 & 1 & 4 & \\
s^2 & 1 & 4 & \\
s^1 & 0 & 0 & \text{Auxiliary polynomial is } s^2 + 4 \\
s^0 & 8 & \\
\end{array}
\]

There are two roots with zero real parts \((s = \pm 2j)\), the other roots have positive real parts.

Example 2. Choose the real parameter \(\lambda\) so that the system whose characteristic equation is

\[ s^4 + \lambda s^3 + (\lambda + 4)s^2 + (\lambda + 3)s + 4 = 0 \]

will be oscillatory. The table of coefficients is

\[
\begin{array}{llll}
s^4 & 1 & & \lambda + 4 & 4 \\
s^3 & \lambda & & \lambda + 3 & \\
s^2 & \lambda^2 + 3\lambda - 3 & & 4\lambda & \\
s^1 & \lambda^2 + 2\lambda^2 + 6\lambda - 9 & & 0 & \\
\end{array}
\]

The terms of the second (or third) row cannot be zero at the same time. The fourth row will vanish if

\[ \lambda^2 + 2\lambda^2 + 6\lambda - 9 = 0 \]

whence the condition \(\lambda = 1\).

9.2.5. Remark. The Routhian conditions for stability are necessary and sufficient, but application becomes cumbersome when the degree of the characteristic equation is high. On the other hand, the conditions that all the coefficients of the characteristic equation must be of one sign are very easy to write, but they are not sufficient. Therefore, “conditions for stability” that are sometimes stated are a mixture of some \(A_i > 0\) conditions and some Routhian conditions. Such “conditions” are of
interest only if they are necessary and sufficient. The following are such necessary and sufficient conditions applying to systems of third, fourth, and fifth order.¹

Third order: \[ s^3 + As^2 + Bs + C = 0 \]
\[ A > 0 \quad B > 0 \quad C > 0 \quad AB - C > 0 \]

Fourth order: \[ s^4 + As^3 + Bs^2 + Cs + D = 0 \]
\[ \begin{cases} A > 0 & B > 0 & C > 0 & D > 0 \\ \text{The last Routh coefficient} > 0 \end{cases} \]

Fifth order: \[ s^5 + As^4 + Bs^3 + Cs^2 + Ds + E = 0 \]
\[ \begin{cases} A > 0 & B > 0 & C > 0 & D > 0 & E > 0 & AB - C > 0 \\ \text{The last Routh coefficient} > 0 \end{cases} \]

9.2.6. Structural Stability. Consider the closed-loop system shown in Fig. 9-8. It consists of a servomotor that applies a torque \( \Gamma \) to a pure inertia \( J \). The torque is proportional to the error but lags the error signal by a time constant \( T \). The equations are, after Laplace-transforming,

\[ \frac{R}{\Gamma} = \frac{1}{Js^3} \quad \frac{\Gamma}{\varepsilon} = \frac{C}{1 + Ts} \]

whence the output/error transfer function is, \( C, J, \) and \( T \) being inherently positive

\[ KG(s) = \frac{R}{\varepsilon} = \frac{C/J}{s^2(1 + Ts)} \]

The application of Routh’s criterion to the characteristic equation \( Ts^2 + s^3 + C/J = 0 \) shows that the system is unstable, whatever the parameters \( T, C, \) and \( J \).

Systems that are always unstable, whatever the values of their parameters, are called structurally unstable systems by Russian authors. Conditions for the structural stability of closed-loop systems can be obtained by applying Routh’s criterion to their characteristic equation \( 1 + KG(s) = 0 \) or, better, by more direct methods. The following results, obtained by M. Ajzerman,² are of interest.


Case 1. Servo system in which the numerator of the output/error transfer function is a constant:

\[ KG(s) = \frac{K}{s^i \left( 1 + \cdots + A_n s^{n-i} \right)} \]

where \( n \) is the order of the system and \( i \) is the number of integrations. The conditions for such a system not to be structurally unstable are

\[ i + u \leq 1 \quad n > 4o \]

in which \( u \) and \( o \) are the numbers of poles of \( KG(s) \) with positive and zero real parts, respectively, that is, of the unstable and oscillatory poles of the open-loop system. These conditions are necessary and sufficient.

Thus it is seen that, for the case in which \( G \) has a numerator equal to unity, the presence of two integrations (as in the case of Fig. 9-8), or of one integration and of an unstable mode in the forward path, causes the system to be structurally unstable. The presence of purely imaginary poles of \( KG \) demands that the degree of the system be sufficiently high \((n > 4o)\) in order that the system be stable.

Case 2. The numerator of \( KG(s) \) incorporates factors of the \( 1 + As + Bs^2 \) type \((A \geq 0, B \geq 0)\) [called lead factors because the phase of \( 1 + A(j\omega) + B(j\omega)^2 \) is positive]:

\[ KG = \frac{K}{s^i} \frac{1 + As + Bs^2}{Q_1(s)} \]

where \( Q_1(s) \) is a polynomial of the \((n - i)\)th degree with \( Q_1(0) \neq 0 \).

In this case, if \( Q_1(s) \) has \( r \) positive real roots and \( c \) complex roots with positive or zero real part, and if \( \gamma \) is the highest integer that is \( \geq c/2 \), the conditions for the system not to be structurally unstable are the following:

a. If \( A > 0, B = 0 \); that is, if the lead factor is of first order,

\[ n > 4\gamma - 3 \quad \text{if } i + r = 0 \text{ or } 2 \]

\[ n > 4\gamma \quad \text{if } i + r = 1 \]

b. If \( A > 0, B > 0 \); that is, if the lead factor is of second order,

\[ n > 4\gamma - 3 \quad \text{if } i + r = 0 \text{ or } 2 \]

\[ n > 4\gamma - 2 \quad \text{if } i + r = 1 \text{ or } 3 \]

c. For the particular case \( A = 0, B > 0 \):

\[ n > 4\gamma - 2 \quad \text{if } i + r = 0 \text{ or } 1 \]

Comparison with Case 1 shows the stabilizing effect of lead factors, since the introduction of such factors can stabilize a structurally unstable system. Also, these conditions show the limitations of stabilization by a lead factor: a first-order lead factor cannot stabilize a system with more than two integrations, and, for unstable roots, a second-order lead factor cannot compensate for more than three.

9.2.7. Limitations in the Use of Algebraic Criteria. Altogether, the application of algebraic stability criteria is limited to a rather small number of problems, for the following reasons:

1. The above criteria settle only the question whether or not all the roots of the characteristic equation lie in the left half plane. They give no indication of where these roots lie, which is a question of considerable importance for evaluating the performance of the system.
2. When the degree of the system is high, Routh's conditions become complicated expressions which are hard to interpret in terms of the parameters of the problem.

3. For these two reasons, algebraic criteria do not lend themselves easily to the synthesis of complex systems.

4. Finally, algebraic criteria can be applied only when the mathematical expression for the characteristic equation is known. This makes it impossible to make direct use of experimental response data.

9.3. THE POLE-ZERO CONFIGURATION

9.3.1. Geometrical Interpretation of \( H(s) \) in the Laplace Plane. Let a transfer function \( H(s) \) be written in the following factored form to illustrate its \( m \) zeros \( z_1, z_2, \ldots \) and its \( n \) poles \( p_1, p_2, \ldots \)

\[
H(s) = k \frac{(s - z_1)(s - z_2) \ldots}{(s - p_1)(s - p_2) \ldots}
\]

In this expression the coefficient \( k \) has no physical significance (Sec. 7.1.5) and is related to the gain \( K \) by Eq. (7-4).

Let \( Z_1, Z_2, \ldots, P_1, P_2, \ldots \) be the locations in the \( s \) plane (or Laplace domain) of the points represented by the complex numbers \( z_1, z_2, \ldots, p_1, p_2, \ldots \). If \( M \) is the point represented by the complex number \( s = \alpha + j\omega \), the expression for \( H(s) \) shows that its value for \( s = \alpha + j\omega \); that is, its value at the point \( M \), is

\[
H(s) = k \frac{Z_1M \cdot Z_2M \cdots}{P_1M \cdot P_2M \cdots} \quad \quad (9-1)
\]

where

\[
|H(s)| = k \frac{Z_1M \cdot Z_2M \cdots}{P_1M \cdot P_2M \cdots} \quad \quad (9-2)
\]

\[
\chi H(s) = \chi Z_1M + \chi Z_2M + \cdots - \chi P_1M - \chi P_2M - \cdots \quad \quad (9-3)
\]

Thus, if the function \( H(s) \) is characterized by its poles and zeros, its value for any complex value of \( s \) can be determined geometrically (Fig. 9-9).

9.3.2. Application to the Frequency Response. In particular, let \( s = j\omega \), that is, let \( M \) be on the positive imaginary axis.

As an example, consider the case shown in Fig. 9-10. It is the pole-zero configuration derived in Sec. 7.6.1 for the transfer function of an airplane for which the pitch rate is now considered as output in order to eliminate the pole at the origin.
For the condition $\varepsilon = 0.075$ and $\omega_n = 20$ (see Fig. 8-3), and if in addition the values of $\varepsilon$ and $p$ are dimensionless, that is, $p$ is divided by $\omega_n$, then

\[ z_1 = -\alpha_1 = -0.05 \]
\[ p_1 = -\alpha_2 + j\omega_2 = -0.075 + j1.0 \]
\[ p_2 = -\alpha_2 - j\omega_2 = -0.075 - j1.0 \]

The following properties are easily seen:

a. For low $\omega$, $H(s)$ has a relatively small modulus, since the vector $Z_1M$ has a small modulus.

b. When $\omega$ approaches the region $\omega = \omega_2$, the vector $P_1M$ in the denominator of $H(s)$ goes through a minimum; as a result, $H(s)$ goes through a maximum. At the same time the phase of $H(s)$ changes rapidly, abruptly decreasing by $180^\circ$. This maximum of magnitude and abruptness of change in phase is known as the resonance due to the $-\alpha_2 \pm j\omega_2$ mode, the maximum being greater and the phase shift being more abrupt when the corresponding mode is poorly damped ($\alpha_2$ small).

c. As $\omega$ approaches infinity, the magnitudes of the three vectors $Z_1M$, $P_1M$, $P_2M$ approach infinity and become of the same order of magnitude. Therefore $H(s)$ approaches zero. This indicates a filtering out of high frequencies.

This example can easily be generalized, the frequency response $H(j\omega)$ being obtained from the pole-zero configuration of $H(s)$. In particular, it is possible to visualize the following:

1. The effect of poorly damped modes which exhibit resonance.
2. The effect of zeros lying near the imaginary axis which produce a dip in the magnitude-vs.-frequency curve.
3. The variation of phase vs. frequency when $\omega$ varies from zero to infinity. It can be seen that, if poles lie in the right-hand half plane, corresponding to unstable modes, they will result in phase shifts that are of the opposite direction to those due to poles lying in the left-hand half plane (Sec. 9.3.5).

9.3.3. Regular Systems. It is easily seen that, if some poles lie near the imaginary axis, their effect is predominant on the behavior of $H(j\omega)$. For example, if a system has two complex conjugate poles $-\alpha \pm j\omega$ near the imaginary axis, all other poles being much farther to the left (Fig. 9-11), it will behave like the second-order system

\[ H_1(s) = \frac{1}{(s + \alpha)^2 + \omega^2} \]
over a wide range of frequency. The poles \(-\alpha \pm j\omega\) are said to be the dominant, or the controlling, poles (also, the dominant mode) of the system.\(^1\)

Systems complying with the above assumptions are regular systems, as defined in Sec. 8.3.9.

9.3.4. Application to Steady State. The pole-zero configuration for a system provides information concerning its steady state. First, the presence or the absence of a pole at the origin immediately shows whether or not integration is involved in the system. Hence, it is immediately known whether the response of the system to a constant input ("position" input) will be a position, corresponding to no pole at the origin; a velocity, corresponding to a single pole at the origin; or an acceleration, corresponding to a double pole at the origin. It is possible, in addition, to extend this analysis further by finding the steady-state errors of the system from its pole-zero configuration.

Let a system \(H(s)\) have a unity static gain \(H(0) = 1\), so that it will have no position error. Its velocity and acceleration constants are given by the series expansion

\[
1 - H = \frac{s}{C_v} + \frac{s^2}{C_a} + \cdots
\]

Hence, from Maclaurin’s equation

\[
1 - H = 1 - H(0) - H'(0)s - \frac{H''(0)s^2}{2!} - \cdots
\]

\[
\frac{1}{C_v} = -H'(0) \quad \frac{1}{C_a} = -\frac{H''(0)}{2}
\]

where \(H'\) and \(H''\) are the first and second derivatives of \(H\) with respect to \(s\).

The derivatives of \(H\) in factored and second derivatives of \(H\) with respect to \(s\). For the velocity constant,

\[
\frac{H'}{H} = \frac{1}{s - z_1} + \frac{1}{s - z_2} + \cdots - \frac{1}{s - p_1} - \frac{1}{s - p_2} - \cdots
\]

and

\[
\frac{H'(0)}{H(0)} = H'(0) = -\frac{1}{z_1} - \frac{1}{z_2} \cdots + \frac{1}{p_1} + \frac{1}{p_2} + \cdots
\]

Therefore,

\[
\frac{1}{C_v} = \frac{1}{z_1} + \frac{1}{z_2} + \cdots - \frac{1}{p_1} - \frac{1}{p_2} = \sum_{i=1}^{m} \frac{1}{z_i} - \sum_{i=1}^{n} \frac{1}{p_i} \quad (9-4)
\]

Similarly, for the acceleration constant

\[
\frac{H''}{H} - \frac{H'^2}{H^2} = -\frac{1}{(s - z_1)^2} - \frac{1}{(s - z_2)^2} - \cdots
\]

\[
+ \frac{1}{(s - p_1)^2} + \frac{1}{(s - p_2)^2} + \cdots
\]

and
\[
H''(0) - H''z(0) = - \frac{2}{C_a} - \frac{1}{C_v^2} - \frac{1}{z_1^2} - \frac{1}{z_2^2} - \cdots + \frac{1}{p_1^2} + \frac{1}{p_2^2} + \cdots
\]

Therefore,
\[
- \frac{2}{C_a} = \frac{1}{C_v^2} + \sum_1^n \frac{1}{p_i^2} - \sum_1^m \frac{1}{z_i^2} \quad (9-5)
\]

Both constants are thus expressed in terms of the location of the zeros and poles.

9.3.5. Application to Transients. Let a system with a transfer function \( H(s) \) be subjected to a step input. The response will be
\[
R(s) = \frac{1}{s} H(s) = \frac{1}{s} \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots}
\]
or
\[
r(t) = A + A_1 e^{p_1 t} + A_2 e^{p_2 t} + \cdots
\]

where the \( A_n \) are the residues of the partial-fraction expansion of \( R(s) \):
\[
\frac{1}{s} \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots} = \frac{A}{s} + \frac{A_1}{s - p_1} + \frac{A_2}{s - p_2} + \cdots
\]
The residue \( A_1 \), for example, is obtained by multiplying both sides of the equation by \((s - p_1)\)
\[
\frac{1}{s} \frac{(s - z_1)(s - z_2) \cdots}{(s - p_2) \cdots} = \frac{A(s - p_1)}{s} + A_1 + \frac{A_2(s - p_1)}{s - p_2} + \cdots
\]
and setting \( s = p_1 \). Proceeding in this manner yields

\[
A_1 = \frac{1}{p_1} \frac{(p_1 - z_1)(p_1 - z_2) \cdots}{(p_1 - p_2)(p_1 - p_3) \cdots}
\]

which can be interpreted as

\[
A_1 = \frac{1}{OP_1 P_2 P_1 \cdots} \quad (9-6)
\]

The residue corresponding to a pole is the product of the vectors from all zeros to that pole divided by the product of the vectors from the other poles of \( R(s) \). The residues can be readily computed and, therefore, the transient response can be drawn from the pole-zero configuration (Fig. 9-12).
The same method applies to the impulse response. In the latter case \( R(s) = H(s) \), whence
\[
A_1 = \frac{Z_1 P_1 \cdot Z_2 P_1 \cdots}{P_2 P_1 \cdot P_2 P_1 \cdots}
\]

Example 1. Find the unit-step response \( q(t) \) for the system \( S_1 \) defined by
\[
H(s) = \frac{10}{(s + 0.5)(s + 1)^2 + 4}
\]
The poles are \( s = -0.5 \) and \(-1 \pm 2j\); there is no zero. The Laplace transform of \( q(t) \) has the form
\[
\frac{1}{s} H(s) = \frac{A}{s} + \frac{B}{s + 0.5} + \frac{C}{s + 1 - 2j} + \frac{C^*}{s + 1 + 2j}
\]
where \( C^* \) denotes the complex conjugate of \( C \).
The residues \( A, B, C \), as read from Fig. 9-13, are
\[
A = 10 \quad \frac{10}{0.5(1 - 2j)(1 + 2j)} = 4
\]
\[
B = -\frac{10}{(-0.5)(0.5 - 2j)(0.5 + 2j)} = -\frac{80}{17} = -4.7
\]
\[
C = \frac{10}{(-1 + 2j)(-0.5 + 2j)(2(2j))} = \frac{10}{2j} \frac{10}{(7^2 + 6^2)^{1/2}} \exp\left(-j \arctan \frac{6}{7}\right)
\]
whence
\[
r(t) = 4 - \frac{80}{17} e^{-0.6t} + \frac{10}{(7^2 + 6^2)^{1/2}} e^{-t} \sin\left(2t + \arctan \frac{6}{7}\right)
\]
\[
r(t) = 4 - 4.7e^{-0.6t} + 1.08e^{-t} \sin\left(2t + 40^\circ 30'\right)
\]

Example 2. Find the unit-step response \( q(t) \) for the system \( S_2 \) defined by
\[
H(s) = \frac{10(s + 0.6)}{s(s + 0.5)(s + 1)^2 + 4}
\]
The pole-zero configuration is shown in Fig. 9-14. The poles are the same as for system \( S_1 \), but a real negative zero \( s = -0.6 \) has been incorporated. The partial-
fraction expansion of \( Q(t) \) has the same form. The residues, as read from Fig. 9-14, are

\[
\begin{align*}
A &= \frac{(10)(0.6)}{0.5(1 - 2j)(1 + 2j)} = 2.4 \\
B &= \frac{10(-0.5 + 0.6)}{-0.5(0.5 - 2j)(0.5 + 2j)} = -\frac{8}{17} \\
C &= \frac{10(-0.4 + 2j)}{2j(-7 - 6j)} = -0.965 + j0.541
\end{align*}
\]

whence

\[
r(t) = 2.4 - 0.47e^{-0.6t} + 2.21e^{-t} \sin (2t - 60^\circ 40')
\]

The effect of the zero \( s = -0.6 \) has been to divide by 10 the term corresponding to the pole \( s = -0.5 \) that lies close to it. This accounts for the main difference between the transient responses of the systems \( S_1 \) and \( S_2 \).

**9.3.6. Frequency Response and Stability.** Consider a polynomial

\[
Q(s) = a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0
\]

Suppose all the \( a \)'s are real and that all the zeros \( z_1, z_2, \ldots, z_n \) lie in the left half plane. One has

\[
Q(j\omega) = a_n (j\omega - z_1)(j\omega - z_2) \cdots (j\omega - z_n)
\]

\[
\zeta Q(j\omega) = \zeta Z_1 M + \zeta Z_2 M + \cdots + \zeta Z_n M
\]

where \( M \) is any point \( j\omega \) on the imaginary axis.

Consider, in particular, two points \( M_1 \) and \( M_2 \) that correspond to \( -j\omega_1 \) and \( +j\omega_2 \), respectively. One has

\[
\zeta Q(j\omega_2) - \zeta Q(-j\omega_1) = \zeta Z_1 M_2 - \zeta Z_1 M_1 + \cdots + \zeta Z_n M_2 - \zeta Z_n M_1 \tag{9-7}
\]

If now \( \omega_1 \) and \( \omega_2 \) tend toward infinity, each difference of the right-hand side becomes equal to \( +\pi \). Therefore,

\[
\Phi(+\infty) - \Phi(-\infty) = n\pi
\]

or, equivalently, since \( Q(s) \) is a real polynomial and therefore the whole figure is symmetric with respect to the real axis,

\[
\Phi(+\infty) - \Phi(0) = n\frac{\pi}{2} \tag{9-8}
\]

If, on the contrary, \( n' \) of the \( n \) roots should lie in the right-hand half plane, \( n' \) terms of the right-hand side in Eq. (9-7) would be \(-\pi\) instead of \( \pi \), and hence

\[
\Phi(+\infty) - \Phi(-\infty) = (n - 2n')\pi
\]

\[
\Phi(+\infty) - \Phi(0) = n\frac{\pi}{2} - n'\pi \tag{9-9}
\]

If the \( Q(j\omega) \) locus is drawn in the complex plane, the above remarks enable one to draw conclusions on the stability of the zeros of \( Q(s) \) from the \( Q(j\omega) \) locus. In the case of stability the phase \( \Phi \) of \( Q(j\omega) \) increases continuously from 0 to \( n\pi/2 \) as the \( Q(j\omega) \) locus is traced out from \( \omega = 0 \) to
\( \omega = +\infty \). Conversely, the presence of unstable roots results in a net phase variation which is less than \( n\pi/2 \); furthermore, the phase may not always increase as \( \omega \) varies from 0 to \( +\infty \), but may decrease in certain intervals (near unstable zeros).

These properties are sometimes presented as a graphical stability criterion\(^1\) for the zeros of a polynomial \( Q(s) \). The criterion can be applied without knowing the values of the zeros of \( Q(s) \). In fact, if \( Q(j\omega) \) is written as

\[
Q(j\omega) = A(\omega^2) + j\omega B(\omega^2)
\]

the zeros of \( A(\omega^2) \) and \( B(\omega^2) \) correspond to the intersections of the \( Q(j\omega) \) locus with the imaginary and the real axes, respectively. It can thus be seen that the condition for stability is equivalent to the roots of the equations \( A(\omega^2) = 0 \) and \( B(\omega^2) = 0 \) being simple, real, positive, and alternating according to the sequence: zero of \( A \), zero of \( B \), zero of \( A \), etc.\(^2\)

It should be remembered that this criterion applies in the form just stated only to polynomials or to the inverse of polynomials. Therefore, it does not apply to transfer functions of the general type \( P(s)/Q(s) \). In Part 2 (Sec. 16.1) the Nyquist stability criterion will be developed. It is applicable to the transfer loci of feedback systems and is based on a similar idea.

9.3.7. The Hydraulic Analogy. The fundamental equation (9-1), or its equivalents Eqs. (9-2) and (9-3), can be visualized by making use of a hydraulic analogy.\(^3\) Given a transfer function \( H(s) \), the loci in the \( s \) plane of those points \( M(s = \alpha + j\beta) \) for which

\[
\text{argument } H(s) = \text{const}
\]

can be shown to be the streamlines of a flow generated by the presence of sources and sinks. The sources are located at the poles of \( H(s) \) and are characterized by a magnitude, i.e., an outflow proportional to the corresponding residues. The sinks are located at the zeros of \( H(s) \) and are characterized by a magnitude proportional to the corresponding residue. It is assumed that the flow is two-dimensional and the liquid has negligible viscosity. Figures 9-15 to 9-18 show typical simple examples that are easy to understand qualitatively.

It can be shown that the loci \( |H(s)| = \text{const} \) are orthogonal to the constant-argu-

\(^1\) This criterion is known to Russian authors as the \textit{A. Mikhailov criterion} ("Metod garmonicheskovo analiza v teorii regulirovaniya," \textit{Avtomatika i Telemehanika}, 3:27 (1938)). It is known in France as the \textit{Leonhard criterion}, in Germany as the \textit{Cremer-Leonhard criterion}, after Cremer and Leonhard, who formulated it independently of each other and of Mikhailov. The original references are A. Leonhard, "Neues Verfahren zur Stabilitätsuntersuchung," \textit{Arch. der Elektrotechnik}, 38:17–28 (1944); and L. Cremer, "Ein neues Verfahren zur Beurteilung der Stabilität linearer Regelungs-systeme," \textit{Z. fur angew. Mathematik und Mechanik}, 25–27:5(6) (1947).


In the hydraulic analogy they are lines of constant velocity or pressure. Other analogies can be developed. The electrostatic analogy is the best known. The poles and zeros are represented by positive and negative quantities of electricity, respectively. The streamlines become the lines of force of the electric field. The orthogonal curves are the isopotential lines. This has led to techniques using electrolytic tanks or graphite paper for the study of analytic functions. The reader interested in these methods is referred to five articles, with numerous bibliographical references, by T. J. Higgins.

9.4. CONCLUDING REMARKS ON THE POLE-ZERO APPROACH

As compared with the harmonic approach, the pole-zero approach has the following two advantages: (a) the pole configuration is immediately

---

1 This is obvious from the theory of functions of complex variables. The constant-magnitude and constant-phase loci correspond, by conformal mapping, to circles centered at the origin (|s| constant) and straight lines passing through the origin (argument s constant), respectively. The orthogonality of these circles and straight lines is preserved by conformal mapping, whence the property stated.


related to the stability of the system; (b) the pole-zero configuration enables one to study both the frequency response and the transient response. Conversely, the harmonic approach has the advantage of being based on the physical notion of frequency response and also of being applicable, therefore, to systems for which the differential equation is of very high order or is not known.

So far as feedback control systems are concerned, the pole-zero approach gives rise to the root-locus technique (Chap. 14) as opposed to the Hall or Nichols chart technique (Chap. 13) based on the harmonic approach. The respective domains of application of these two techniques resulting from the above are:

a. For very simple systems (second- and third-order systems), all methods are equivalent.

b. For less simple systems (say, third to seventh order) it is advisable to use the two approaches in parallel, as they often complement one another.

c. For more complicated systems (say higher than the seventh order) or for systems that involve components characterized by experimental data, the harmonic approach is the essential tool for the feedback-control engineer.
CHAPTER 10
TRANSFER MATRICES

Summary

1. Transfer matrices.
2. Application to cascade-system synthesis.
3. Application to impedance matching for maximum power transfer.

The concept of transfer function for a linear system \( S \) assumes that the system involves one input and one output, that is, is adequately represented by a two-terminal “black box,” sometimes called a dipole. This assumption is, however, an idealization. In particular, it overlooks the fact that the elements actuated by the system \( S \) play their part in the input-vs.-output relation; taking this into account leads to the consideration of the system \( S \) as a four-terminal network element, or quadripole. It is the purpose of this chapter\(^1\) to show:

a. How the concept of transfer function can be generalized by the concept of transfer matrix for a quadripole

b. How transfer matrices are helpful for cascade-system synthesis

c. How transfer matrices aid in the problem of impedance matching

10.1. TRANSFER MATRICES

10.1.1. Why Complements to the Concept of Transfer Function Are Necessary. 1. First Example. The electric network shown in Fig. 10-1 is made up of a resistance \( R \) and of a capacitance \( C \). It is often used in servo systems as a compensating network (Chap. 19) and is generally placed between the sensing device and the amplifier.

Let the input \( e(t) \) be the voltage across the two terminals \( A \) and \( B \) (Fig. 10-1) and the output \( r(t) \) the voltage across the terminals \( D \) and \( E \). The transfer function is obtained readily by applying the node equation and neglecting the grid current toward pole \( D \). Thus,

\[
\frac{r(t) - e(t)}{R} + C \frac{dr}{dt} = 0
\]

\(^1\) This chapter is an elementary introduction to the concepts of input functions and quadripoles. It may be omitted by readers already trained in these fields.
whence, assuming zero initial conditions,
\[
\frac{R(s)}{E(s)} = \frac{1}{1 + RCs}
\]
(10-1)

However, if two such networks are connected in cascade so that the output of the first is the input of the second (Fig. 10-2), it can be seen in the same way that the over-all transfer function is
\[
\frac{R(s)}{E(s)} = \frac{1}{(1 + R_1C_1s)(1 + R_2C_2s) + R_1C_2s}
\]
(10-2)

It is not the product
\[
\frac{1}{1 + R_1C_1s} \cdot \frac{1}{1 + R_2C_2s}
\]
as might be expected from Sec. 7.5.5.

It is possible to show more precisely whence the difference arises. When the transfer function of one of the networks was determined, the current flowing from D to E through the load resistance was supposed to be negligible; in other words, the load resistance (or, more generally, impedance) was assumed to be infinite. If this assumption does not hold and if Z is that load impedance, the circuit becomes that of Fig. 10-3 and the node equation gives for the expression of the transfer function:
\[
\frac{R(s)}{E(s)} = \frac{1}{1 + RCs + R/Z}
\]

This expression is identical to \(1/(1 + RCs)\) only if \(Z\) is infinite. This fact can be expressed by saying that, in determining the transfer function of a system, the load impedance is implicitly assumed to be infinite.

The question now arises, when two networks are connected in cascade, when is the over-all transfer function equal to the product of the two component transfer functions? Following from the above, it will be the case if the load impedance seen from the first network is very high. This happens, for the above example:

a. If \(R_2 \gg R_1\), as can be immediately checked from Eq. (10-2).

b. If a buffer amplifier is inserted between the two networks (Fig. 10-4), its grid current is very small, and hence provides the required high load impedance.
2. Second Example. Consider a field-controlled d-c motor. The command is the field current $i$. One can consider as the main output the angular velocity $\omega(t)$ (Fig. 10-5). Now the output $\omega(t)$ also depends on the load torque (i.e., the sum of those torques which act on the output shaft and which are different from the electric driving torque $\tau_m$). When the torque variations are negligible with respect to the driving torque $\tau_m$, the output is independent of the load and one can then speak of a transfer-function $\Omega(s)/I(s)$. However, in general, since $\omega(t)$ is a function both of $I$ and of the output load $C$, it is incorrect to speak of an $\omega$-vs.-$I$ transfer function.

Generalizing from the first example, it can be stated that, in writing the motor transfer-function $\Omega(s)/I(s)$, it is assumed that the motor mechanical output impedance $\Omega(s)/C(s)$ is infinite. Furthermore, the present example suggests that this impedance problem is related to energy considerations, since energy conditions relate $\omega$ and $\tau$, their product being the power developed by the motor.

3. Conclusion. In writing the equations of a system with one input and one output, certain conditions are implicitly assumed. They are equivalent to isolating the system from its physical environment. When these conditions are not fulfilled, one has to consider two variables (voltage and current for electric systems, angular velocity and torque for mechanical systems).

10.1.2. Quadripoles, Transfer Matrices. 1. General. Introductory Example. The previous chapters have dealt with systems possessing one input and one output. Such systems are represented by a two-terminal network and may, if they are linear, be characterized by a transfer function which relates the Laplace transforms of the input and the output (Fig. 10-6).

In the present chapter, the systems which will be considered possess two inputs (and two outputs) the product of which represents the input (or output) power: voltage and current for electric systems, velocity and force or angular velocity and torque for mechanical systems. Such systems with two inputs and two outputs are called quadripoles. They are symbolically represented by a rectangle with two terminals at the input and two at the output. If the systems are linear, the differential equations relating inputs and outputs (Fig. 10-7 for electric systems) are expressed in the form of a transfer matrix. Quadripoles are often repre-
sented in the form of Fig. 10-8; the terminals then have a physical meaning, whereas in Fig. 10-7 the arrows are pure abstractions. In the system represented in Fig. 10-8,

\[ V_e - V_r = RI_e \]

\[ I_e = C \frac{dV_r}{dt} + I_r \]

Laplace-transforming gives

\[ V_e = (1 + BC_s)V_r + RI_e \]

\[ I_e = C_sV_r + I_r \]

This is conventionally written in a more compact form as

\[
\begin{vmatrix}
V_e \\
I_e
\end{vmatrix} = \begin{vmatrix}
1 + RC_s & R \\
C_s & 1
\end{vmatrix} \times \begin{vmatrix}
V_r \\
I_r
\end{vmatrix}
\]

where the operator

\[
\begin{vmatrix}
1 + RC_s & R \\
C_s & 1
\end{vmatrix}
\]

is termed the inverse transfer matrix or, briefly, the matrix of the quadripole.

If the two networks shown in Fig. 10-7 are now connected in cascade (Fig. 10-9), one can write, calling \( V \) and \( I \) the intermediate variables,

\[ V_e = (1 + R_1C_{18})V + R_1I \]

\[ I_e = C_{18}V + I \]

\[ V = (1 + R_2C_{28})V_r + R_2I_r \]

\[ I = C_{28}V_r + I_r \]

whence, eliminating \( V \) and \( I \),

\[ V_e = [(1 + R_1C_{18})(1 + R_2C_{28}) + R_1C_{28}]V_r + [(1 + R_1C_{18})R_2 + R_1]I_r \]

\[ I_e = [C_{18}(1 + R_2C_{28}) + C_{28}]V_r + [R_1C_{18} + 1]I_r \]
The over-all inverse transfer matrix is

\[
\begin{vmatrix}
(1 + R_1 C_1 s)(1 + R_2 C_2 s) + R_1 C_2 s & (1 + R_1 C_1 s) R_2 + R_1 \\
C_1 s(1 + R_2 C_2 s) + C_2 s & R_2 C_1 s + 1
\end{vmatrix}
\]

It is the product of the matrices of the two quadripoles (obtained by the same technique as the product of two determinants). Therefore, cascade connection results in exact matrix multiplication.

Fig. 10-10. Electrical quadripole.

2. Some Definitions Concerning Electrical Quadripoles. We shall make use of the following usual notations and conventions (Fig. 10-10):

\[
\begin{align*}
V_e & = \text{input voltage} & V_r & = \text{output voltage} \\
I_e & = \text{input current} & I_r & = \text{output current}
\end{align*}
\]

We shall assume:

a. That the quadripole \( Q \) involves only passive lumped-constant elements

b. That there is no space variable (Sec. 2.2.2)

c. That nonlinear phenomena can be neglected

The four variables \( V_e, I_e, V_r, \) and \( I_r \) are then related by

\[
\begin{align*}
V_e &= AV_r + BI_r \\
I_e &= CV_r + DI_r
\end{align*}
\]

where \( A, B, C, \) and \( D \) are in general functions of \( s \) and are related by

\[AD - BC = 1\]

under the above assumptions. These relations are written in the matrix form

\[
\begin{vmatrix}
V_e \\
I_e
\end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \times \begin{vmatrix} V_r \\
I_r \end{vmatrix}
\]

The matrix

\[
\begin{vmatrix} A & B \\ C & D \end{vmatrix}
\]

is the inverse transfer matrix or, briefly, the matrix of the quadripole. The output is expressed in terms of the input as

\[
\begin{vmatrix} V_r \\
I_r
\end{vmatrix} = \begin{vmatrix} D & -B \\ -C & A \end{vmatrix} \times \begin{vmatrix} V_e \\
I_e \end{vmatrix}
\]

A quadripole is said to be symmetrical if its properties remain unchanged when the input and the output are interchanged. In this case \( D = A \), provided the sign conventions chosen for the currents are reversed. It
can be shown that the development of the theory of quadripoles in general leads to the theory of symmetrical quadripoles.

**QUADRIPOLE CONNECTIONS.** Cascade, series, parallel, and series-parallel connections of quadripoles are shown in Table 10-1.

**Table 10-1. Quadripole Connections**

**Cascade Connection**

\[ \begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} A_1 & B_1 \\ C_1 & D_1 \end{vmatrix} \times \begin{vmatrix} A_2 & B_2 \\ C_2 & D_2 \end{vmatrix} \]

**Series Connection**

\[ \begin{vmatrix} A & -1 \\ C & C \end{vmatrix} = \begin{vmatrix} A_1 & -1 \\ C_1 & C_1 \end{vmatrix} + \begin{vmatrix} A_2 & -1 \\ C_2 & C_2 \end{vmatrix} \]

**Parallel Connection**

\[ \begin{vmatrix} D & -1 \\ B & B \end{vmatrix} = \begin{vmatrix} D_1 & -1 \\ B_1 & B_1 \end{vmatrix} + \begin{vmatrix} D_2 & -1 \\ B_2 & B_2 \end{vmatrix} \]

**Series-parallel Connection**

\[ \begin{vmatrix} 1 & -B \\ A & A \end{vmatrix} = \begin{vmatrix} 1 & -B_1 \\ A_1 & A_1 \end{vmatrix} + \begin{vmatrix} 1 & -B_2 \\ A_2 & A_2 \end{vmatrix} \]

**QUADRIPOLE IMPEDANCES.** The load impedance of a quadripole characterizes the elements that follow it: \( Z_r = V_r/I_r \). The input impedance \( Z_e = V_e/I_e \) depends both on the quadripole and on its load impedance, since it is possible to obtain \( I_e \) for a given \( V_e \) only if \( Z_r \) is known. More precisely, \( Z_e \) and \( Z_r \) are related by

\[ Z_e = \frac{AZ_r + B}{CZ_r + D} \]

which is a homographic relation. In particular, \( Z_e = Z_r \) when \( Z_r \) is a root of

\[ CZ^2 + (D - A)Z - B = 0 \]

The positive real root of this equation is termed the iterative impedance \( Z_i \) of the quadripole. Thus, when the load impedance is chosen to be \( Z_i \), the input impedance of the quadripole is also equal to \( Z_i \).

When the quadripole is symmetrical \( (D = A) \), the iterative impedance is termed the characteristic impedance \( Z_e \) of the quadripole, given by \( Z_e^2 = B/C \).
3. Mechanical Quadrupoles. By making use of the $V \sim v$ (or $V \sim \omega$) analogy (Sec. 2.4), it is possible to define the mechanical impedance of a mechanical system as the ratio of velocity to force (or angular velocity to torque):

$$Z = \frac{V(s)}{F(s)} \quad \text{or} \quad Z = \frac{\omega(s)}{\tau(s)}$$

This leads to the notion of mechanical quadrupoles. To be specific, let us consider the case of a rotational mechanical system (Fig. 10-11) in which $\tau_e$ is the torque (exerted on the input shaft by the preceding system).

![Diagram of mechanical quadrupole](image)

**Fig. 10-11. Mechanical quadrupole.**

and $\tau_r$ is the torque which the output shaft transmits to the following system. If the system is assumed to be linear, its equations can be written

$$\begin{align*}
\omega_e &= A\omega_r + B\tau_r \\
\tau_e &= C\omega_r + D\tau_r
\end{align*}$$

or, in terms of matrices,

$$\begin{bmatrix} \omega_e \\ \tau_e \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} \omega_r \\ \tau_r \end{bmatrix}$$

The matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is the inverse transfer matrix or, briefly, the matrix of the quadrupole. The condition $AD - BC = 1$ is in general satisfied. (It is not, however, when gyroscopic coupling is present, but we shall exclude this case.) All the notions relative to electric quadrupoles can be extended to mechanical quadrupoles.

**Note.** The introduction of disturbances into a control system is generally represented by a block diagram as in Fig. 10-12. When the disturbing torque depends on the output, such block diagrams are no longer valid; it is then better to represent the whole system as a mechanical quadrupole.

**Examples.** The use of mechanical quadrupoles is helpful in the study of servo-mechanisms in the following cases:

a. When the torsional stiffness of the shafts cannot be assumed to be infinite; also when elastic couplings are present between two or more rigid shafts (see below, Example A)
b. When mechanical clutches, which produce a torque proportional to relative angular velocity, are present (Example B)

c. When gears, together with inertia and viscous friction, are present

This list is, of course, not complete.

**Fig. 10-13. Noninfinite stiffness of a shaft.**

**Example A** (Fig. 10-13)

The laws of dynamics, applied to each of the inertias, gives

\[
\tau_r = J_1 \frac{d\omega_r}{dt} + k \int (\omega_e - \omega_r) \, dt
\]

\[
k \int (\omega_e - \omega_r) \, dt = J_2 \frac{d\omega_r}{dt} + \tau_r
\]

Assuming zero initial conditions, the Laplace transform is

\[
\omega_r = \frac{J_2 s^2}{k} + 1 \omega_r + \frac{s}{k} \tau_r
\]

\[
\tau_r = \left[ \frac{J_1 J_2 s^2}{k} + (J_1 + J_2) s \right] \omega_r + \left( \frac{J_1 s^2}{k} + 1 \right) \tau_r
\]

In other words,

\[
\begin{bmatrix}
\omega_e \\
\tau_e
\end{bmatrix} = |Q| \times \begin{bmatrix}
\omega_r \\
\tau_r
\end{bmatrix}
\]

with

\[
|Q| = \begin{vmatrix}
\frac{J_2 s^2}{k} + 1 & \frac{s}{k} \\
\frac{J_1 J_2 s^2}{k} + (J_1 + J_2) s & \frac{J_1 s^2}{k} + 1
\end{vmatrix}
\]

The condition \( AD - BC = 1 \) is satisfied.

**Example B** (Fig. 10-14): clutch

Similarly, one gets

\[
|Q| = \begin{vmatrix}
\frac{(J_1/f)s + 1}{f} & \frac{1/f}{(J_1 J_2/f)s + (J_1 + J_2)s} \\
\frac{(J_1 J_2) s^2}{f} + (J_1 + J_2)s & \frac{(J_1/f)s + 1}{(J_1 J_2/f)s + (J_1 + J_2)s}
\end{vmatrix}
\]

The condition \( AD - BC = 1 \) is again satisfied.

**Fig. 10-14. Mechanical clutch.**
EXAMPLE C (Fig. 10-15): gear

The system equations are:

\[ J_1 \frac{d\omega_e}{dt} + f_1\omega_e = \tau_e - \tau \]
\[ J_2 \frac{d\omega_r}{dt} + f_2\omega_r = \alpha \tau - \tau_r \]

The quadripole matrix is:

\[
\begin{bmatrix}
\alpha & 0 \\
\alpha(J_1s + f_1) + J_2(s + f_2)/\alpha & 1/\alpha
\end{bmatrix}
\]

The open-circuit input impedance (\(\tau_r = 0\)) is:

\[ Z_e = \frac{\omega_e}{\tau_e} (s) = \frac{1}{(J_1 + J_2/\alpha^2)s + (f_1 + f_2/\alpha^2)} \]

That is, when \(J_1\) and \(f_1\) are neglected:

\[ Z_e = \frac{\alpha^2}{J_2s + f_2} = \alpha^2 Z_r \]

This expression can be compared to the impedance of a transformer: \(Z_e\) is the secondary impedance, \(Z_r\) is the secondary impedance as seen from the primary, and \(\alpha\) is the turns ratio of the transformer.

4. Electromechanical Systems. Up to now we have considered pure electrical or pure mechanical systems. Let us now consider the coupling component, assuming that the coupling is electromechanical.

The physical laws for coupling are the Laplace law giving the force acting on a current-carrying conductor placed in a magnetic field \(F = kI\), and the Maxwell law giving the voltage induced in a conductor moving in a magnetic field \(V = k_1v\). These two laws are simultaneously valid; so, the law of energy yields \(Fv = VI\), whence \(k = k_1\) (which can be made equal to unity by proper choice of the units).
In particular, let $e$ be the back emf of a field-controlled electric motor (Fig. 10-16), $\tau$ its driving torque, $I$ the armature current, and $\Phi$ the magnetic flux:

\[ e = k\Phi \omega \quad \tau = k\Phi I \]

The equations

\[ V_e = R I_e + L \frac{dI_e}{dt} + k\Phi \omega, \]

\[ k\Phi I_e = \tau, \]

can be written in matrix form as

\[
\begin{bmatrix}
V_e \\
I_e
\end{bmatrix} =
\begin{bmatrix}
k\Phi & \frac{R + L\omega}{k\Phi} \\
0 & \frac{1}{k\Phi}
\end{bmatrix}
\begin{bmatrix}
\omega \\
\tau
\end{bmatrix}
\]

The inverse transfer matrix is the product of the two matrices $Q_1$ and $Q_2$; thus,

\[ Q_1 = \begin{bmatrix} 1 & R + L\omega \\ 0 & 1 \end{bmatrix} \quad Q_2 = \begin{bmatrix} k\Phi & 0 \\ 0 & \frac{1}{k\Phi} \end{bmatrix} \]

$Q_1$ is the matrix of a purely electrical quadrupole; $Q_2$ represents the coupling quadrupole (Fig. 10-17).

The homographic relation (Sec. 10.1.2, par. 2) between the input and output impedances is

\[ \frac{Z_e}{(k\Phi)^2} = Z_r + \frac{R + L\omega}{(k\Phi)^2} \]

that is,

\[ \frac{Z_e}{(k\Phi)^2} = Z_{\text{mech}} + \frac{1}{(k\Phi)^2} Z_{\text{elec}} \]

where $1/(k\Phi)^2$ is a pure numerical factor.

In conclusion, then, it may be said that it is no longer necessary to specify whether a part of a control system is electrical or mechanical. The term impedance of a quadrupole may be used without specifying whether the elements are electrical or mechanical.

10.1.3. Conclusion. The study of a quadrupole should take into account the elements following the quadrupole; information concerning them is summarized in the load impedance of the quadrupole. Hence, it is important that quadrupoles and the corresponding impedances should be properly chosen with respect to one another: this is the general problem of impedance matching.

In the course of servo-system synthesis this problem arises in many ways; we shall focus our attention on these: (a) how to synthesize a cascade of elements whose transfer functions are predetermined and (b) how to obtain maximum power transfer. Section 10.2 will be devoted to the first problem, Sec. 10.3 to the second. As will be seen, both lead to the use of the concept of iterative (or characteristic) impedance.

10.2. CASCADE-SYSTEM SYNTHESIS WITH PREDETERMINED TRANSFER FUNCTIONS

10.2.1. General. This problem arises when one wishes to improve the performance of a servo system by designing a compensating network
(Chap. 18). Strictly speaking, the transfer function is dependent on the elements following the network; using the notations of Sec. 10.1.2, par. 2, \( V_s/V_r = A + B/Z_r \). In two cases, however, the equation simplifies.

Case 1. If the output impedance \( Z_r \) is very large, one has only \( V_s/V_r = A \), which is the usual open-circuit transfer function. This can be obtained in two ways:

- By choosing the output impedance \( Z_r \) very large with respect to \( A \) (or, conversely, \( A \) very small with respect to \( Z_r \), if \( Z_r \) is given)
- By inserting an electronic amplifier (buffer amplifier) between the output of the quadripole and the following stage

Fig. 10-18. Cascade-system synthesis.

Case 2. If a cascade of quadripoles satisfies the following assumptions (see Fig. 10-18):

1. All quadripoles have the same iterative impedance \( Z_i \)
2. The impedance across the output is precisely \( Z_i \)

then the output impedance of each quadripole is equal to \( Z_i \).

10.2.2. Application to the Synthesis Problem. The synthesis problem of cascade systems can be stated as follows: It is desired that a cascade of systems \( Q_1, Q_2, Q_3, \ldots \) be represented by a given inverse transfer function \( H(s) = H_1(s)H_2(s)H_3(s) \) for a given load impedance \( Z \) on the last system (\( Z \) may be the input impedance of the following electrical stage, the armature impedance of the motor, the mechanical impedance of the system to be driven, etc.). This problem can be solved by determining each system \( Q_i \), separately so that it satisfies the two following conditions:

1. Its iterative impedance is equal to \( Z \).
2. Its inverse transfer function is \( H_i(s) \) when the load is \( Z \).

Thus, it is certain that, when the systems are placed in cascade and when the load \( Z \) is attached to the output of the latter, the over-all transfer function will be the product of the transfer functions of the two systems (Fig. 10-18).

In the case of a symmetrical quadripole the following two relations hold:

\[
AD - BC = 1 \quad A = D
\]

The quadripole is completely determined by the knowledge of two terms of the matrix, for example, \( A \) and \( B \). The two equations

\[
\frac{B}{(A^2 - 1)^{3/2}} = Z \quad A + \frac{B}{Z} = H(s)
\]

enable one to determine the quadripole (its realizability is to be checked afterwards).

Example. The following example concerns the synthesis of networks having circular transfer loci (for instance, homographic transfer func-
Such networks are often used as compensating networks (see Chap. 18) in servo systems and placed between the sensing device and the preamplifier.

For instance, the matrix of the quadripole representing the double lattice network shown in Fig. 10-19 is

\[
\begin{vmatrix}
Z_B + Z_A & 2Z_AZ_B \\
Z_B - Z_A & Z_B - Z_A \\
2 & Z_B + Z_A \\
Z_B - Z_A & Z_B - Z_A
\end{vmatrix}
\]

It can be easily verified that

\[AD - BC = 1 \quad A = D\]

Let the following conditions be imposed on this quadripole:

1. Its characteristic impedance should be real and equal to \( R \) (the reason for this condition is given below).

2. Its transfer function when the load is \( R \) should be \( k(1 + ars)/(1 + rs) \), where \( a \) and \( r \) are given position quantities (\( k \) is not given).

All calculations made, these conditions yield:

a. \( Z_AZ_B = R^2 \); that is, the impedances \( Z_A/R \) and \( Z_B/R \) are reciprocal.

b. \( k(1 + ars)/(1 + rs) = (R - Z_A)/(R + Z_A) \); that is, \( Z_A \) is a homographic function of \( s \), and the corresponding transfer locus is a semicircle.

Thus, \( Z_A \) may be taken as the impedance shown in Fig. 10-20; its reciprocal \( Z_B \) is shown in Fig. 10-21 (interchanging \( Z_A \) and \( Z_B \) is equivalent to changing \( k \) to \(-k\)). \( r_1 \) represents the resistance of the coil \( L \), and \( r'_1 \) the leakage resistance of the capacitor \( C \).

One must consider numerical values to assure that the impedances can be physically realized. The network shown in Fig. 10-19 gives

\[
k = \frac{(R - r_2)(r_1 + r_s) - r_1r_s}{(R + r_2)(r_1 + r_s) + r_1r_s}
\]

\[a = \frac{1}{k} \frac{R - r_1 - r_2}{R + r_1 + r_s}
\]

\[\tau = \frac{L}{(R + r_2)(r_1 + r_s) + r_1r_s}
\]

where \( \theta = \frac{L}{r_s} \) (equivalent time constant of the inductance)

\[\lambda = \frac{a + 1 - 2ar/\theta}{a - 1}\]
The conditions \( L > 0, r_1 > 0, \theta > 0, x > 0 \) demand, if
\[
L = \frac{\tau}{1 - \tau/\theta} \frac{r_1(R + r_2)}{R + r_1 + r_2}
\]
that
\[
\theta > \tau \quad \lambda < x < 1 \quad \text{or} \quad 1 < x < \lambda
\]
Numerically, for a phase-lead network (see Sec. 18.2.2) with \( a = 10, \tau = 5 \times 10^{-2} \) sec, and \( R = 10,000 \) ohms, the following values are acceptable:
\[
L = 20 \text{ henrys} \quad r_1 = 4,800 \text{ ohms} \\
r_2 = 11,000 \text{ ohms} \quad r_3 = 300 \text{ ohms}
\]

10.3. IMPEDANCE MATCHING FOR MAXIMUM POWER TRANSFER

10.3.1. General. The iterative-impedance approach developed in Sec. 10.2 provides a method for synthesizing a cascade of elements with predetermined transfer function. This method consists in starting from the last element and then proceeding backward to the left. It thus leads to the problem of matching the internal impedance \( Z_i \) of an energy source with a load impedance \( Z \) (Fig. 10-22). In particular, it is often desirable that matching be performed in order to obtain maximum power transfer.

10.3.2. Examples of Impedance Matching Involving a Source of Energy. 1. Temperature Measurement. Consider a temperature resistance pick-off. Variations in temperature above and below 0°C result in variations in current in a circuit made up of the pick-off, a galvanometer, and a battery (Fig. 10-23). Let \( E \) be the emf of the battery, \( x \) the resistance of the galvanometer, and \( R \) the resistance of the pick-off at a temperature of 0°C. Consider finding the value of \( x \) which provides maximum sensitivity \( dB/d\theta \) (\( dB \) is the galvanometer deviation and \( d\theta \) is the temperature change; \( dB/d\theta \) is proportional to \( dB/dR \)). If the galvanometer winding has \( n \) turns,
\[
B = k_1 in = k_2 E \frac{x}{R + x}
\]
because \( x \) is proportional to \( n \). Then
\[
\frac{dB}{dR} = -k_2 E \frac{x}{(R + x)^2}
\]
The condition for maximum sensitivity
\[
\frac{\partial}{\partial x} \left( \frac{dB}{dR} \right) = 0 \quad \text{yields} \quad x = R
\]
If changes in temperature around 0°C are to be measured, the internal resistance of the galvanometer must be equal to $R_0$.

2. **Voltage Source.** Let a generator (emf $E$, internal resistance $R$) be connected to a circuit with resistance $x$ (Fig. 10-24). Consider finding the value of $x$ which allows maximum d-c power transfer $P_u$ into $x$. The following relation holds:

$$P_u = xi^2 = x\left(\frac{E}{R + x}\right)^2$$

$E$ and $R$ being fixed, $P_u$ is a maximum for

$$x = R$$

that is, when the internal resistance of the generator is equal to the resistance of the output circuit. This can be generalized for alternating current. The instantaneous power in the circuit is

$$P(t) = i(t)v(t) = V_mI_m \sin \omega t \sin (\omega t - \phi) = \frac{V_mI_m}{2} [\cos \phi - \cos (2\omega t - \phi)]$$

$P(t)$ is the sum of (a) a constant term: the power $P_a$, which corresponds to the power dissipated in the circuit, and (b) a periodic term: the "reactive" power $P_r$. Using complex representation, if

$$P^* = \frac{1}{2}v \cdot \bar{i}^*$$

where $\bar{i}$ is the conjugate of $i$, one has

$$P_a = \text{real part of } P \quad P_r = \text{imaginary part of } P$$

It is easily shown that maximum active power transfer occurs when the impedance of the circuit is the imaginary conjugate of that of the generator. Since both these impedances are functions of the frequency $\omega$, matching occurs in general for only one frequency. However, if pure resistances are used, the matching will hold, whatever the frequency (since pure resistances are not frequency-dependent); hence it will hold when arbitrary inputs are considered.

3. **Transformer** (Fig. 10-25). It is well known that, when a transformer has a turns ratio $n = N_2/N_1$, an impedance $Z_2$ in the secondary circuit is seen from the primary as is $Z_2/n^2$. Now, it may be shown that the power transmitted is a maximum when $n^4 = |Z_2/Z_1|^2$; that is, the generator and the circuit impedances, as seen from the generator, have equal magnitude.

4. **Impedance Matching in a Transmission Line.** Similar considerations hold for distributed-constant circuits. For example, a transmission line has a characteristic impedance

$$Z_c = \left(\frac{R + jL}{G + jC}\right)^{1/2}$$

\[\text{Fig. 10-24. Example 2 of impedance matching: voltage generator.} \quad \text{Fig. 10-25. Example 3 of impedance matching: transformer.}\]
where \( R, L, C, \) and \( G \) are coefficients of resistance, inductance, capacitance, and conductance per unit length. Maximum power transfer occurs when the load impedance is chosen to be equal to the characteristic impedance of the line (equivalent to the condition for zero reflection).

10.3.3. Conclusions from These Examples. The above four examples have shown that maximum active power transfer occurs when the impedances of the source and its output circuit are equal. Furthermore, if these impedances are purely resistive, the matching holds regardless of how the source may vary in time (all-frequency matching). These results are general where a source and its output circuit are concerned; that is, they apply at the source end of a cascade of elements.

10.3.4. Matching of Quadripoles for Maximum Power Transfer. The question now may be asked, what is the condition for maximum power transfer through a quadripole, i.e., through any element of the cascade of elements? There is no general answer to this question; however, for a-c conditions, the power dissipated in the quadripole is a minimum when the load impedance is equal to the quadripole iterative impedance (the power factor remaining unchanged). This statement results from the definition of the iterative impedance

\[
Z_i = \frac{V_e}{I_e} = \frac{V_r}{I_r}
\]

and is of great practical importance. However, it should not be overestimated: choosing the load impedance of a given quadripole equal to \( Z_i \) does not change \( \cos \varphi \). Some other load impedance, although not matched, might have led to a more favorable value of \( \cos \varphi \). In most cases the iterative impedance is close to the optimum, hence considering it can be useful at least as a first approximation.

The general problem of impedance matching for maximum power transfer is difficult because many other factors must be taken into account at the same time. However, matching by means of iterative (or characteristic) impedances generally constitutes an acceptable compromise. Furthermore, the advantage of using, whenever possible, impedances independent of frequency is to be kept in mind.
CHAPTER 11
LINEARITY DOMAIN

Summary
1. Magnitude domain.
2. Frequency domain.

It has been shown in previous chapters how a linear system can be completely characterized by its transfer function or, more generally, by its transfer matrix. Such a representation is extremely convenient, and one is easily tempted to identify the system completely with its transfer function, that is, to forget that linear operation can be assumed only on the basis of certain simplifying assumptions—in other words, that a transfer function is valid only in a certain linearity domain.

This fact and its consequences, however obvious they may seem, are so easily and so frequently overlooked, and the results of such negligence can be so important, that this entire chapter is devoted to simply emphasizing the fact that transfer functions are just an alternative manner of writing linear differential equations and, as a consequence, are valid only in so far as these equations are valid. These limits of validity will be considered in the following sections.

11.1. AMPLITUDE LINEARITY DOMAIN

11.1.1. Amplitude Characteristics of a Linear System. As previously pointed out, linearity essentially implies proportionality of output to input. Therefore, the static characteristic of a linear system, with the input as abscissa and the output as ordinate, is a straight line that extends to infinity and whose slope represents the static gain of the system (Fig. 11-1).

11.1.2. Dead Zone. Actually, no physical system is perfectly linear. First, in the neighborhood of zero input condition there always exists a threshold of accuracy termed dead zone or inactive zone, that is, a small range of variation for the input to which the system is insensitive. In other words, the response \( r \) of the system will be zero not only when the input \( e \) is zero, but whenever \( e \) satisfies

\[-e_m < e < e_m\]

The existence of a dead zone appears in the characteristic curve as shown in Fig. 11-2.
The dead zone may be due to different physical phenomena. In many mechanical systems, such as the accelerometer of Sec. 6.1.1, it is caused by *coulomb friction*, or *static friction*, that is, friction developing a force constant in magnitude and opposite to the direction of the motion. A nonzero value of the input is necessary to overcome this type of friction force. Frequently, especially in electrical systems, the threshold is a consequence of phenomena that exist on the microscopic scale and can be characterized only in statistical terms. Such phenomena are referred to as *noise* (Secs. 12.1.2 and 29.1.3 to 29.1.8) and can result in the equivalent of a dead zone (for more detail, see Sec. 22.3.5). They lead to a fundamental limitation in the accuracy of systems.

### 11.1.3. Saturation

Similarly, a proportional output-vs.-input relationship never holds, in practice, for "infinite" values of the input. In most cases, for high values of the input, the output increases less than proportionally, or even remains constant. This phenomenon is called *saturation*; it appears in the characteristic curves as shown in Fig. 11-3. The case of pure saturation (Fig. 11-3b) is expressed by

\[
\begin{align*}
    r &= Ke & \text{for } e < e_M \\
    r &= r_M & \text{for } e > e_M
\end{align*}
\]

In mechanical systems the presence of limit stops for an accelerometer, a rate gyro, etc., results in saturation. For the longitudinal motion of an airplane (Sec. 7.5.2), saturation results in the stall phenomenon; that is, the lift force due to a positive control-surface deflection ceases to increase and may even diminish when the deflection exceeds a certain value. In electronic systems the saturation effect of amplifiers is well known.

### 11.1.4. Generalization

More generally, the output-vs.-input amplitude relation \( r = f(e) \) can be any function. However, the function \( f(e) \) can in many cases be approximated by considering a threshold, a saturation, and a static gain in the region between threshold and saturation (Fig. 11-4).

### 11.1.5. Magnitude Linearity Domain

Summarizing these discussions, it is seen that, so far as static behavior is concerned, a so-called linear
system is actually linear only when the input magnitude is greater than the threshold and does not exceed the saturation, that is, when

\[ e_m < |e| < e_M \]

The interval \((e_m, e_M)\) is called the magnitude linearity domain for the system.

The ratio \(e_M/e_m\) is a figure of merit for the linear operation of the system. It is generally evaluated in decibels as \(20 \log_{10} (e_M/e_m)\). For example, a measuring instrument may have a magnitude linearity domain of \(e_M/e_m = 100\), that is, 40 db. Linear domains of 1,000, that is, 60 db, are, in general, to be found only in precision instruments. Very frequently the threshold and saturation of a system are adjustable, but their ratio remains constant and is a characteristic of the instrument. For example, the threshold and maximum deflection of a rate gyro can be adjusted by changing a resistance placed in series with the restoring circuit, which results in shifting the magnitude linearity domain by a certain number of decibels, but does not change its range.

Fig. 11-4. Approximation of curved characteristic by dead zone and saturation.

It is very important to understand clearly that the amplitude linearity domain of a system is a characteristic of primary importance, since it expresses the range of linear operation for the system. When characterizing a physical system, its magnitude linearity domain should always be indicated at the same time as its transfer function. When prescribing given performance for a system or for a component, the magnitude linearity domain should be considered, first, even prior to the transfer function.

11.2. FREQUENCY LINEARITY DOMAIN

11.2.1. Introductory Examples. 1. Electric Lumped-parameter Circuit. It has been pointed out in Secs. 2.2.2 and 2.3.1 that the differential equations of electric lumped-parameter circuits are no longer valid when the phenomena involved vary very rapidly, i.e., when the frequencies involved are too high. For example, potentiometers are not pure resistances, they have a self-inductance. The effect of this self-inductance is negligible only at low frequencies, i.e., when the currents that flow through them do not vary too rapidly. (The terms of the form \(L \, di/dt\) can then be neglected when writing the equations.) But as the frequency increases, potentiometers will behave more and more like resistances plus self-inductances. Similarly, a vacuum tube is known to have a capacitance that is no longer negligible at high frequencies, etc. Thus, if the behavior of the system at high frequencies is to be
studied, the equations written for low-frequency phenomena must be rewritten in more complicated form including unwanted capacitances, etc. At very high frequencies, the system is no longer governed by the linear Kirchhoff laws. Maxwell’s equations must then be applied; this is the technique of microwave circuits. In conclusion, the equations obtained by writing Kirchhoff’s laws are valid only at low frequencies, i.e., in a frequency band \((0, F)\), where \(F\) may be of the order of a few megacycles.

2. Mechanical Lumped-constant Systems. Mechanical systems made up of inertias, springs, and dashpots have been briefly described in Sec. 2.4.4. The equations written on the lumped-parameter basis are valid only at low frequencies; for when such a mechanical system is subjected to rapidly varying inputs, the inertias of the springs and dashpots are no longer negligible; furthermore, owing to the important stresses developed, the stiffness of the shafts can no longer be considered as infinite. In a manner analogous to the electrical case, unwanted inertias and stiffness coefficients must be introduced. But for mechanical systems the upper limit of the frequency range \((0, F)\) in which the usual linear equations (analogous to Kirchhoff’s laws) are valid is considerably lower than in the electrical case. It may be of the order of 10 or 15 cps in ordinary mechanical systems, or of 100 or 150 cps in mechanical systems that are especially designed for linear operation at high frequencies, e.g., high-speed recorders. Analogous considerations apply to hydraulic, acoustic, and other types of systems.

3. Airplane in Longitudinal Motion. Similarly, it can be seen that the transfer functions derived in Sec. 7.5.2 for the longitudinal motion of an airplane are not valid at high frequencies. The primary reason for this is that, in writing these equations, the airplane was considered as a rigid body, which is a legitimate assumption so long as the effect of noninfinite stiffness remains below the over-all threshold of accuracy implied. But if the control surface is deflected with a usual amplitude at a frequency of, say, 5 cps, considerable stresses are developed. Deformations of the airplane wings, tail, and body result and can no longer be neglected. Taking them into account leads to writing more complicated equations.

4. Lags and Time Constants. The exact transfer function of a pure time lag (delay) \(T\) is known to be \(e^{-tT}\). However, the transfer function \(1/(1 + Ts)\) constitutes a satisfactory approximation when the frequencies under consideration are sufficiently low (Sec. 5.5). Therefore, when a small time lag \(T\) is present in a servomechanism, it is usually permissible to approximate it by a time constant \(T\) with a transfer function \(1/(1 + Ts)\) when analyzing the effect of it in the resonance region of the servomechanism (Sec. 13.4.2); but this approximation may lead to incorrect conclusions for the analysis of phenomena that involve higher frequencies.\(^1\)

11.2.2. Upper Limit \(F\) of the Linearity Domain. Thus, transfer functions for linear systems are no longer valid when the frequencies involved become high. The transfer functions for the above systems are valid only in a frequency band \((0, F)\), where \(F\) can be of the order of a few dozens of cps, or of a few cps in the case of mechanical systems. The word valid means that the linear differential equations or the corresponding transfer functions will properly represent the behavior of the system, e.g., will yield for the frequency response results that agree with experience within, say, 3 to 5 per cent for the amplitude and, say, 5° for the phase response.

\(^1\) This is often the case when one wishes to investigate the possibility of the occurrence of limit cycles when the gain is increased or when nonlinearities of the on-off type are introduced (Sec. 24.3.5).
11.2.3. Lower Limit $F_1$ of the Linearity Domain. It happens sometimes (but this is not a general case) that the transfer function used for describing a linear system is not valid for very slowly varying phenomena, that is, when the frequencies involved are lower than a fixed frequency $F_1$.

For example, the transfer functions derived in Sec. 7.5.2 for the longitudinal motion of an airplane were based on the assumption of a constant forward velocity for the airplane. This assumption cannot be considered as true at low frequencies—in particular, for the static behavior—since different steady states of flight involving different angles of attack necessarily imply different values of the forward velocity (see Sec. 7.5.2, par. 3). Thus, the low-frequency part of the transfer function, i.e., essentially the $1/s$ integrating factor, does not describe the behavior of the airplane at low frequencies. In fact, a more detailed analysis shows that the integrating factor $1/s$ should be replaced by a quadratic factor with a low natural frequency (of the order of 0.5 cpm for an airplane flying at 500 mph, and inversely proportional to the speed of the airplane\(^1\)) and a low damping ratio (of the order of 0.1). This quadratic factor accounts for the so-called phugoid oscillation of the airplane center of gravity in a vertical plane.

In conclusion, the transfer function derived in Sec. 7.5.2 is valid only in a frequency band $(F_1,F)$. The lower limit (say a fraction of 0.1 cpe) results from neglecting the low-frequency mode of the airplane. The upper limit $F$ (say a few cpe) corresponds to the general limitation of transfer functions at high frequencies. Within these two limits, the transfer functions and loci obtained provide an excellent approximation for studying the behavior of the airplane and can be utilized if it is desired, e.g., to stabilize the high-frequency mode of the airplane, whose resonant frequency usually lies between 0.2 and 0.8 cpe.

11.2.4. Practical Limitations Due to the Upper Limit $F$. The existence of an upper limit $F$ for the frequency range in which the transfer function of a system is valid results in certain limitations concerning the manipulation of this transfer function. Such limitations are important to know and to be aware of, but frequently, especially in servo problems, they appear in a subtle manner that requires careful thought on the part of the engineer. The following discussion attempts to explain why, when representing a system by its transfer function, it is, in general, necessary to consider higher frequencies than would be expected a priori.

Suppose a system to be controlled has a resonant frequency $\omega_n$. The quantity $\omega_n$ provides an order of magnitude for the time scale of the system. For example, if the system is an airplane that is to be autopiloted, $\omega_n$ may range from 1 or 2 rad/sec for a cargo freighter to 10 or 15 rad/sec for a fast missile. Assuming an $\omega_n$ of 3 rad/sec $= 0.5$ cpe, it might seem sufficient to make sure that the transfer function used when describing it is valid in the frequency band from 0 to $2\omega_n = 1$ cps. However, once the aircraft is autopiloted, the resonant frequency is no longer 0.5 cpe; it becomes higher.\(^2\) for example 0.7 cpe, or even 0.9 cpe if the piloting is "stiff." Therefore, it is necessary that the aircraft transfer function be valid up to $2 \times 0.9 = 1.8$ cpe, that is, in a frequency band

---


\(^2\) This fact will be explained in Sec. 13.2.3, Note.
twice as large as would have been expected. As an application of this, if the transfer function of the aircraft is obtained by performing frequency-response tests, these tests should be carried out at much higher frequencies than would be necessary if the aircraft were considered as an open-loop system. When the autopilot is in operation, the important parameter for defining the time scale of the system is the over-all resonant frequency of the closed-loop system.

This manner of reasoning applies to most servomechanisms. It always leads to considering much higher frequencies (often 10 or 20 times) than would seem reasonable, especially in these two cases: (a) when the servo system is stiff, that is, when its open-loop gain is high;\(^1\) and (b) when compensating networks that result in an increase in the system natural frequency are used.

Furthermore, once a servo system has been adjusted, unexpected changes in the values of certain parameters may cause the system behavior at high frequencies to play an important part in its performance.

![Magnitude vs Frequency](image)

**Fig. 11-5.** Schematic linearity domain.

An example will be given in Sec. 34.2.2 (last paragraph) in which a satisfactorily stable servo system with a resonant frequency of 1 rad/sec runs into instability as a consequence of its behavior in the 10- to 20-rad/sec frequency band.

**11.3. SYNTHESIS: LINEARITY DOMAIN**

In conclusion, when a system is described by linear equations or, equivalently, by a transfer function, it should be borne in mind that the transfer function is valid only in a given magnitude and frequency domain. Considering input angular frequency \(\omega\) as the abscissa and input magnitude \(e\) (logarithmic) as the ordinate, the linearity domain is that area of the \(\omega, e\) plane which is contained between the two horizontal lines \(e_m\) and \(e_M\) corresponding to threshold and saturation, respectively, and the two vertical lines corresponding to the lower boundary \(F_1\) (often zero) and the upper boundary \(F\) of the frequency domain (Fig. 11-5). To characterize a linear system, it is necessary to specify its linearity domain and its transfer function.

Actually, the boundaries of the linearity region are not straight lines.

\(^1\) See Sec. 13.2.3. Note that the open-loop gain of most servomechanisms is much higher than the open-loop gain of autopiloted aircraft (the latter is usually between 0.5 and 2).
The lower limit has the shape of a curved line, since the threshold is lower at medium frequencies. In fact, low-frequency inputs can break the coulomb friction that occurs under static condition; this is known as the dither effect.¹ The upper frequency boundary is generally imprecise, and in the corresponding zone the amplitude linearity domain becomes small. As a result, the linearity domain looks somewhat like a lemon (Fig. 11-6) tapered at its left and right ends. Therefore, it is sometimes called the linearity lemon.

![Figure 11-6. The linearity lemon.](image)

![Figure 11-7. Schematic linearity domain. The cross hatched area represents the information capacity of the system.](image)

In practice, it is sufficient to consider approximate average values for the threshold and the saturation, that is, to consider a linearity rectangle. For systems that can be roughly described as low-pass filters with a cutoff frequency $F_c$, the linearity rectangle is divided (Fig. 11-7) into (a) a utilization rectangle below $F_c$ and (b) a zone in which the system can still be considered to operate linearly but in which it essentially filters out the inputs. It will be seen at the end of the next chapter (Sec. 12.4.5) that the area of the utilization rectangle is a measure of the information capacity of the system.

¹ A possible application of the dither effect is the following: platforms supporting certain mechanical equipment, e.g., airborne gun turrets, are deliberately made to vibrate in order to eliminate the effect of static friction.
CHAPTER 12

STATISTICAL CONSIDERATIONS

Summary
1. Introduction.
2. Notions on frequency spectra.
3. Fundamental relation and applications.
4. Introduction to information theory.

12.1. INTRODUCTION

12.1.1. Comments. Up to now we have studied the response of linear systems to particular types of input, essentially harmonic inputs (frequency response) and step or impulse inputs ( transient response). We have shown how, by means of the concept of transfer functions, the response to an arbitrary input \( e(t) \) can be deduced from one of these results.

Actually, numerous systems, including all feedback control systems, receive what are essentially unexpected inputs. These inputs, which can at most be characterized by a probability function, are called random, or stochastic.

12.1.2. Noise as an Input to Feedback Control Systems. Servo-mechanisms are subjected to random inputs. If this were not the case, complex feedback control systems would not be required (Sec. 1.3.8). System disturbances are always random; the control input, or command, is usually random. Among the disturbances special mention should be made of the internal noises in the system. In any electrical system, the measured noise is always greater than the background noise, owing to motion of the electrons.

In Chap. 29, the background noise will be spoken of as constituting a microscopic noise, with respect to other types of noise, generally depending upon the degree of care inherent in the design of the system. The other types of noise will be called macroscopic noise.

In mechanical or hydraulic systems, the noise component due to the motion of molecular particles (Brownian motion) is not measurable. However, backlashs, pressure fluctuations due to the pumps, bursts of air bubbles in the oil, etc., constitute disturbances of random nature which may also be called macroscopic noise.
12.1.3. Conclusion, Summary of the Chapter. But if inputs to servo systems are unexpected, what can be said about them? As unexpected as they might be, they belong to given categories, and it is possible to characterize them, in probability, by certain quantities. In the case of linear systems, the most important quantity characterizing a random input is its frequency spectrum, which will be defined and studied in the first half of this chapter.

The problem which must be faced is to study the response of a linear system with a random input characterized as stated above, that is, to determine the frequency spectrum of the output. This problem, which will be treated in the second part of the chapter, will lead to important theoretical and practical consequences. Finally, some aspects of information theory that will be described will enable expressing statistical notions in an especially interesting form.

In what follows, importance has been attached to physical reasoning and to the significance of the relations presented rather than to mathematical rigor. For further details and proofs, the reader is referred to works listed in Bibliography.

12.2. CONCEPT OF FREQUENCY SPECTRUM

12.2.1. Intuitive Notions. Let us consider a person speaking. One cannot talk of the frequency of his voice, since his intonation in speech varies the pitch of sounds and each sound is in itself a complex mixture of harmonics (due to the structure of vocal cavities) which characterizes the tone of his voice. However, some persons have more or less high voices, others have more or less deep voices, while still others have what is called a bitonal voice. It is conceivable that a voice can be characterized by a curve (Fig. 12-2) which expresses the energy contained by the voice. (a) The average value, or area of the curve, indicates the average intensity of the voice and (b) the maxima indicate the existence of predominant frequencies. This curve represents the frequency spectrum of the corresponding voice; its area corresponds to the total emitted energy, and its shape to the distribution of this energy in the spectrum of frequencies (0 to \( \infty \)).

It is evident that we do not have the same spectrum for a particular person when he sings or when he speaks; it is clear that it would be meaningless to define the frequency spectrum of his voice from a 10-min recording of both conversation and singing. It can be said that a vocal
broadcast by a person including only conversation is a *stationary* stochastic phenomenon, that is, coherent enough for the frequency spectrum to be the same whether it is, for instance, determined during the first or last 3 min of the broadcast. Frequency spectra are defined only for such phenomena.

In the limit, the utterance of a pure sound of angular frequency \( \omega_0 \), which is not a random function any more, will be represented by a curve formed by a very narrow and high-amplitude pulse with area larger for larger intensity (Fig. 12-3). The utterance of a sound together with harmonics would give a curve of the shape indicated in Fig. 12-4 (line spectrum). The total intensity of the sound is represented by the sum of the shaded areas.

At the other extreme, Fig. 12-5 represents the spectrum of a noise stretching over the whole spectrum of frequencies without any particular frequency predominating. This is called *white noise*.

12.2.2. Mathematical Definitions. 1. Notion of a Stationary Stochastic Process. A random function of time is one whose value at any instant is governed by chance. A *sample* of a random function is the value of the function during a certain interval of time measured, for example, in the form of a recording. As a rule, the information contained in the sample defined the stochastic process as long as the two following conditions are met:

1. The stochastic process is stationary.
2. The sample is taken over a large enough interval of time.

The hypothesis of stationary status assumes that the stochastic process will not change. In this case, the knowledge of a long enough sample is equivalent to the knowledge of the stochastic function itself. From the *ergodic hypothesis* the ensemble averages are the same as the time aver-
ages. (An ensemble is ergodic if all its members have the same statistical properties.)

This property is the basis for most of the calculations done on stochastic functions, which are determined in practice by means of samples. The stationary-status condition assumes that the underlying physical causes of the random function remain permanent in all the experiments of interest.

Since a stationary random function is defined in practice by a long enough sample, the current language uses, somewhat abusively, the term random function to describe a sample.

Example. Figure 12-6 represents a portion of the recording of the output $f(t)$ of a rate gyro located in a guided missile which is flying in a turbulent atmosphere.

![Figure 12-6](image)

Fig. 12-6. Sample of random function (rate gyro mounted on guided missile in turbulent atmosphere).

2. Frequency Spectrum. Let $y(t)$ be a stationary stochastic function of time, or rather a long enough sample of this stochastic function. Let us choose an arbitrary origin of time on this sample. Let $y_T(t)$ be the function defined by:

$$
\begin{align*}
y_T(t) &= y(t) & -T < t < T \\
y_T(t) &= 0 & t < -T, \ t > T
\end{align*}
$$

This function $y_T(t)$ has a Fourier transform

$$
A_T(\omega) = \int_{-T}^{T} y(t) e^{-j\omega t} \, dt
$$

The frequency spectrum $\Phi(\omega)$ of $y(t)$ is defined as the function

$$
\Phi(\omega) = \lim_{T \to \infty} \frac{1}{2\pi} \frac{1}{T} < |A_T(\omega)|^2 >
$$

The sign $<>$ indicates ensemble average.

This definition has practical meaning only if it converges for a value of $2T$ smaller than the length of the sample, and if, in addition, the value of the limit is independent of the chosen origin of time. This is then the stationary-status condition of the random function $y(t)$.

Note. According to the authors the term frequency spectrum is applied to the power density (Truxal, Laning and Battin) or to the power itself (Davenport and Root).

12.2.3. The Mean-square Theorem. The mean-square value of $y(t)$ is, when the mean value of $y(t)$ is zero (an assumption made in all that follows),
\[ \overline{y^2} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} y^2(t) \, dt \]

it can be shown that
\[ \overline{y^2} = \int_{-\infty}^{+\infty} \Phi(\omega) \, d\omega \]

12.2.4. Examples. 1. It can be verified that the frequency spectrum of a sinusoidal function of frequency \( \omega_0/2\pi \) is an impulse function

\[ \Phi(\omega) = A \delta(\omega - \omega_0) \]

defined by
\[ \begin{align*}
\Phi(\omega) &= 0 \quad \text{for } \omega \neq \omega_0 \\
\Phi(\omega) &\text{is infinite} \quad \text{for } \omega = \omega_0 \\
\int_{-\infty}^{\infty} \Phi(\omega) \, d\omega &\text{is finite}
\end{align*} \]

This is the limiting case of a random function, the energy being concentrated at the frequency \( \omega_0 \) (single-line spectrum).

2. A periodic function \( f(t) = f(t + T) = f(t + 2T) = \ldots \) has a line spectrum of the form

\[
\Phi(\omega) = \Phi_0(\omega) + A_1 \delta \left( \omega - \frac{2\pi}{T} \right) + A_2 \delta \left( \omega - \frac{2\pi}{2T} \right) + \ldots
\]

3. A true random function has no periodicity, but exhibits prevailing frequencies determined by the maxima of its frequency spectrum. For instance, the random function, a sample of which was given above (Fig. 12-6), has the spectrum represented in Fig. 12-7. The latter shows a prevailing frequency of 12 rad/sec and secondary frequencies of 7 and 24 rad/sec; half of the energy lies between 11 and 16 rad/sec and only less than 5 per cent lies outside the interval 4 to 20 rad/sec.

4. As opposed to periodic functions, white noise, for which

\[ \Phi(\omega) = \text{const} \]

is a random function whose energy is uniformly distributed in the frequency spectrum. This is the case for several microscopic stochastic phenomena.

12.2.5. Practical Determination of Frequency Spectra, Autocorrelation Function. 1. General. The frequency spectrum of a random function \( y(t) \) is obtained from a sample \( y_T(t) \) by using the autocorrelation function of \( y(t) \), namely the function whose Fourier transform is the frequency spectrum \( \Phi(\omega) \) of \( y(t) \). If \( y(t) \) is a stationary random function, its autocorrelation function \( \varphi(\tau) \) is defined as the mean value of the product

\[ y(t)y(t + \tau) \]

This definition leads to the equation

\[
\varphi(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} y(t)y(t + \tau) \, dt
\]
which is independent of the sample used \((-T,+T)\) (condition for stationary status). In practice, it is customary to "normalize" \(\varphi(\tau)\) by referring it to \(\varphi(0)\), which is the mean square value of \(y(t)\).

For a stationary stochastic process, \(\varphi(\tau)\) is independent of the sample used \((-T,+T)\); in this case, the \(\varphi(\tau)\) measured and calculated for a long enough sample really represents the autocorrelation function of the stochastic process to which the sample belongs.

2. **Some Mathematical Properties of the Autocorrelation Function of a Stationary Stochastic Process**

   a. It is an even function of \(\tau\).

   b. Its value for \(\tau = 0\) is the mean square of \(y\): \(\varphi(0) = \bar{y}^2\).

   c. If \(y(t)\) has a period \(T\), then \(\varphi(\tau) < \varphi(0)\) for \(t \neq kT\) \((k\) integer).

   d. Finally, the frequency spectrum of \(y(t)\) is the Fourier transform of its autocorrelation function:

   \[
   \Phi(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \varphi(\tau) e^{-i\omega \tau} d\tau
   \]

   and

   \[
   \varphi(\tau) = \int_{-\infty}^{+\infty} \Phi(\omega) e^{i\omega \tau} d\omega
   \]

   In particular, we find again

   \[
   \varphi(0) = \int_{-\infty}^{+\infty} \Phi(\omega) d\omega = \bar{y}^2
   \]

3. **Its Significance.** The notion of autocorrelation function can be considered from several points of view. It is a particular case of the **correlation function**. The correlation function between two random functions \(f(t)\) and \(g(t)\) is defined as:

   \[
   \varphi_{fg}(\tau) = \overline{f(t)g(t + \tau)}
   \]

   which measures the correlation between the value of \(f(t)\) at the instant \(t\) and the value of \(g(t)\) \(\tau\) sec later. If there is no correlation, \(g(t + \tau)\) varies in an incoherent manner when \(f(t)\) is given, therefore \(\varphi_{fg}(\tau)\) is very small; if on the contrary there is a correlation [for instance, a relation of cause to effect, \(f(t)\) determining \(g(t)\) with a lag of \(\tau_0\)], then \(\varphi_{fg}(\tau)\) is maximum for \(\tau_0\), which generalizes the notion of time constant for the case of random functions. The autocorrelation function of a function \(y(t)\) is its correlation with itself. One can say briefly that \(\varphi(\tau)\) is related to the coherence of the stochastic process as time progresses and, if its past is known, to its degree of predictability. In this manner, a periodic function is perfectly predictable (it is not a random function). As expected, white noise, with \(\varphi(\tau) = 0\)—whatever the value of \(\tau\) except for \(\tau = 0\)—is the least predictable of all random functions (see Sec. 19.2.6).

4. **Its Practical Determination.** The autocorrelation function of a random function \(y(t)\) can be determined from the recording of a sample \(y_T(t)\) by means of an apparatus called an autocorrelator. One type of autocorrelator is based on the following principle: The area under the recorded sample is shaded as shown in Fig. 12-8. Two carts, displaced
with respect to each other by a constant distance, are driven with equal speed in the direction of increasing \( t \). From each cart a light beam is slaved, by means of a photoelectric cell, to follow the curve \( y(t) \). The beam-deflecting voltages are respectively proportional to \( y(t) \) and \( y(t + \tau) \); multiplying and integrating them gives \( \varphi(\tau) \). Then the operation is repeated for different displacements \( \tau_0 \), thus giving the autocorrelation function \( \varphi(\tau) \).

5. Practical Remarks. Once the autocorrelation function has been determined, it is necessary to be able to verify that the sample presents satisfactory characteristics of stationary status. This necessitates that

![Fig. 12-8. Principle of an autocorrelator.](image)

![Fig. 12-9.](image)

the sample be of sufficient length. For example, in the case of Fig. 12-6, it is necessary that the sample contain at least 20 pseudo periods for obtaining convergence of \( \varphi(\tau) \) within 5 per cent. There follows an important practical consequence. When the autocorrelation function of a supposedly stationary stochastic process is to be determined, one should neglect those pseudo-periodic components whose periods are not sufficiently small with respect to the length of the sample. If these components have significant amplitude, considerable errors would result unless an “average function” of \( y(t) \) is evaluated empirically (see Fig. 12-9) and deviations of \( y \) about this average function are considered.

12.3. FUNDAMENTAL RELATION, APPLICATIONS

12.3.1. Fundamental Relation. If a linear system with a transfer function \( H(s) \) is subjected to a stationary stochastic input \( e(t) \) with a
frequency spectrum $\Phi_e(\omega)$ (Fig. 12-10), it can very easily be shown that the frequency spectrum of the output is

$$\Phi_r(\omega) = |H(j\omega)|^2\Phi_e(\omega)$$

![Fig. 12-10.](image)

Note that the phase of $H(j\omega)$ does not appear in this equation.

The above equation can be interpreted physically as follows: The prevailing frequencies at the output are the frequencies prevailing at the input modified by the weighting coefficient $|H(j\omega)|^2$ which amplifies those frequencies about the resonant frequencies of the system.

For example, if the input is a white noise, $\Phi_e(\omega) = K$, the frequency spectrum of the output will be

$$\Phi_r(\omega) = K|H(j\omega)|^2$$

This enables one to obtain the transfer function of a system from its output when it is subjected to a white noise. As an application of this, a noise of arbitrary frequency spectrum $\Phi(\omega)$ can be obtained from a white-noise source by introducing the latter to a linear system with a transfer function $F(s)$ so that $\Phi(\omega) = |F(j\omega)|^2$

**12.3.2. Application to Disturbances.** Let a linear system be subjected to two noncorrelated inputs:

a. A command, or control input, with a frequency spectrum $\Phi_E(\omega)$

b. An unwanted input, for instance a noise $n(t)$ with a spectrum $\Phi_N(\omega)$

Let

$$H_1(s) = \frac{R}{E}(s) \quad H_2(s) = \frac{R}{N}(s)$$

be the two transfer functions of the system. The output will be the sum of a term due to the command, with a spectrum

$$\Phi_1(\omega) = |H_1(j\omega)|^2\Phi_E(\omega)$$

and of an unwanted term due to the disturbance

$$\Phi_2(\omega) = |H_2(j\omega)|^2\Phi_N(\omega)$$

The mean-square value of the latter term is given by

$$\overline{\varepsilon_N^2} = \int_{-\infty}^{+\infty} \Phi_2(\omega) \, d\omega = \int_{-\infty}^{+\infty} |H_2(j\omega)|^2\Phi_N(\omega) \, d\omega$$

**12.3.3. Output Noise.** Let us consider the very important particular case characterized by the following assumptions:

1. The noise is introduced at the same point in the system as the command $e(t)$.
2. The frequency spectrum $\Phi_N(\omega)$ of the unwanted input can be approximated by a spectrum of white noise.
3. The system can be considered as a low-pass filter.

The first of these three assumptions is merely intended to simplify what follows. Hence $H_2(s) = H_1(s) = R/E(s)$.

The second assumption expresses, in an idealized way, the fact that in
the great majority of cases the noise spectrum contains frequencies much higher than the command spectrum and the natural frequency of the system. The disturbance spectrum given by this second assumption is $\Phi_N(\omega) = K$.

The third assumption states that, as far as the noise is concerned, the system is largely comparable to a low-pass filter which allows low-frequency transmission but attenuates frequencies higher than its cutoff frequency $\omega_c$. This is frequently the case for systems such as servomechanisms in which the last mechanical stage attenuates frequencies greater than a few cps. This assumption leads to writing (see Fig. 12-11):

$$H(j\omega) = H_0 \quad \omega < \omega_c$$
$$H(j\omega) = 0 \quad \omega > \omega_c$$

With these assumptions, the above equation gives for the mean-square value of the output noise

$$\overline{\varepsilon_N^2} = \int_{-\infty}^{\omega_c} H_0^2 K \, d\omega = A \omega_c$$

In other words, the mean-square value of the output noise is proportional to the bandwidth.

12.3.4. Consequence: Choice of Bandwidth. The result just obtained enables one to suggest a choice for the bandwidth of a system as a function of its purpose.

1. First of all, it is necessary that the system have a sufficiently large bandwidth—that is, speed of response—in order that it will not filter out small and rapid input variations. If such input variations have frequency spectra, the greatest part of which lie in a frequency band $(0,F)$ or $(F_1,F)$, it is desirable that the distortion at the frequency $F$ should not exceed, for example, 10 per cent in amplitude and $10^\circ$ in phase. In numerous cases this leads to stipulating that the natural frequency of the system be at least $2F$ or $3F$. For instance, if a guided missile has a natural frequency of 1.5 cps, its natural frequency when stabilized will be of the order of 2 cps; so it is necessary that the sensing device of the autopilot have at least a natural frequency of 8 cps. (A frequently adopted value is 12 cps.)

2. But once a sufficient speed of response is assured, there is no need to further increase the bandwidth; for the resulting increase in output noise would more than outweigh the improvement obtained. In the preceding example, there is no point in exceeding a natural frequency of 16 to 20 cps; the result of boosting the natural frequency up to 40 cps would be the detection of structure vibrations and thrust variations.\(^1\)

\(^1\) An objection to this is the fact that the detector is followed by a mechanical stage and it is the latter that fixes the bandwidth independently of the natural frequency of the detector. This reasoning, which could apply to perfectly linear systems, is faulty because practical systems always involve saturations. The effect of adding noise to a signal is to produce a mixing of the two, in which frequencies lower than the over-all system bandwidth are obtained.
3. Thus, any excess of bandwidth should be avoided, since this involves
needless high-performance devices and increase in output noise, resulting
in a deterioration of accuracy. In the majority of servomechanism
applications it is not recommended to exceed twice the bandwidth con-
sidered as adequate when neglecting noise.

In order to illustrate this important point, two actual examples, which seem to the
authors to be particularly outstanding, will now be given. The first one concerns
flight tests of airplanes in which the outputs of the longitudinal and transverse acceler-
ometers were to be recorded. Since the frequencies of the phenomena to be mea-
sured did not exceed a few cps, accelerometers of natural frequencies 20 to 25 cps
would have been quite adequate. Actually, use was made of accelerometers of very
high natural frequency (180 to 200 cps), originally intended for vibration measure-
ments, but it was thought that an excess of quality could not hurt. The result was
that several hundred meters of recording paper were swamped by the hash from the
vibration of the airplane, with frequencies of about 100 cps.

The second example is the misadventure of a group of experimenters in the early
days of radio remote control. Before starting the experiment, it was decided to test
the radio link by voice transmission. But reception was obliterated by static, and
attempts to remedy the situation failed. At the point of packing up, the command
link was nevertheless connected. To the experimenters’ great surprise, the system
worked perfectly: its bandwidth was a few cps and was insensitive to static.

Note. It has just been seen that an excess in bandwidth for a measuring instru-
ment results in a decrease in accuracy. Remembering that bandwidth and response
time are inversely proportional quantities, this can be interpreted as follows: when a
certain quantity is to be measured with given accuracy ε, the measurement must take
place at least over a certain time $T$, which becomes greater as ε becomes smaller.
This result can be expressed quantitatively in terms of information theory. It can
also be understood by intuition, since a certain time is necessary for random disturb-
ances, which are a source of error, to be averaged out.

12.3.5. Average Required Power. Let an electrical quadrupole or dipole be sub-
jected to a stationary random voltage $U(t)$ and current $I(t)$, or a mechanical quadri-
pole to a stationary random torque and angular velocity.$^1$ One may consider the
average value of the product

$$P(t) = U(t)I(t)$$

or, more generally, the correlation function

$$\varphi_{UI}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} U(t)I(t + \tau) \, dt$$

The value of $\varphi_{UI}(\tau)$ for $\tau = 0$ is the average value $P_m$ of $U(t)I(t)$

$$P_m = \varphi_{UI}(0)$$

which generalizes the concept of power dissipated for sinusoidal inputs.

It can be shown that

$$P_m = R\varphi_{II}(0) = R \int_{-\infty}^{+\infty} \Phi(\omega) \, d\omega$$

where $R$ is the resistance of the dipole, $\varphi_{ii}(\tau)$ is the autocorrelation function, and
$\Phi(\omega)$ is the frequency spectrum.

$^1$ Note that under these conditions the output is not necessarily random. In fact,
it would be periodic in the limiting case of systems with infinite resonance ratio $Q$. 
12.3.6. **Fluctuating Power.** The instantaneous value of the product $U(t)I(t)$ differs from its average value. The quantity

$$Q(t) = U(t)I(t) - P_m$$

has zero average value and represents the power that is not dissipated. This generalizes the concept of reactive volt-amperes in the sinusoidal case.

The correlation function $\varphi_{UI}(t)$, or $\varphi_{IC}(t)$ in the case of a mechanical quadripole, generally has the shape shown in Fig. 12-12. In fact, if the two input variables $U$ and $I$, or $\omega$ and $C$, varied at each instant in a similar manner, the correlation function would have the shape of an autocorrelation function with a peak at $\tau = 0$. This would be the case for a purely resistive electric network. But the presence of inductances and capacitances (or inertias and springs) results in two input variables not varying similarly at each instant. These variables are correlated, but the correlation is a maximum for a certain value $\tau_0$ which expresses a generalized lag of, for example, $I$ with respect to $U$, which generalizes for random functions the concept of phase shift for sinusoidal variables. This is equivalent to saying that, if the $I$-vs.-$t$ curve were shifted by $\tau_0$ sec, the correlation would be a maximum.

In conclusion, the autocorrelation function $\varphi_{II}(\tau)$ or $\varphi_{CC}(\tau)$ provides information concerning the average required and dissipated powers; and the cross-correlation function $\varphi_{UI}(\tau)$ or $\varphi_{IC}(\tau)$ enables one to generalize the concepts of acting and reacting power in sinusoidal systems.

12.4. **FUNDAMENTALS OF INFORMATION THEORY**

12.4.1. **Generalities.** The physical nature of the input to a system can be of many different types: force, torque, position, electric potential, luminous flux. Nevertheless, it is astonishing to realize the possible varieties of signals which describe the same input. One can, depending on the phenomenon one uses, record a temperature curve from an electrical (variation of resistance), mechanical (expansion), or optical detector. Similarly, the same concept may be expressed in different languages, i.e., by different words, sounds, or printed letters. These common-sense ideas were the starting point for C. E. Shannon's information theory, the fundamentals of which will be briefly outlined in the following paragraphs.

12.4.2. **Definition of the Quantity of Information.** If one attempts to define intuitively the quality of a message, it is realized that, the more the contents of the message are anticipated, the poorer it will be in information. If the message is just to say yes or no (the anticipation is 50 per cent), only a very simple signal is needed for its transmission—for instance, the emission of a dash for yes and a dot for no, or of an electric impulse for yes and the lack of impulse for no. On the contrary, the message "shall arrive tomorrow by car at 6:30 p.m." contains more information. There is less anticipation of its contents, and it will be harder to transmit. Therefore it is natural to define the quantity of information involved in a message by its degree of unpredictability.

Consider, to begin with, a particularly simple signal source, one that
can furnish only the two symbols yes and no, the production probabilities of the two symbols being equal. One says that, at each emission of a signal, the source gives a unit of information which is called a bit or binit (contractions of binary digit).

Consider now, as a further example, the message that consists of indicating a position \( x \). The position to be transmitted may be the input of a measuring device, or of a position-control system. In practice, the accuracy with which the position can be defined is finite, because noise is always present. The position \( x \) can be defined only within a certain small quantity which is often called the threshold due to noise (Sec. 22.3.5). As a result, \( x \) can have only a finite number of values that differ from one another by the threshold in exactly the same way that the number expressing a length can have only a finite number of values if, say, a millimeter is assumed to be smallest length discernible.

How many bits will be involved in such a message? If only two values are possible, defining \( x \) will be a one-bit message. If four \( (2^2) \) values are possible, defining \( x \) will amount to the following: First, specify whether the position \( x \) lies in the upper or in the lower half; this corresponds to 1 bit. Second, specify in what division of the retained half \( x \) lies; this is one more bit. Thus the total information will consist of 2 bits. More generally, it can be seen by similar reasoning that, if \( N \) values are possible, specifying the value of \( x \) involves \( \log_2 N \) bits, in other words, the number of figures needed to write \( x \) in binary notation. Thus, the quantity of information involved will be

\[
H = \log_2 N
\]

where \( N \) is the number of transmissible possibilities.

This definition can be extended to the case in which different transmissible positions have different probabilities of being transmitted. If \( p_i \) is the probability for the position \( i \) to be transmitted (in the above example, \( i \) ranges from 1 to \( N \) and \( p_i \) is equal to \( 1/N \), whatever \( i \) may be), then one is led to define the quantity of information involved when specifying a position as

\[
H = - \sum_{i=1}^{N} p_i \log_2 p_i
\]

For example, specifying a position to within 1 per cent in a domain of uniform probability density involves

\[
\log_2 \frac{100}{1} = 6.7 \text{ bits}
\]

Specifying it to within 0.1 per cent involves

\[
\log_2 \frac{1000}{1} = 9.8 \text{ bits}
\]

**12.4.3. Information Flow.** Up to now, only isolated signals have been considered. In practice, inputs of measuring instruments or control systems consist of many such signals that are to be transmitted with the
passing of time. If each signal involves a quantity of information of \( H \) bits, and if \( n \) signals are to be transmitted each second, it is said that the message has an information flow or an information speed of \( nH \) bits/sec.

For instance, suppose that a temperature between 0 and 100°C is to be recorded with a precision of 1° and is varying slowly enough that a signal every 2 sec will give sufficient information on the evolution of the phenomenon. The transmission of the corresponding message will involve a flow of information of

\[
\frac{1}{2} \times 6.77 = 3.4 \text{ bits/sec}
\]

If, however, it is desired to measure within 0.5° the position of an airplane near the horizon and to make the measurement 10 times per second, the number of bits involved in transmission of the corresponding message will be

\[
10 \times \log_{10} \frac{360}{0.5} = 95 \text{ bits/sec}
\]

12.4.4. Information Theory. The preceding notions establish the basis of information theory, which studies from an "informational" viewpoint all the problems involved in the transmission of messages. This theory, which is of great importance to the theory of communications, was initially applied to problems of telegraphic transmissions, especially to the development of an optimal code, that is, to the best methods from the informational viewpoint of transmitting the message. But because of the generality of the notion of information, this theory can shed light on a great variety of domains.

In particular, servo systems can be looked upon from the viewpoint of information theory. In fact, the most general definition of a servo system is the following: a system whose purpose is to transmit information with power amplification. This definition is seen to be equivalent to that given in Chap. 1 if it is recalled (Sec. 1.3.8) that feedback systems are used only when the inputs involve a certain degree of unexpectedness, that is, when information is to be transmitted (see Sec. 33.9.1).

12.4.5. Information Flow through a Linear System. One of the main interests of information theory is to describe quantitatively notions that are more or less evident qualitatively. However, such quantitative applications are beyond the scope of the present chapter. We shall restrict ourselves to indicating briefly how linear systems can be considered in the light of information theory. It has been seen in Chap. 11 that a linear system is fundamentally characterized by two quantities:

a. Its linear region from threshold \( e_m \) to saturation \( e_M \) with a saturation-to-threshold ratio \( e_M/e_m \).

b. Its bandwidth, i.e., the frequency range from zero to cutoff frequency \( F_c \) for systems that can roughly be described as low-pass filters.

When an input is fed into the system, the latter only distinguishes at a given instant \( e_M/e_m \) possible values for it (admitting that the threshold is the same all over the measuring domain). Furthermore, saying that the bandwidth is \( F_c \) cps is roughly equivalent to saying that inputs can be properly transmitted by the system at a rate of the order of \( F_c \) times per second. In conclusion, it can be said that the system is able to
transmit a quantity of information equal to $\log_2 \left( \frac{e_M}{e_m} \right)$ bits at a rate of $F_c$ times in a second. This will be expressed by saying that the system has an informational capacity

$$C = \log_2 \frac{e_M}{e_m} F_c \text{ bits/sec}$$

the flow of information that the system is capable of transmitting. Note that $C$ is the area of the utilization rectangle of the system (defined in Chap. 11; see Fig. 12-13) within the factor $20 \log_{10} 2 = 6$.

12.4.6. Informational Matching. Consider a servo system that is asked to follow its command, or a recording device that is used to study the evolution of a physical phenomenon. From the viewpoint of information theory, such a system constitutes a channel for the transmission of the information involved in the input. The message is characterized by an information flow $nH$ bits/sec. The system is characterized by its informational capacity $C = \log_2 \left( \frac{e_M}{e_m} \right) F_c$ bits/sec.

![Fig. 12-13. Information capacity of linear system is represented by shaded area.](image)

It can be shown that it is desirable to approach conditions where the capacity $C$ of the system is equal to the flow to be transmitted: $C = nH$. In fact:

a. Let $C < nH$. This means that the system is not able to transmit the message properly, owing to a lack of information capacity. This can happen in two ways. Either (1) $e_M/e_m$ is too small, then the system will clip off the high values of the input if $e_M$ is not sufficiently large and/or ignore its minute variations if $e_m$ is too great, or (2) $F_c$ is too low, then the system will filter out frequencies that ought to be transmitted without appreciable distortion, by lack of sufficient speed of response. In either case, the input message will not be transmitted properly.

b. Let $C > nH$. This means that the system would be eventually able to transmit more information per second than is actually called for. Either only a fraction of the total scale $e_M/e_m$ is taken advantage of or the system has a greater bandwidth than is necessary. (See Sec. 12.3.4 concerning the drawbacks of such a situation.) Thus, the system is not taken advantage of to the full extent of its possibilities. Hence the notion, first introduced by J. Loeb,\(^1\) of informational efficiency $\eta = nH/C$.

This analysis has the merit of presenting in a general and concise

manner the notion of adapting an instrument to a particular application. It can be extended somewhat further. The informational efficiency can be made to approach unity if the slope of the output-vs.-input characteristic of the first element of the system, instead of being constant, is chosen for each value of the input proportionally to the probability that the value will appear.

12.4.7. Importance of the Theory of Information. Information theory is of great importance in communications. It has the merit of extracting from a signal of any physical nature the message it transmits. The message is the fundamental concept, rather than the signal, whose form may change from one case to another. It would be difficult to find a better comparison than with the discovery of thermodynamics. This discovery revolutionized the concepts of heat engines. Before Carnot, people were interested only in the mechanics of the moving parts of the machines. Now, however, the notions of energy and entropy introduced by Carnot have become of primary importance and have permitted a complete understanding of the operation of heat engines of all types. It is the same with information theory; it permits the grouping together of phenomena heretofore considered individually and the defining quantitatively of concepts and methods previously considered only qualitatively and empirically.

The notion of information is capable of throwing new light on some aspects of servomechanism theory; for a servo system is, after all, just a system for transmitting information with power amplification. The functioning of a servomechanism is evidently often difficult to represent faithfully by means of abstract concepts because of the numerous limitations and complications of the mechanics. However, the application of the informational point of view is often successful.
PART TWO

LINEAR SERVO SYSTEMS

CHAPTER 13

TRANSFER FUNCTIONS AND TRANSFER LOCI
OF FEEDBACK CONTROL SYSTEMS

Summary

1. Transfer functions of feedback control systems.
2. Graphical interpretation of the equation \( H = KG/(1 + KG) \).
3. Open-loop gain adjustment.
4. Influence of time lags.

Notation. When it is possible throughout this chapter and the following ones, the following notation will be used:

- \( e(t), E(s) \) Main input which controls the system
- \( r(t), R(s) \) Output
- \( e(t) - r, \varepsilon(s) \) Error, or deviation between input and output
- \( d(t), D(s) \) Secondary input or disturbance

13.1. OPEN-LOOP AND CLOSED-LOOP TRANSFER FUNCTIONS

13.1.1. Definitions. 1. Transfer Function of a Feedback Control System. Let us consider any feedback control system in which a control input \( e(t) \) controls an output or response \( r(t) \) (Fig. 13-1). If the relationship between \( e(t) \) and \( r(t) \) is a linear differential equation with constant coefficients, which will be assumed in all of Part 2, a transfer function can be defined as

\[
H(s) = \frac{R}{E}(s)
\]

This function has a static gain \( H(0) \). Graphically, it is represented by the locus of the vector \( H(j\omega) \), with \( \omega \) variable from 0 to \( \infty \). This locus is the transfer locus of the feedback control system. The gain is obtained from the transfer locus as the value at \( \omega = 0 \). The maximum value of \( H(j\omega) \) is the peak or resonance ratio \( (|H(j\omega)|_{\text{max}}) \) and represents the maximum value of the transfer locus. The resonant frequency \( \omega_R \) is the value of frequency at which the maximum occurs.

2. Open-loop Transfer Function. To study the functioning parts of the feedback control system, the system may be represented by a block diagram (Fig. 13-2) which presents the shape of a closed loop and consists of:

215
a. Forward path, with the error \( \varepsilon(t) = e(t) - r(t) \) as an input and the same output \( r(t) \) as the whole feedback control system.

b. Feedback loop, which feeds back the value of the output \( r(t) \) so as to compare it with the control input \( e(t) \) in the case of unity feedback.

The differential relationship between \( \varepsilon(t) \) and \( r(t) \) can be written as the transfer function of the forward path

\[
KG(s) = \frac{R}{\varepsilon}(s)
\]

By definition, this is the open-loop transfer function of the feedback control system. As opposed to this definition, the transfer function of the whole feedback control system is called the closed-loop transfer function or system function.

![Fig. 13-2.](image)

3. Open-loop Gain. As for all transfer functions, one is in the habit of putting the gain into evidence by writing the open-loop transfer function as \( KG(s) \). The \( K \) coefficient, to be defined in a moment, is termed the open-loop gain of the feedback control system. If the open-loop transfer function has no integration, then

\[
KG(s) = K \frac{1 + a_1s + \cdots + a_ms^m}{1 + b_1s + \cdots + b_ns^n}
\]

\( KG(0) = K \)

\( K \) is, then, the gain of the system when the feedback is discontinued. It is an abstract number, since \( r \) and \( \varepsilon \) are of the same nature. Its range of values varies considerably with the problem under consideration. In an autopiloted aircraft, its value is about 1, but for most systems the value is generally much higher.

If the open-loop transfer function has one integration, the open-loop gain is obtained from

\[
KG(s) = \frac{K}{s} \frac{1 + a_1s + \cdots + a_ms^m}{1 + b_1s + \cdots + b_ns^n}
\]

\( \lim_{s \to 0} sKG(s) = K \)

For positional servo systems in which \( e, r, \) and \( \varepsilon \) are linear or angular positions, it is often said that \( K \) is the open-loop velocity gain of the feedback control system.

Similarly, if there are two integrations, the open-loop gain is defined by

\[
\lim_{s \to 0} s^2KG(s) = K
\]

For positional servos, this is the open-loop acceleration gain. The influence of these quantities on the performance of feedback control systems will be shown at the end of this chapter and in the following one.
Fig. 13-3. Elementary second-order servomechanism.

Fig. 13-4. Nyquist loci for \(1/s(1 + s)(1 + as)\).

Note that the open-loop gain is a dimensionless factor if there is no integration. In the other cases, it has the dimension of a negative power of time.

13.1.2. Example. Consider the angular-positional servo system of Fig. 13-3, which is designed to control the angular displacement of a load, characterized by an inertia \(J\) with a viscous friction \(f\), by means of an amplifier stage producing a torque proportional to the error \(\Gamma = C\varepsilon\). The output angle \(r(t)\) is given by

\[
\Gamma = C\varepsilon = J \frac{d\tau}{dt} + f \frac{dr}{dt}
\]

Hence the closed-loop transfer function, or system function, is

\[
H(s) = \frac{R}{E}(s) = \frac{C}{Js^2 + fs + C}
\]
and the open-loop transfer function is

\[ KG(s) = \frac{R}{e}(s) = \frac{C}{J_0s^2 + f_0} \]

that is,

\[ KG(s) = \frac{C/f}{s[1 + (J/f)s]} \]

This is, therefore, a second-order system, with a gain \( H(0) = 1 \), an undamped natural frequency \( \omega_n = (C/J)^{1/2} \), and a damping ratio \( \zeta = f/2(CJ)^{1/2} \). Hence, the considerations developed in Chap. 6 can be applied to it.

The open-loop transfer function has one integration, hence a velocity gain of \( K = C/f \).

The open-loop transfer locus is the locus, \( \alpha = 0 \), of Figs. 13-4 (Nyquist locus) and 13-5 (Nichols locus).

---

**13.1.3. Significance of the Two Transfer Functions.** The function of interest to the user is the closed-loop function, because it describes the over-all performance of the servo system. The servomechanism engineer, on the other hand, is primarily concerned with the internal behavior of the system, and hence with its open-loop transfer function. In practice,
the design of a feedback control system consists of a cascade of elements which control the output so as to correct the error. Accordingly, it is natural to study the open-loop transfer function. Moreover, the influence of variations in the value of the parameters appears more clearly in that function.

13.1.4. Relation between the Two Transfer Functions, Fundamental Formula. Eliminating \( \varepsilon(s) \) between

\[
R(s) = KG(s)\varepsilon(s) \quad \text{and} \quad \varepsilon(s) = E(s) - R(s)
\]

yields (Sec. 7.5.6) the very important relation between the closed-loop and the open-loop transfer functions

\[
H(s) = \frac{KG(s)}{1 + KG(s)}
\]

In terms of inverse transfer functions, one can write

\[
H^{-1}(s) = (KG)^{-1}(s) + 1
\]

Eliminating \( R(s) \), one obtains the error for a given input

\[
\varepsilon(s) = \frac{1}{1 + KG(s)} E(s)
\]

which can be regarded as the definition of an error-input transfer function equal to \( 1/[1 + KG(s)] \).

![Fig. 13-6. Nonunity feedback.](image)

13.1.5. Case of a Transfer Function in the Feedback Path. A feedback control system with an element in the feedback path is a system in which one detects not the difference between input and output, but the difference between the input and a quantity related to the output by a functional equation, here supposed to be linear (Fig. 13-6). The forward path is thus actuated by

\[
\varepsilon(s) = E(s) - R(s)F(s)
\]

where \( F(s) \) is the transfer function of the feedback loop. This being so, one has

\[
R(s) = KG(s)[E(s) - F(s)R(s)]
\]

whence the closed-loop system function
\[ H(s) = \frac{R(s)}{E(s)} = \frac{KG(s)}{1 + KG(s)F(s)} \]

In terms of inverse transfer functions, this can be written as

\[ H^{-1}(s) = (KG)^{-1}(s) + F(s) \]

As will be seen later, the study of the feedback control system with a transfer function in the feedback loop can be reduced to that of a feedback control system with unity feedback.

![Fig. 13-7. Longitudinal control of aircraft.](image)

**Example.** Figure 13-7 represents the block diagram for a longitudinal aircraft control system. The sensing device compares the aircraft pitch angle \( \theta \), given by a pitch indicator, to the pilot's order \( \epsilon \). The transducer and controller convert this error into a voltage \( v \) which, by means of a servomotor \( M \), deflects the rudder by an angle \( \delta \), hence producing a pitch angle \( \theta \). Setting

\[ H_e = \frac{V}{\epsilon}, \quad H_M = \frac{\Delta}{V}, \quad H_A = \frac{\Theta}{\Delta}, \quad H_i = \frac{\Theta}{\Theta} \]

one has

\[ KG = K_e H_M H_A \]

whence

\[ H^{-1} = H_e^{-1} H_M^{-1} H_A^{-1} + H_i \]

This equation can also be written with direct functions:

\[ H = \frac{H_e H_M H_A}{1 + H_e H_M H_A} \]

**13.1.6. Multiple Loops.** When many loops are superposed, the over-all closed-loop transfer function is easily obtained by the relations between inverse transfer functions.\(^1\) For example, the inverse system function for the system shown on Fig. 13-8 can be read directly from the block diagram:

\[ H^{-1} = \frac{E}{R} = H + A^{-1} F^{-1}[\Delta G + B^{-1} E^{-1}(D + C^{-1})] \]

**13.1.7. Introduction of Disturbances.** 1. If \( d(t) \) is a disturbance which is introduced in the forward path, as shown in Fig. 13-9, where \( B = A + D \), one has

\[ R(s) = K_1 G_1(s) K_2 G_2(s) E(s) + K_2 G_2(s) D(s) \]

whence

\[ R = \frac{K_1 G_1 K_2 G_2}{1 + K_1 G_1 K_2 G_2} E + \frac{K_2 G_2}{1 + K_1 G_1 K_2 G_2} D \]

\(^1\) Examples can be found in P. Naslin, “Les systèmes asservis,” p. 112, Revue d’Optique, Paris, 1951.
Similarly, one has

$$
\varepsilon(s) = \frac{1}{1 + K_1G_1K_2G_2} E(s) - \frac{K_2G_2}{1 + K_1G_1K_2G_2} D(s)
$$  \hspace{1cm} (13-1)

For a regulator \((E = 0)\) this is often referred to as the *regulator transfer function*

$$
\varepsilon(s) = \frac{-K_2G_2}{1 + K_1G_1K_2G_2} D(s)
$$  \hspace{1cm} (13-2)

2. More generally, if the disturbance is introduced somewhere in order to have as an open-loop transfer function \(R/D = K_3G_3\) (Fig. 13-10), one can write

$$
\varepsilon(s) = \frac{1}{1 + K_1G_1K_2G_2} E(s) - \frac{K_3G_3}{1 + K_1G_1K_2G_2} D(s)
$$  \hspace{1cm} (13-3)

For a regulator \((E = 0)\), this can be written

$$
\varepsilon(s) = \frac{-K_3G_3}{1 + K_1G_1K_2G_2} D(s)
$$  \hspace{1cm} (13-4)
Note that in these equations $K_s G_s$ takes the place of $K_s G_2$ of Eq. (13-1). The product $K_s G_1 K_s G_2$ is always the transfer function of the open-loop feedback control system.

Note 1. It is easily seen from Eq. (13-1) that the error due to a disturbance $D(s)$ (Fig. 13-9) is the same as the error due to a disturbance $K_s G_2 D(s)$ applied after the $K_s G_2$ box, or to a disturbance $D(s)/K_s G_1$ applied before the $K_s G_1$ box (Fig. 13-11). This property can be very useful for many applications.

![Figure 13-11. These two diagrams are equivalent to Fig. 13-9.](image)

Note 2. It is to be noted that all transfer functions pertaining to the system, including servo transfer functions ($R/E$ and $E/E$) as well as regulator transfer functions ($R/D$ and $E/D$), have the same characteristic equation, $1 + K_s G_1 K_s G_2 = 0$, and hence the same nature of stability. This property can be looked upon from a philosophical viewpoint. When a feedback control system is subjected to any disturbance, its reaction always involves the characteristics of the whole system. For the case shown in Fig. 13-12, in which a disturbance $D$ is introduced close to the very right-hand end of the forward path, the whole system takes part in the response, although the disturbance "seems" to result in a response $R(s)$ simply because of $K_s G_1(s)$:

$$R(s) = \frac{H_s G_4(s)}{1 + K_s G_1(s) K_s G_2(s) K_s G_3(s) K_s G_4(s)}$$

![Figure 13-12.](image)

This is similar to what happens when a living being is subjected to a stress: however superficial the stress may seem to be, the organism as a whole will be involved in the resulting phenomena. Any person reacts according to his total personality, just as any response of a feedback control system involves all the components as combined in the characteristic equation.

Note 3. It is sometimes advantageous to resort to flow graphs rather than block diagrams. The concept of flow graphs, or signal flow diagrams, was devised and
applied to servo problems by S. Mason. It does not give results which one could not obtain by more classical methods. But flow graphs involve more advanced thinking than do block diagrams, and they can be helpful when handling complicated systems.

13.2. HARMONIC APPROACH. GRAPHICAL INTERPRETATION OF $H = KG/(1 + KG)$

13.2.1. Determining $H(s)$ from the Open-loop Transfer Locus. Let $M$ (Fig. 13-13a) be an ordinary point (angular frequency $\omega$) on the locus $KG(j\omega)$. Let $M'$ be the point on the locus $H(j\omega)$ for the same frequency $\omega$. They are related by

$$H(j\omega) = \frac{KG(j\omega)}{1 + KG(j\omega)}$$

$1 + KG(j\omega)$ represents the vector $AM$, $A$ being the point $(-1)$.† Hence the equation means that

$$OM' = \frac{MO}{MA} \quad (Ox,OM') = (Ox,OM) - (Ox,AM)$$

Hence

$$(Ox,OM') = 2\pi - (MO,MA)$$

This enables one to locate $M'$, and therefore to sketch the locus $H(j\omega)$.

To do this, it is very helpful to draw the loci

$$\frac{MO}{MA} = \lambda \quad \text{const} \quad OMA = \phi \quad \text{const}$$

The constant-$\lambda$ loci are defined in Table 13-1 and plotted in Fig. 13-13b. They form a family of circles. For $\lambda = 0$ the corresponding circle is the limit point $0$, for infinite $\lambda$ it is the limit point $A$. For $\lambda = 1$


2 Servo problems are, in general, not complicated enough to bring out the advantage of the flow-graph approach. In the authors’ opinion, the most fruitful field of application for flow graphs is that of electrical networks and their analogs (e.g., lattice constructions). The technique of flow graphs enables one to formulate the response of any variable of a complicated linear system to an input applied anywhere to it much more rapidly than by use of conventional methods. See M. Boisvert and L. Robichaud, "Direct Analysis of Electrical Networks," Rapport de Recherches 9, Laval University, Québec, 1956, and other publications by Dr. Boisvert in the Annales des Télécommunications: "Diagrammes de fluence: théorie élémentaire," "Diagrammes de fluence et topologie," and "Diagrammes de fluence et théorie de la réaction," 13(3/4):50–77 (1958).

† Called critical point for reasons which will be given later.
### Table 13-1. Geometrical Definition of Constant-\( \lambda \) Circles (Hall’s Chart)*

<table>
<thead>
<tr>
<th>( \lambda = \frac{MO}{MA} )</th>
<th>Center ( \frac{\lambda^2}{\lambda^3 - 1} )</th>
<th>Angle ( \psi = \arcsin \frac{1}{\lambda} ) (see Fig. 13-23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>+0.10</td>
<td>+</td>
</tr>
<tr>
<td>0.4</td>
<td>+0.19</td>
<td>+</td>
</tr>
<tr>
<td>0.5</td>
<td>+0.33</td>
<td>+</td>
</tr>
<tr>
<td>0.6</td>
<td>+0.56</td>
<td>+</td>
</tr>
<tr>
<td>0.7</td>
<td>+0.96</td>
<td>+</td>
</tr>
<tr>
<td>0.8</td>
<td>+1.78</td>
<td>+</td>
</tr>
<tr>
<td>0.9</td>
<td>+4.26</td>
<td>+</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>1.1</td>
<td>-5.77</td>
<td>65</td>
</tr>
<tr>
<td>1.2</td>
<td>-3.27</td>
<td>56.5°</td>
</tr>
<tr>
<td>1.3</td>
<td>-2.45</td>
<td>50</td>
</tr>
<tr>
<td>1.4</td>
<td>-2.04</td>
<td>45.5°</td>
</tr>
<tr>
<td>1.5</td>
<td>-1.80</td>
<td>42</td>
</tr>
<tr>
<td>1.6</td>
<td>-1.64</td>
<td>38.5°</td>
</tr>
<tr>
<td>1.7</td>
<td>-1.53</td>
<td>36</td>
</tr>
<tr>
<td>1.8</td>
<td>-1.47</td>
<td>33.5°</td>
</tr>
<tr>
<td>1.9</td>
<td>-1.38</td>
<td>31.7°</td>
</tr>
<tr>
<td>2.0</td>
<td>-1.33</td>
<td>30</td>
</tr>
<tr>
<td>2.25</td>
<td>-1.24</td>
<td>26.5°</td>
</tr>
<tr>
<td>2.50</td>
<td>-1.19</td>
<td>23.5°</td>
</tr>
<tr>
<td>2.75</td>
<td>-1.15</td>
<td>21.5°</td>
</tr>
<tr>
<td>3.00</td>
<td>-1.12</td>
<td>19.5°</td>
</tr>
<tr>
<td>3.50</td>
<td>-1.10</td>
<td>16.5°</td>
</tr>
<tr>
<td>4.00</td>
<td>-1.07</td>
<td>14.5°</td>
</tr>
<tr>
<td>5.00</td>
<td>-1.04</td>
<td>11.5°</td>
</tr>
</tbody>
</table>


the locus has infinite radius and is a straight line parallel to the imaginary axis at a distance \(-0.5\). The constant-\( \Phi \) curves are their orthogonal trajectories, they are circles passing through \( O \) and \( A \). The chart just obtained is called the Hall chart.

If the \( KG(j\omega) \) locus is sketched as an amplitude-phase plot, the loci

\[
\frac{MO}{MA} = 20 \log \lambda \text{ db} \quad OMA = \Phi \text{ degrees}
\]

can be drawn for constant \( \lambda \) and \( \Phi \). One thus obtains the Nichols chart, shown in Fig. 13-13c. One of these charts is given in Chart 3, at the back of the book. The \( \lambda = \) constant contours are called the Nichols contours.¹

¹ For practical use, one should cut out a cardboard or paper template of the 2.3-db contour (corresponding to \( Q = 1.3 \)). Its use will be shown later on.
Fig. 13-13c. The Nichols chart. (Twofold reduction of Chart 3 at the back of the book.)
The Nichols and Hall charts have similar uses. They are transforms of each other.

13.2.2. Application to the Steady State. The value of the closed-loop system function for the steady state is inferred from the position of point $M_0 (\omega = 0)$ on the $KG$ locus. Its magnitude is given by $M_0 O / M_0 A$.

Case 1. The $KG$ function has at least one integration (Fig. 13-14). Point $M_0$ is then at infinity (Sec. 8.1.1, par. 5) and

$$H(0) = \frac{M_0 O}{M_0 A} = 1$$

That is to say, for the static behavior, $r(t) = e(t)$, that is, there is no error; the system is said to involve zero position error. Examples of a feedback control system with one integration are the second-order system

$$\frac{H(0)}{H(j\omega)}$$

or

$$\frac{KG(j\omega)}{H(0)}$$

Fig. 13-14.
defined in Sec. 13.1.2 and a roll-controlled aircraft without dihedral effect.

Case 2. The KG function has no integration (Fig. 13-15). The point $M_0$ is then on the real axis at a finite distance $K$. Hence

$$H(0) = \frac{K}{K + 1}$$

For this case, there is a steady-state error equal to $1/(K + 1)$. In particular, for the steady response to a step input (Fig. 3-13), one has $r(t) \neq e(t)$. The system is said to involve a position error which is proportional to $1/(K + 1)$.

![Fig. 13-15.](image)

Examples of a feedback control system without integration are the system described in Sec. 13.1.2 when a restoring torque is added at the output, and a pitch-controlled aircraft.

The above important results\(^1\) can be expressed as follows:

1. The steady-state error for a step input, or position error, decreases when the open-loop gain is increased.

2. The position error is suppressed by the presence of an integration\(^2\) in the forward path.

13.2.3. Application to the Resonance. At resonance, $|H(j\omega)|$ or $MO/MA$ is a maximum (Fig. 13-16). This occurs at the one or more frequencies for which the $KG(j\omega)$ locus is tangent to a circle of the family $MO/MA = const$. The corresponding resonance ratio is the $\lambda$ of the circle. Similarly, on the Nichols chart the resonance corresponds to tangency to a Nichols contour. One resonance is shown in Fig. 13-17a, two in Fig. 13-17b.

![Fig. 13-16.](image)

It can be seen that, the closer the $KG$ locus approaches point $A$, the less damped is the system. When the $KG$ locus is too near point $A$, the servo tends to oscillate; this is called hunting, in a broad sense. (For this concept of hunting, see Sec. 16.2.3.) If this same $KG$ locus goes through point $A$, one has for the corresponding frequency

$$1 + KG(j\omega) = 0$$

\(^1\) A generalization will be made in the following chapter.

\(^2\) Which is equivalent to an infinite static gain.
In other words, the system oscillates; there is *hunting* in a strict sense. Finally, it will be proved in Chap. 16 that, if the $KG$ locus goes on the other side of point $A$, the servo system is unstable. For these reasons, the point $A$ is termed the *critical point*.

*Note.* It is important to distinguish between the *resonant frequency* of a closed-loop system, i.e., the frequency at which $|KG/(1 + KG)|$ is a maximum, and any frequency at which $|KG|$ is maximum, which constitutes a resonant frequency for the open-loop system. To avoid confusion, the former frequency is usually called the resonant (or natural) frequency of the servo system, whereas the latter is called its *open-loop resonant, or natural, frequency*.

In general, the resonant frequency of the servo is *higher* than its open-loop resonant frequency, their difference being greater when the open-loop gain is high. Figures 13-18 and 13-19 show this better than an explanatory text could.

*Example 1.* Figure 13-18 shows

$$KG = \frac{K}{1 + (2\pi/\omega_n)s + s^2/\omega_n^2}$$

using the Hall chart.
Example 2. Figure 13-19 shows the transfer locus of a missile in vertical flight, using the Nichols chart.

13.2.4. Bandwidth of a Servo System. For most servomechanisms the transfer function $KG(s)$ has the following properties:

a. At low frequencies, $|KG(j\omega)| \gg 1$. This is obvious when $KG(s)$ implies integration; it is also generally true when $KG(s)$ does not imply integration, because $K$ is much greater than unity in most servomechanisms.\(^1\)

b. At high frequencies, $|KG(j\omega)| \to 0$. This is common to all mechanical systems because of inertia (Sec. 8.1.1, par. 4).

As a consequence:

\[
\begin{align*}
a: & \quad |H(j\omega)| \approx 1 \quad \text{at low frequencies} \\
b: & \quad |H(j\omega)| \approx 0 \quad \text{at high frequencies}
\end{align*}
\]

Thus, most servomechanisms can be roughly considered as low-pass filters (Secs. 6.4.5 and 8.3.5).

The bandwidth is the frequency range $(0, \omega_c)$, where $\omega_c$ is generally taken as the 6-db cutoff frequency, i.e., the frequency at which the magnitude is 6 db down from the zero-frequency value. $\omega_c$ is the frequency at which the $KG(j\omega)$ locus intersects the $\lambda = 0.5$ circle on the Hall chart, or the $\lambda = -6$-db contour on the Nichols chart; it can be read directly off the open-loop Nyquist or Nichols plot.

Use is also often made of the frequency where $KG(j\omega)$ is unity; this frequency is called the gain crossover frequency, and it can be read directly from the Nyquist or Nichols open-loop plot. It lies at the intersection of the $KG(j\omega)$ locus with the circle of unit magnitude centered at the origin, or of the Nichols plot with the 0-db horizontal line.

13.2.5. Transient Characteristics. Knowledge of the open-loop transfer function completely determines the behavior of the $H(j\omega)$ function. Thus information about the transient state of the system can be deduced from the inspection of the open-loop transfer locus.

1. To begin with, the size of the bandwidth determines the approximate duration of the transient state (Secs. 6.4.5 and 8.3.5).

2. Then the resonance sharpness, measured by the resonance ratio $Q$, indicates the damping ratio of the transient state (Sec. 8.3.4). As has been explained in Sec. 8.3.9, these considerations are valid only for “regular systems,” the $KG$ locus of which has a behavior comparable with that of a second-order feedback control system in the region where it approaches the critical point. This is the general case for positional servomechanisms. In the other cases, one must consider the problem with greater care, according to each specific case. It is then possible, for a given class of problem, to draw out a group of common properties which are visible on the $KG$ locus. A typical example is the problem of automatic aircraft control, with its two modes corresponding to the two movements of the aircraft (center of gravity and about the center of gravity).

\(^1\) A noteworthy exception is a piloted aircraft, for which $K$ is of the order of unity.
3. Finally, it will be shown in Chap. 16 how the position of the $KG$ locus with respect to the critical point indicates the stability of the system by means of the left-hand criterion, or of Nyquist's criterion.

4. Quantitatively, one can compute the whole transient response of the feedback control system from the $KG(j\omega)$ locus by use of Floyd's or Proust's method (Sec. 8.3.7), starting with the magnitude and the phase shift of $KG/(1+KG)$ given by the Hall or the Nichols chart.

13.3. OPEN-LOOP GAIN ADJUSTMENT

13.3.1. Effect of the Open-loop Gain upon the Transient Performance when all the Other Parameters Are Fixed. Consider the following problem: One has a feedback control system with an open-loop transfer function $KG(s)$, having zero, one, or more integrations. Let the only free parameter be the gain $K$, the other elements of the servo, hence the $G(s)$ function, being fixed. One intends to study the effect of variations of $K$ upon the performance of the servo system. This problem must always be solved in the course of the design of a feedback control system. It is essentially the problem of setting the gain of the amplifier, which is always adjustable, when the other elements of the system are fixed. The study is easy when one reasons from the Hall or the Nichols chart.

To change the open-loop gain $K$ on Hall's chart is equivalent to transforming the $KG(j\omega)$ locus from the original $G(j\omega)$ locus corresponding to $K = 1$. Typical cases are shown in Fig. 13-20. It is seen that the $KG$ locus approaches the critical point when the gain increases. For systems
of an order higher than two, it can even pass on the other side of the critical point for values of $K$ larger than the critical one. Hence increasing the open-loop gain ends in creating hunting and destabilizing the system. This result is constantly verified by experience.¹

Note. In order to simplify the graphical discussion, one only draws that open-loop transfer locus which corresponds to $K = 1$. Changing the open-loop gain $K$ is then equivalent to taking the critical point at $-1/K$ and studying its position with respect to the fixed locus. For small values of $K$ the critical point lies at great distance on the negative real axis. As $K$ increases, the critical point approaches the locus, which may cause hunting or instability (Fig. 13-21).

13.3.2. The Stability-Accuracy Dilemma. From other considerations, it is known that, the larger the open-loop gain, the stiffer the servo and the better its performance at low frequencies, especially the steady-state or static accuracy. This is obvious for a feedback control system without

![Diagram showing increasing open-loop gain](image)

Fig. 13-21. Increasing the open-loop gain.

integration (Sec. 13.2.2), and we will see in the following chapters that it is also true for a feedback control system with integration. In conclusion, the actual setting of the open-loop gain results from a compromise. Either one chooses $K$ small in order to be safe with respect to stability, whence the servo is "soft" and not very accurate, or one stiffens the servo by increasing $K$ in order to improve the static accuracy, but this causes hunting. This is the old stability-accuracy, or stiffness-hunting, dilemma, which servomechanism experts knew long before the concept of the transfer function was applied to feedback-control-system analysis.

Example. For a second-order feedback control system $KG = K/(1 + Ts)$, the effect of the gain setting upon the resonance ratio and the transient response is shown by Table 13-2 and Fig. 13-22. The value $K = 1.4/T$, which yields a resonance frequency $\omega_R = 1.1/T$ and corresponds to $Q = 1.3, \zeta = 0.43$, is a reasonable compromise for most practical applications.

¹ As theory shows that a second-order system cannot become unstable, no real servo is represented correctly by means of a second-order block diagram. We will see a little later on that one of the principal reasons for this is the presence in the forward path of unwanted time lags, neglected when the equations yielding the second-order system were written.
### Table 13-2. Gain Adjustment for Second-order Servo

<table>
<thead>
<tr>
<th></th>
<th>Setting 1 (K small)</th>
<th>Setting 2 (K medium)</th>
<th>Setting 3 (K large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>$0.8/\tau$</td>
<td>$1.4/\tau$</td>
<td>$4/\tau$</td>
</tr>
<tr>
<td>$Q$</td>
<td>1.1</td>
<td>1.3</td>
<td>2.0</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>0.57</td>
<td>0.43</td>
<td>0.24</td>
</tr>
<tr>
<td>$\tau_0$</td>
<td>0.53</td>
<td>0.95</td>
<td>1.9</td>
</tr>
</tbody>
</table>

#### 13.3.3. Practical Gain Adjustment

How to settle the stability-accuracy dilemma depends on the case in point. However, for most cases one is satisfied with the compromise consisting in the choice of an open-loop gain which gives for the closed-loop system a *resonance ratio* $Q$ equal to about 1.3; that is to say,

$$1.2 < Q < 1.5$$

This gain will be written $K_{1.3}$. It is sometimes improperly called "optimum gain" (implication: for the stability-accuracy compromise). As this problem has great practical importance, solutions will be successively shown for Hall’s and Nichols’ charts.

1. Using Hall’s chart, it is easy to note for each value of $K$ which circle $MO/MA = \lambda$ is tangent to the corresponding KG locus. There is a different circle for each value of the gain $K$, hence a different value of $Q$. One chooses the value of $K$ which yields $Q = 1.3$.

The usual practice is to plot only the $G(j\omega)$ locus. The circle corresponding to the desired $Q$ varies with respect to the origin. Table 13-1 is very helpful for this; it gives, for different values of $\lambda$, the center abscissa, the radius, and the $\psi$ angle. The $\psi$ line is drawn first (Fig. 13-23), and then, by trial-and-error procedure, the circle centered on $Ox$ and tangent to the $\psi$ line and to the $G(j\omega)$ locus is found. The gain is determined by noting that the center abscissa is the number given in the table. Or, the perpendicular can be dropped from the point of tangency of the circle with the $\psi$ line: the abscissa is equal to $1/K$. 

---

**Fig. 13-22. Influence of $K$ on transient:**

(a) $K$ small, (b) $K$ medium, (c) $K$ large.
**Example.** Figure 13-24 represents the $G(j\omega)$ locus. It is desired to set the gain for a $Q$ of 1.5. The circle tangent to the locus and to the 42° line has its center at $-3.68$. As $M/(M^2 - 1) = 1.8$, one has $K_{1,1} = 1.8/3.68 = 0.49$. More simply, tangency of the circle and the 42° line occurs at $2.04$ of abscissa, whence $K_{1,1} = 1/2.04 = 0.49$.

**Remark.** If inverse loci are used, the technique is still easier. The resonance ratio is equal to the reciprocal of the radius of the circle centered at the critical point $A$ and tangent to the $1/KG(j\omega)$ locus (Fig. 13-25).

2. On the *Nichols chart* a variation of the open-loop gain corresponds to a vertical translation of the $KG$ locus, with upward motion for an increase in gain. To set the gain, the $G(j\omega)$ locus is sketched, and one writes down the decibels necessary to bring the locus tangent to the contour $\lambda = 2.3$ db, which is the transform of the circle $MO/MA = 1.3$. Usually, only the $G(j\omega)$ locus is plotted on graph paper. A $\lambda = 2.3$ db contour template is vertically moved on the Nichols chart until it is
tangent to the $G$ locus. The difference in decibels between the origin of the contour and the 0-db ordinate of the plane is then measured. This difference is the optimum gain $K_{1.3}$ (greater than 1 if the contour has been moved downward). The only necessary equipment is graph paper and the 2.3-db contour, which serves as a substitute for the Nichols chart.

If the studied feedback control system differs too much from a regular one, or if one wants to see on the Nichols chart what the performances are like, it is better to proceed as follows, using the whole chart. First sketch the $G(j\omega)$ locus on transparent graph paper, then superimpose the transparent plot on a Nichols chart and slide it vertically with respect to the chart until the locus is tangent to the $\lambda = 2.3$ db contour, all of its points being exterior to the contour. The $K_{1.3}$ gain is the necessary translation as measured in decibels.

**Important Practical Remark.** When one sketches $G(j\omega)$, one does not take into account the gains of the different stages which form the forward

![Fig. 13-25.](image_url)

![Fig. 13-26. Open-loop gain adjustment using the Nichols chart.](image_url)
path (they are all made equal to 1). The translation carried out on the $G(j\omega)$ locus in order to make it tangent to the $Q = 2.3$ db contour gives immediately the total open-loop gain corresponding to this resonance ratio. This total gain is the product of the partial gains, as stated above; but one is not concerned with the latter when working on the Nichols chart.

**Example** (solved with the first method). Figure 13-26 shows the open-loop transfer locus $G(j\omega)$ of a stabilized missile in vertical flight after launch, with $K = 1$. The transfer function accounts, therefore, for the gyroscopic sensing device, the compensating network, the autopilot, and the dynamics of the missile. To bring the 2.3-db contour into tangency with the locus, it is necessary to move the template vertically upward, and there is a distance of 6 db between the origin of the contour and the 0-db ordinate of the plane. This corresponds to the setting $20 \log K_{1,3} = -6$ db, that is, $K_{1,3} = 0.5$.

**13.3.4. Consequence.** A given function $G_1(s)$ gives rise to an optimum value $K_1$ of the open-loop gain. Another function $G_2(s)$ would give rise to another value $K_2$. If $K_2 > K_1$, it is natural to say that the $G_2(s)$ function is better than the $G_1(s)$ function with respect to gain, that is to say, with respect to the static accuracy. Therefore, one can improve the performance of a feedback control system by modifying the frequency-dependent part $G(s)$ of its open-loop transfer function. This can be done by inserting a transfer function $J(s)$ into the forward path so that the new function $G(s) \times J(s)$ has a higher optimum open-loop gain. This idea will be stated more precisely and developed in Chap. 18, under Compensation of Feedback Control Systems.

**13.4. INFLUENCE OF TIME LAGS**

**13.4.1. The Problem.** When the equations of a system are written, some elements are always neglected, for they are not accurately known or are difficult to handle in the calculations. But even if one does not consider them completely in the calculations, one must have an idea of the modifications they cause in the computed performance of the system. A particularly important case in the field of feedback control systems is that of the varied time lags which can be found in almost all stages and are often difficult to work out.

One therefore writes an open-loop transfer function $KG(s)$ which neglects some unwanted time lags and works with this function. More particularly, the gain is set assuming that the transfer function is exact. It is then necessary to know the effect of the neglected time lags.

**13.4.2. Graphical Study.** Suppose that the expression $KG(j\omega)$ for the open-loop transfer locus has been obtained by neglecting some time lags. Also, suppose $K$ is set at its optimum value. The locus is therefore tangent to the $\lambda = 1.3$ circle at $\omega = \omega_R$ and is completely exterior to it (Fig. 13-27). Now, suppose that the neglected time lags can be represented by a time constant $1/(1 + Ts)$. Generally $1/T$ is much larger than $\omega_R$, so that the resonant frequency of the system is on the shaded part of the locus $1/(1 + jT\omega)$ sketched in Fig. 13-28. Hence, it can be
seen that the time lags essentially cause a phase shift

$$\Phi = -\arctan T\omega_R \approx -T\omega_R$$

in the resonance region. For a regular servo system the result is that the KG locus approaches the critical point by entering into the $\lambda = 1.3$ circle (Fig. 13-29). The result is a decrease of the damping and of the

![Fig. 13-27.](image)

![Fig. 13-28. Unwanted time constant. Resonant frequency of servo system lies in the shaded part of locus.](image)

![Fig. 13-29. Effect of (a) pure time lag $e^{-Tt}$ and (b) time constant $1/(1 + Ts)$ on servo. Note that, if $T$ is small, both are equivalent in the resonance region; that is, small lags can be approximated by time constants so far as gain adjustment is concerned.](image)

stability. If the phase shift is large, the locus may even cross the critical point, therefore producing instability.

Generally speaking, **time lags increase the risk of hunting.** A more accurate statement can be made by noting the $\lambda$ of the circle tangent to the open-loop locus. This can show to what extent there is loss of damping.

**Remark.** If KG is a second-order function, the transfer locus is tangent to the negative real axis at the point $O$. Hence, the system is stable whatever the gain, as the $KG(j\omega)$ locus cannot cross the critical point. This result disagrees with experience. On the contrary, if the time lags are accounted for, the open-loop transfer
locus is tangent to the positive imaginary or real axis at point $O$, and there is a risk of instability if the gain is increased. This result agrees with experience. As a matter of fact, there are no second-order servos, for time lags are always present.

13.4.3. Application to Second-order Systems. Consider the second-order feedback control system with the following transfer function, where nondimensionalized variables are used: $G(s) = 1/(s + 1)$. Also suppose the presence of an unwanted time lag $1/(1 + Ts)$. The locus $1/s(1 + s)(1 + Ts)$ is sketched in Nyquist and Nichols coordinates (Figs. 13-4 and 13-5). It is easily shown that $T$ modifies the resonance frequency only slightly, but that the phase shift increases with $T$, while the damping ratio decreases (Table 13-3). If a maximum $Q$ of 1.5 is permissible, 0.1 is therefore the tolerable upper boundary for $T'$.

1 This is a particular result, which must not be generalized.
CHAPTER 14

POLE-ZERO-CONFIGURATION APPROACH:
THE ROOT-LOCUS METHOD

Summary

1. The concept of root locus.
2. Construction of root loci.

The transfer functions of feedback control systems have been studied in Chap. 13. It has been shown that the closed-loop system function \( H(s) \), that is, the output-input transfer function \( R(s)/E(s) \), is given by

\[
H(s) = \frac{KG(s)}{1 + KG(s)} \quad (14-1)
\]

where \( KG(s) \) is the open-loop or output-error transfer function. It was stated that \( KG(s) \) is the usual tool for the design engineer, whereas \( H(s) \) characterizes the behavior of the system. Equation (13-1) was then expressed in harmonic approach terms; that is, the closed-loop transfer locus \( H(j\omega) \) was derived from the open-loop transfer locus \( KG(j\omega) \). The purpose of the present chapter is to express Eq. (14-1) in terms of its pole-zero configuration, that is, essentially to show how the poles of \( H(s) \) can be obtained from the pole-zero configuration of \( KG(s) \).

14.1. THE CONCEPT OF ROOT LOCUS

14.1.1. Introductory Example. Consider the elementary second-order servo system of Sec. 13.1.2. The error \( \varepsilon \) results in a torque \( \Gamma \) which is applied to an output shaft characterized by inertia and viscous friction:

\[
C\varepsilon = J \frac{d^2\Gamma}{dt^2} + f \frac{d\Gamma}{dt}
\]

The open-loop transfer function is

\[
\frac{C}{Js^2 + fS} = \frac{K}{s(1 + Ts)} = \frac{k}{s + 1/T}
\]

where the time constant \( T \) is \( J/f \). The coefficient \( K \) is the open-loop velocity gain. The coefficient \( k \) has no physical significance; it is sometimes called the gain (Sec. 7.1.5) and is equal to \( K/T \).

The closed-loop system function is

\[
H(s) = \frac{kG(s)}{1 + kG(s)} = \frac{k}{s^2 + s/T + k} \quad (14-2)
\]
Its numerator has no zero and its denominator is the characteristic equation of the system,

\[ s^2 + \frac{s}{T} + k = 0 \quad (14-3) \]

The behavior of the servo system is determined by the location of the poles of \( H(s) \), which are the roots of the characteristic equation (14-3). These roots can be computed by solving the corresponding second-degree equation. If \( k \) is smaller than \( 1/4T^2 \) they are real:

\[
\begin{align*}
s_1 &= -\frac{1}{2T} - \frac{1}{2} \left( \frac{1}{T^2} - 4k \right)^{1/2} \\
s_2 &= -\frac{1}{2T} + \frac{1}{2} \left( \frac{1}{T^2} - 4k \right)^{1/2}
\end{align*}
\]

If \( k \) is greater than \( 1/4T^2 \), they are complex conjugates:

\[
\begin{align*}
s_1 &= -\frac{1}{2T} - j\frac{1}{2} \left( 4k - \frac{1}{T^2} \right)^{1/2} \\
s_2 &= -\frac{1}{2T} + j\frac{1}{2} \left( 4k - \frac{1}{T^2} \right)^{1/2}
\end{align*}
\]

This situation can be visualized by solving Eq. (14-3) graphically, real roots \( s = \alpha \) corresponding to intersections of the parabola

\[ y = \alpha^2 + \frac{\alpha}{T} \]

with the horizontal line \( y = -k \) (with \( k > 0 \)). The parabola intersects the \( \alpha \) axis at the points \( \alpha = 0 \) and \( \alpha = -1/T \) and is a minimum with an ordinate of \( -1/4T^2 \) for \( \alpha = -1/2T \).

It is seen in Fig. 14-1 that:

a. If \( k = 0 \), the roots are \( \alpha = 0 \) and \( \alpha = -1/T \).

b. If \( k \) is small, there are two intersections, whence two real roots, which merge into one for the critical value \( k_0 = 1/4T^2 \).

c. If \( k \) is greater than the critical value, the equation has no real root.

It is possible to show the location in the complex plane of the roots of Eq. (14-3) for different values of \( k \).

a. When \( k \) is zero, \( s_2 = -1/T \) and \( s_1 = 0 \). The roots of the characteristic equation are the poles of \( kG(s) \).

b. When \( k \) is small, \( s_1 \) and \( s_2 \) are real. They move along the real axis toward the \( -1/2T \) point while remaining symmetric with respect to that point.

c. For \( k = 1/4T^2 \), the roots \( s_1 \) and \( s_2 \) merge into a double root located at the \( -1/2T \) point.

d. When \( k \) is greater than \( 1/4T^2 \), the roots become conjugate complex, their real part always remaining \( -1/2T \). The root \( s_1 \) moves downward and the root \( s_2 \) upward on the vertical \( -1/2T \) line.

Thus, the locus of the roots of the characteristic equation consists of the segment \((0, -1/T)\) of the real axis and the vertical \( -1/2T \) line. This locus (Fig. 14-2), graduated in \( k \), is called the root locus of the system. Knowing the root locus, it is possible to draw conclusions concerning the behavior of the system as \( k \) varies.

14.1.2. The Root Locus. The above procedure can be generalized. When \( k \) varies, the roots of the characteristic equation, that is, the poles of
the closed-loop system function, describe a locus in the \( s \) plane (Laplace domain). This locus, \textit{graduated in} \( k \), is called the \textit{root locus} of the system. The root-locus technique was conceived and developed by W. R. Evans.\(^1\) It is seen from the above that the knowledge of the root locus of a servo system makes a complete study of the system behavior possible.

14.2. CONSTRUCTION OF ROOT LOCI

14.2.1. General. If \( s \) is a root of the characteristic equation (14-3), one has

\[
G(s) = -\frac{1}{k} \quad (14-4)
\]

which is equivalent to the two equations

\[
|G(s)| = \frac{1}{k} \quad (14-5)
\]

\[
\arg G(s) = \pi + 2\lambda \pi \quad (14-6)
\]

where \( \lambda \) is an integer.

If now \( G(s) \) is written in factored form,

\[
G(s) = \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots}
\]

then Eqs. (14-5) and (14-6) become (Sec. 9.3.1)

\[
\frac{|s - z_1| \cdot |s - z_2| \cdots}{|s - p_1| \cdot |s - p_2| \cdots} = \frac{1}{k} \quad (14-7)
\]

\[
\arg (s - z_1) + \cdots + \arg (s - p_1) - \cdots = \pi + 2\lambda \pi \quad (14-8)
\]

which are called the \textit{magnitude condition} and the \textit{angle condition}, respectively.

The root locus is defined from the \textit{angle condition} as the locus of the points \( M \) that satisfy the condition

\[
\arg Z_1M + \cdots - \arg P_1M - \arg P_2M - \cdots = \pi + 2\lambda \pi \quad (14-9)
\]

The magnitude condition then enables one to obtain \( k \) at each point \( M \) of the locus, that is, to effect the graduation of the locus. Note that the angle condition is a particular case of

\[
\arg Z_1M + \cdots - \arg P_1M - \arg P_2M - \cdots = 0 \quad (14-10)
\]

which defines the streamlines corresponding to the function $kG(s)$ in the hydraulic analogy (Sec. 9.3.7). The root locus is that particular streamline which corresponds to

$$
\Phi = \pi + 2\lambda \pi
$$

The construction of root loci is greatly facilitated by the use of some practical rules, which will be outlined.¹

14.2.2. Practical Rules for the Construction of the Root Locus.

General. If the characteristic equation is of the $n$th degree, it has $n$ roots, which may be real or complex. Therefore, the root locus is comprised of $n$ branches, each of which corresponds to one root of the characteristic equation. Since real roots are represented by points lying on the real axis, the branches which correspond to real roots will consist of portions of the real axis. The branches that correspond to complex roots may be more or less complicated. Since the coefficients of $1 + kG(s)$ are real, the complex roots are grouped in pairs; therefore those branches of the root locus that correspond to complex roots are symmetric with respect to the real axis. Altogether, the over-all root locus is thus symmetric with respect to the real axis.

Rule 1. Starting Points ($k = 0$). Let

$$
kG(s) = \frac{kP(s)}{Q(s)} \quad (14-11)
$$

where $P(s)$ and $Q(s)$ are polynomials in $s$. The characteristic equation for this case is

$$
kP(s) + Q(s) = 0 \quad (14-12)
$$

and its degree is that of the denominator $Q(s)$.

For $k = 0$ the roots of the characteristic equation are the roots of $Q(s)$, that is, the poles of $kG$. Therefore, the root locus starts from the poles of the open-loop transfer function, each branch of the root locus starting from one pole.

Rule 2. Terminal Points ($k \to \infty$). As $k$ approaches infinity, Eq. (14-12) becomes $P(s) = 0$. Thus, $m$ roots of the characteristic equation approach the $m$ roots of $P(s) = 0$.

¹ The following sections are restricted to the case in which $G(s)$ is a polynomial fraction. If $G(s)$ incorporates a lag factor,

$$
G(s) = \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots} e^{-Ts}
$$

the magnitude condition remains unchanged and the angle condition becomes

$$
\chi(s - z_1) + \cdots - \chi(s - p_1) - \cdots - T\omega = \pi + 2\lambda \pi
$$

(where $\omega = \text{Im} \ s$). For the plotting of the root locus in this case, see Y. Chu, “Feedback Control Systems with Dead Time Lag or Distributed Lag by Root-locus Method,” Trans. AIEE, 71(II): 291–298 (1952).
At the same time, the other \( n - m \) roots approach infinity. Therefore, the root locus has \( n - m \) asymptotic directions \( \Delta \). These are given by the angle condition (14-9) applied when \( M \) is at infinite distance. It follows that
\[
(Ox, O\Delta) + \cdots - (Ox, O\Delta) - (Ox, O\Delta) - \cdots = +\pi + 2\mu\pi
\]
where \( \mu \) is integer. That is,
\[
m(Ox, O\Delta) - n(Ox, O\Delta) = +\pi + 2\mu\pi
\]
\[
(Ox, O\Delta) = \frac{-\pi - 2\mu\pi}{n - m}
\]
or, replacing \( \mu \) by \( (-1 - \lambda) \),
\[
(Ox, O\Delta) = \frac{\pi}{n - m} (1 + 2\lambda)
\]
where \( \lambda \) is an integer. These remarks are illustrated by the configurations shown in Fig. 14-3, where the asymptotic directions are marked by arrows.

![Fig. 14-3. Asymptotic directions for root locus as \( k \) becomes infinite.](image)

Case a. \( n - m = 1 \)  
Case b. \( n - m = 2 \)  
Case c. \( n - m = 3 \)  
Case d. \( n - m = 4 \)

Summarizing, the \( n \) terminal points for infinite \( k \) are:
1. The \( m \) zeros of the open-loop transfer function.
2. The \( n - m \) asymptotic directions given by
\[
(Ox, O\Delta) = \frac{-\pi}{n - m} (1 + 2\lambda)
\]

**Rule 3. Branches on the Real Axis.** The root locus comprises those portions of the real axis that lie between poles and zeros of \( kG(s) \), starting from the pole or zero that lies farthest to the right, as shown in Fig. 14-4.

---

\(^1\) This is a general result of the theory of algebraic equations. As \( k \) approaches infinity, it is seen by writing the characteristic equation in the form
\[
P(s) + \frac{1}{k} Q(s) = 0
\]
that the coefficients of the highest powers of \( s \), that is, \( n, n - 1, \ldots \), down to \( m + 1 \), approach zero. As a result, \( n - m \) roots approach infinity. The other \( m \) roots become the roots of \( Q(s) = 0 \).
This results from the fact that, when applying the angle condition to a point \( M \) lying on the real axis, complex zeros or poles and also real zeros or poles that lie farther to the left need not be considered, since the terms they introduce into the angle condition are 0 or \( 2\pi \). Hence, the angle condition becomes

\[
\pi Z_R - \pi P_R = \pi(1 + 2\lambda)
\]

where \( Z_R \) and \( P_R \) are the numbers of the zeros and poles of \( kG(s) \) that lie farther to the right than \( M \). As a result, \( Z_R + P_R \) must be odd, which proves the rule stated.

Note. If \( k \) were negative [which can happen if \( G(s) \) is unstable or is nonminimum phase—see the end of Sec. 7.1.5] rules 2 and 3 should be modified as follows: The asymptotic directions are symmetrical with respect to the imaginary axis of those obtained in the \( k > 0 \) case. The portions of the root locus on the real axis are obtained by starting from the pole or zero that lies farthest to the left.

![Fig. 14-4. Branches of root locus on the real axis.](image)

![Fig. 14-5. Breakaway between poles.](image)

![Fig. 14-6. Breakaway between zeros.](image)

**Rule 4. Intersection with the Imaginary Axis.** Intersections of the root locus with the imaginary axis correspond to values of \( k \) for which the characteristic equation has pure imaginary conjugate roots. Such values of \( k \) are obtained by applying Routh's criterion to the characteristic equation and choosing \( k \) in order that all the coefficients of a row be zero. As was seen in Sec. 9.2.4, the purely imaginary roots of the characteristic equation—that is, the intersections of the root locus with the imaginary axis—are the zeros of the auxiliary polynomial that one introduces when applying Routh's criterion through to the last row.

**14.2.3. Other Practical Rules.** The above four rules enable one to draw a sketch of the root locus for many simple systems. The following rules make a more precise plot possible and enable one to draw quantitative results.

**Rule 5. Breakaway Points from the Real Axis.** Such points correspond to critical values of \( k \), i.e., to values at which two real roots merge before they separate at right angles and constitute an oscillatory mode (Figs. 14-5 and 14-6). The easiest method for determining the breakaway points consists of plotting \( P(\alpha)/Q(\alpha) \) as a function of the real variable \( \alpha \). The transition from real to imaginary intersections by the horizontal \(-1/k\) line occurs when \( P(\alpha)/Q(\alpha) \) is a minimum. (See the introductory example in Sec. 14.1.1. Also, see the note after Example 2 in Sec. 14.2.5.) Therefore, the abscissas of the breakaway points are roots of
\[
\frac{d}{d\alpha} \left( \frac{P}{Q} \right) = 0
\]
that is, of
\[
\frac{dP/d\alpha}{P} = \frac{dQ/d\alpha}{Q}
\] (14-13)

This equation can be readily written from the expressions of \(P(s)\) and \(Q(s)\). It can also be expressed in pole and zero terms as follows: Note that
\[
\frac{dP/d\alpha}{P} = \frac{d\ln P}{d\alpha} \quad \frac{dQ/d\alpha}{Q} = \frac{d\ln Q}{d\alpha}
\]
where \(\ln\) is the natural logarithm. If now \(P\) and \(Q\) are written in factored form
\[
\ln P(s) = \ln (s - z_1) + \ln (s - z_2) + \cdots + \ln k
\]
\[
\ln Q(s) = \ln (s - p_1) + \ln (s - p_2) + \cdots
\]
Eq. (14-13) becomes
\[
\frac{1}{\alpha - z_1} + \frac{1}{\alpha - z_2} + \cdots = \frac{1}{\alpha - p_1} + \frac{1}{\alpha - p_2} + \cdots
\] (14-14)
which is the angle condition applied to a point lying close to the real axis with the abscissa \(\alpha\).

Equation (14-14) is of the \(nm - 1\) degree. The roots, lying on the segments of the real axis on which breakaway is expected to occur, can be found by trial-and-error technique. The corresponding value of \(k\) is immediately obtained from the value of the minimums of \(P(\alpha)/Q(\alpha)\).

If \(P(s)\) is unity, the Eq. (14-14) reduces to \(dQ/d\alpha = 0\).

**Rule 6. Tangent at Complex Starting or Terminal Point.** The tangent to the root locus at a complex pole or zero of \(kG(s)\) is given by the angle condition applied to a point close to the pole or zero.

For the example shown in Fig. 14-7, the angle condition gives
\[
\theta_4 - (\theta_1 + \theta_2 + \theta_3) = \pi + 2\lambda \pi,
\]
whence
\[
\eta = \theta_4 - \theta_3 + \pi/2
\]

**Rule 7. Location of the Asymptotes.** In most cases the asymptotes intersect the real axis at the point whose abscissa is
\[
\alpha = \frac{(p_1 + p_2 + \cdots) - (z_1 + z_2 + \cdots)}{m - n}
\]

---

This point is the center of gravity of the poles and zeros of \( kG(s) \), each pole being considered as having a unit mass and each zero a unit "negative mass." This result can be proved by letting \( s = \alpha + s' \) in the angle condition and equating to zero the term in \( s'(d-n-1) \).

14.2.4. Graphical Determination of Any Point of the Root Locus.

1. Determining Any Point. Once the root locus has been sketched by using the seven rules which have been outlined, it may be desirable to improve its accuracy by graphically determining a few points. This is performed by applying the angle relation

\[
\chi kG = 180^\circ + n360^\circ
\]

which defines the root locus (Sec. 14.1). At a point \( s_i \) this equation is equivalent to:

\[ \chi \text{ vectors from zeros to } s_i - \chi \text{ vectors from poles to } s_i = 180^\circ + n360^\circ \]

which enables one to determine the exact position of the point \( s_i \) by trial and error.

2. Determining \( k \). The relation \( kG(s) = -1 \) can be written in factored form

\[
k \frac{(s - z_1)(s - z_2) \cdots (s - z_m)}{(s - p_1)(s - p_2) \cdots (s - p_n)} = -1
\]

The coefficient \( k \) at the point \( s_i \) is then given by

\[
k = \frac{1}{|G(s)|} = \frac{|s_i - p_1|, |s_i - p_2|, \cdots, |s_i - p_n|}{|s_i - z_1|, |s_i - z_2|, \cdots, |s_i - p_m|}
\]

where \( |s_i - p_k| \) represents the length of the vector from the pole \( p_k \) to the point \( s_i \). Thus \( k \) is obtained by multiplying the length of the vectors from the poles to the point \( s_i \) and dividing the result by the product of the length of the vectors from the zeros to the point \( s_i \) (Sec. 9.3.1). The gain \( K \) is given by [see Eq. (7-4)]

\[
K = k \begin{vmatrix} \frac{z_1z_2 \cdots z_m}{p_1p_2 \cdots p_{n-1}} \end{vmatrix}
\]

3. Example. Given

\[
KG(s) = \frac{k(s - z_1)(s - z_2)}{s(s - p_1)(s - p_2)(s - p_n)}
\]

Figure 14-8 shows that the point \( s_i \) is determined by the condition

\[
\alpha + \beta - \gamma - \delta - \phi - \theta = 180^\circ + n360^\circ
\]

Once the point \( s_i \) is determined, \( k \) is easily obtained by measuring the lengths of the vectors \( A, B, C, D, E, F \), shown in the figure

\[
k = \frac{CDEF}{AB}
\]

1 The operations of adding angles and multiplying vectors can be facilitated by the use of protractor-like devices, the best known of which is the spirule (Evans, op. cit., sec. 7.7 and appendix C).
14.2.5. Examples. 1. Second-order Servo System. The root locus for the elementary second-order servo

\[ kG(s) = \frac{k}{s(s + 1/T)} \]

has been drawn in Fig. 14-2. For second-order servos without integration

\[ kG(s) = \frac{k}{s^2 + as + b} \]

the root loci are sketched in Figs. 14-9 and 14-10.

2. Second-order-plus-lag Servo System. This example will be dealt with in detail because it illustrates many features common to practically all servo systems. Let

\[ kG = \frac{k}{s(s + 1/T')(s + 1/T'')} \]

be the open-loop transfer function. Applying the practical rules given above yields the following results. Figure 14-11a and b is drawn for \( T = 1, \ T' = 0.25 \).

![Fig. 14-8. Construction of any point.](image)

![Fig. 14-9. Root locus of second-order servo system without integration when poles are real.](image)

![Fig. 14-10. Root locus of second-order servo system without integration when poles are complex.](image)
General. The root locus consists of three branches, since the denominator of \( kG \) is of the third degree.

**RULE 1. STARTING POINTS.** The root locus starts from the poles of \( kG(s) \):

\[
s_1 = 0 \quad s_2 = -\frac{1}{T} \quad s_2 = -\frac{1}{T'}
\]

**RULE 2. TERMINAL POINTS.** The numerator of \( kG(s) \) has no zero. Therefore, the three roots of the characteristic equation approach infinity as \( k \) becomes infinite. In fact, since \( n - m = 3 - 0 = 3 \), there are three asymptotic directions at \( \pi/3 \), \( 3\pi/3 \), and \( 5\pi/3 \).

![Fig. 14-11](image)

Fig. 14-11. Step-by-step construction of root locus for \( kG = 1/(s+1)(s+4) \). (a) Application of rules 1, 2, and 3. (b) Application of rules 4, 5, and 7.

**RULE 3. BRANCHES ON THE REAL AXIS.** The portion of the real axis between \( s = 0 \) and \( s = -1/T \) and the portion between \(-1/T'\) and minus infinity are parts of the root locus (it is assumed that \( T > T' \)).

**RULE 4. INTERSECTION WITH THE IMAGINARY AXIS.** The characteristic equation is

\[
s^3 + \left(\frac{1}{T} + \frac{1}{T'}\right)s^2 + \frac{s}{TT'} + k = 0
\]

The condition for sustained oscillation is obtained by equating to zero the Routh determinant (Sec. 9.2.4)

\[
\left(\frac{1}{T} + \frac{1}{T'}\right)\frac{1}{TT'} - k = 0
\]

which gives \( k = 20 \). The values of \( s \) are the roots of the auxiliary Routh polynomial

\[
\left(\frac{1}{T} + \frac{1}{T'}\right)s^2 + k = 0
\]

that is,

\[
s = \pm j\omega \quad \omega = \left(\frac{k}{1/T + 1/T'}\right)^{1/4} = 2
\]
RULE 5. BREAKAWAY FROM THE REAL AXIS. The abscissa $\alpha$ of the breakaway point located between the 0 and $-1$ poles is a root of

$$\frac{dQ}{d\alpha} = 3\alpha^2 + 2\left(\frac{1}{T} + \frac{1}{T'}\right)\alpha + \frac{1}{TT'} = 0$$

whence $\alpha = -0.47$.

RULE 6. LOCATION OF THE ASYMPOTOTES. They cross at the center of gravity of the poles,

$$s = -\frac{1}{3}\left(\frac{1}{T} + \frac{1}{T'}\right) = -1.67$$

It is now possible to draw the root locus, as shown in Fig. 14-12.

![Root locus diagram](image)

**Fig. 14-12.** Root locus for $kG = 1/s(s + 1)(s + 4)$.

Note. A mathematical way of visualizing the manner in which the roots of the characteristic equation trace out the root locus can be described as follows. Real roots of the characteristic equation

$$KG = k\frac{P}{Q} = -1$$

which in the present case ($P = 1$) is

$$Q(s) = s\left(s + \frac{1}{T}\right)\left(s + \frac{1}{T'}\right) = -k \quad k > 0$$

can be obtained graphically by plotting the cubic

$$\frac{1}{G(\alpha)} = Q(\alpha) = \alpha\left(\alpha + \frac{1}{T}\right)\left(\alpha + \frac{1}{T'}\right)$$
and finding its intersection with the horizontal \(-k\) line. A sketch of the situation is given in Fig. 14-13. \(Q(\alpha)\) is equal to \(+\infty\) when \(\alpha\) is \(+\infty\), and it is equal to \(-\infty\) when \(\alpha\) is \(-\infty\). It intersects the \(\alpha\) axis at 0, \(-1/T\), and \(-1/T'\).

a. When \(k = 0\), the roots of the characteristic equation are 0, \(-1/T\), and \(-1/T'\). They are the poles of \(kG(s)\).

b. When \(k\) is small, three intersections take place; that is, the characteristic equation has three real roots. It is seen from the figure that the root which lies farthest to the left moves to the left as \(k\) increases. The other two roots move toward each other.

c. For a critical value \(k\) equal to \(k_0\), the \(-k\) line becomes tangent to \(G(\alpha)\). The two roots that lie on the right then merge into one.

\[Fig. 14-13.\]

\[Fig. 14-14.\] Root locus for
\[kG = \frac{k}{s[(s + \alpha)^2 + \beta^2]}\].

\[Fig. 14-15.\] Aircraft with ideal autopilot.

d. When \(k\) becomes greater than \(k_0\), there is only one intersection. That is, the characteristic equation has only one real root, which increases negatively as \(k\) increases. The other two roots are imaginary.

To summarize the results of this discussion:

1. When \(k\) varies from zero to \(+\infty\), the mode of the system corresponding to the intersection that lies farthest to the left is always nonoscillatory. It corresponds to an exponential function of time, which decays faster as \(k\) increases.

2. On the other hand, the other two roots constitute two nonoscillatory modes when \(k\) is small; whereas they constitute an oscillatory mode for higher values of \(k\).

3. Third-order Servo with Integration and Complex Poles. Let

\[kG = \frac{k}{s[(s + \alpha^2) + \beta^2]}\]

be the open-loop transfer function. The root locus is easily found to have the shape shown in Fig. 14-14.
**Fig. 14-16.** Root locus for stable aircraft with ideal autopilot.

**Fig. 14-17.** Root locus for unstable aircraft with ideal autopilot.

**Fig. 14-18.** Autopiloted aircraft.

**Fig. 14-19.** Pole-zero configuration for autopiloted stable aircraft.
4. Aircraft in Longitudinal Motion. The transfer function of a stable airplane in longitudinal motion is of the form

\[ kG(s) = \frac{k(s + a)}{s[(s + \alpha)^2 + \beta^2]} \]

when the velocity is assumed constant (Sec. 7.5.2). A real zero \( s = -\alpha \) has been introduced into the transfer function of Example 3. The root locus for the equation \( 1 + kG(s) = 0 \) is the root locus for the feedback system whose forward path consists of the aircraft alone, within a gain factor. This feedback system is the controlled aircraft if the autopilot is assumed to produce a control-surface deflection proportional to the error in pitch angle (Fig. 14-15). The shape of the root locus is shown in Fig. 14-16. Intersections with the imaginary axis occur if the asymptote lies in the right-hand half plane, that is, if \( a > 2\alpha \).

If the aircraft is unstable, its transfer function will have the form

\[ k \frac{s + a}{s(s + b)(s - c)} \quad a, b, c > 0 \]

The corresponding root locus is sketched in Fig. 14-17. Intersections with the imaginary axis take place if \( a + c < b \).

5. Autopiloted Aircraft in Longitudinal Motion. The following examples are adapted from W. Bollay.¹ First consider a stable aircraft with the transfer function

\[ \Theta(s) = \frac{A}{s(s + \alpha)^2 + \beta^2} = \frac{8.3}{s(s + 0.9)^2 + 1.85^2} \]

The autopilot is assumed to consist of (Fig. 14-18) an ideal sensing device, a compensating network with a transfer function²

\[ \frac{H}{\xi} = 6 \frac{s + 1/6T}{s + 1/T} = 6 \frac{s + 1}{s + 6} \]

and a hydraulic motor which can be approximated as a second-order system

\[ \frac{\Delta}{H} = \frac{163}{(s + 6.5)^2 + 11^2} \]


² Such a network is called a lag network or an undercompensated integral controller (Sec. 18.3).
The over-all open-loop transfer function for the controlled aircraft is

\[
\frac{\Theta}{\xi} = \frac{k}{s} \frac{s + 0.87}{(s + -0.9)^2 + (1.85)^2} \frac{s + 1}{s + 6} \frac{1}{(s + 6.5)^2 + 11^2}
\]

The pole-zero configuration is shown in Fig. 14-19.\(^1\) The root locus, with asymptotes at \(\pm 45^\circ\), is shown in Fig. 14-20.

![Fig. 14-21. Pole-zero configuration for autopiloted unstable aircraft.](image)

If now an unstable aircraft with a transfer function

\[
\frac{\Theta(s)}{\Delta(s)} = \frac{5.6}{s} \frac{s + 0.8}{(s + 2.52)(s - 0.92)}
\]

is assumed to be piloted by the same autopilot

\[
\frac{\Delta(s)}{\xi} = \frac{6}{s + 6} \frac{s + 1}{(s + 6.5)^2 + 11^2}
\]

Fig. 14-22. (a) Root locus for autopiloted unstable aircraft. [After W. Bollay, "Aerodynamic Stability and Automatic Control," J. Aeron. Sci., 18(9):569–623 (1951).] (b) Magnification of portion of (a) lying near the origin.
the new pole-zero configuration is shown in Fig. 14-21 and the new root locus in Fig. 14-22a. (Figure 14-22b shows the portion in the neighborhood of the origin.)

14.3. STABILITY OF FEEDBACK CONTROL SYSTEMS

14.3.1. General. A linear servo system is completely characterized by the zeros and poles of its closed-loop system function

\[ H(s) = \frac{kG(s)}{1 + kG(s)} \]

The zeros of \( H(s) \) are those of \( KG(s) \); the poles are the roots of the characteristic equation. Therefore, it is apparent that the knowledge of the root locus enables one to study the behavior of the system for all values of \( k \).

In particular, a glance at the root locus graduated in \( k \) immediately indicates the stability of the system. As shown in Sec. 9.1, the system is stable if all the poles of its transfer function lie in the left half plane. Therefore, the system will be stable for those values of \( k \) for which the roots corresponding to all the branches of the root locus lie in the left half plane. These considerations will now be applied to some cases frequently encountered in practice.

14.3.2. Case of Regular Servo Systems. In the case of a second-order servo system with an open-loop transfer function

\[ kG(s) = \frac{k}{s(s + 1/T)} \]

it has been seen that the root locus lies entirely in the left half plane. Therefore, such a system is stable, whatever its open-loop gain.

In practice, however, experience shows that all servo systems become unstable when their open-loop gain is increased sufficiently. This indicates that real servo systems are not adequately represented by second-order systems. In fact, if one takes into account time lags and writes a new open-loop transfer function

\[ kG(s) = \frac{k}{s(s + 1/T)(s + 1/T')} \]

it has been found (Sec. 14.2.5, Example 2) that the resulting root locus has its asymptotes at ±60°. Therefore, large values of \( k \) cause the corresponding point to move into the right-hand half plane, and the system becomes unstable.\(^1\) The critical value \( k_c \) of \( k \) which corresponds to an oscillatory system is easily obtained by applying Routh's criterion:

\[ k_c = \frac{T + T'}{T^2T'^2} \]

The system is stable for \( k \) smaller than \( k_c \), and unstable for \( k \) greater than \( k_c \).

\(^1\) This is the phenomenon of hunting. It has been already referred to in Sec. 13.2.3 and 13.3.1 and will be studied in Sec. 16.2.3.
Servo systems which are stable for values of \( k \) smaller than and unstable for values of \( k \) greater than a critical value are said to be regular. This means, essentially, that the influence of the open-loop gain on their stability is the same as for a second-order-plus-lag servo system.

14.3.3. Conditionally Stable Systems. Some servo systems, on the contrary, are stable for certain ranges of values for their open-loop gain; that is, whenever \( k \) satisfies conditions such as

\[
 k'_e < k < k''_e \quad \text{with} \quad k'_e \neq 0
\]

the systems are said to be conditionally stable. One of the most important cases of such systems is that of servo systems which are open-loop unstable, that is, those whose open-loop transfer functions have poles with positive real parts. Examples are piloted unstable airplanes and servo systems with superposed loops when the internal servo is unstable.

For such systems, at least one branch of the root locus starts for \( k = 0 \) in the right-hand half plane. It then may cross the imaginary axis (\( k = k'_e \)) and exhibit a stable portion. If \( k \) is increased still further, the locus will move to the right-hand half plane again.\(^1\)

Example 5 in Sec. 14.2.5 is such a conditionally stable servo system in the case of an unstable aircraft (Fig. 14-22; also see below, Fig. 14-30).

14.3.4. Stability Margins. As shown in Part 1 (Sec. 9.1), the fact that the poles of the transfer function lie in the left half plane merely guarantees mathematical stability. If some poles lie too close to the imaginary axis, the system will have poor performance, the transient response being too slow or insufficiently damped. Therefore, it is often advisable to exclude as regions of operation for the system either (a) or (b):

a. The region lying to the right of a line parallel to the imaginary axis, which is said to prescribe an absolute stability margin (Fig. 9-4).

b. The region lying to the right of two oblique lines starting from the origin, which are said to prescribe a relative stability margin (Fig. 9-5). The root-locus technique lends itself very readily to the manipulation of such conditions. In particular, it is easy to adjust the gain to the highest possible value corresponding to a prescribed damping ratio for the dominant oscillatory mode, which is equivalent to prescribing a maximum \( Q \) on the Hall or Nichols chart. For example, relative stability corresponding to \( \xi = 0.4 \) and \( \psi = 25^\circ \) is obtained for the servo system of Example 2 in Sec. 14.2.5 if \( k < 3.2 \) (that is, \( K < 0.8 \)). (See Fig. 14-12.)

14.3.5. First Approach to Stabilization. An unstable, or insufficiently stable, servo system can be stabilized by inserting a compensating network. In practice, the need for such a network arises as follows: Increasing the open-loop gain of a servo system in order to improve the steady-state performance results in hunting, whence arises the idea of multiplying the open-loop transfer function by a complex factor which will make it possible to increase the open-loop gain to a higher value without causing hunting.

\(^1\) Should this not be the case, it would be due to the fact that lag factors, such as inertia or time lags, have been neglected when deriving the transfer function.
Since the presence of time lags, i.e., factors of the form \(1/(1 + Ts)\), has a destabilizing effect, it is logical to attempt to eliminate their influence by introducing factors of the form \(1 + Ts\). Since

\[
(1 + Ts)\mathcal{L}(s) = \mathcal{L}

\begin{bmatrix}
\mathcal{L}(t) + T \frac{d\mathcal{L}}{dt}
\end{bmatrix}
\]

this results in controlling by means of the time derivative of the error, whence the name derivative control. Factors of the form \(1 + Ts + T''s^2\) can also be introduced, which leads to controlling by means of the first two derivatives of the error.

**Fig. 14-23.** Block diagram of roll-stabilized aircraft.

In terms of the pole-zero configuration, the introduction of such factors amounts to adding zeros to the open-loop transfer function. In terms of frequency response, it is equivalent to adding positive quantities to the phase of the open-loop transfer function. In fact, such factors have positive phase angles over the frequency range of interest; they are therefore called phase-lead elements.

In practice, the transfer function of the compensating network has a denominator of degree higher than the numerator; accordingly, derivative control is practically realized by

\[
\frac{1 + T_1s}{1 + T_2s} \quad \text{with} \quad T_1 > T_2
\]

or by

\[
\frac{1 + T_1s + T_2^2s^2}{1 + T_1's + T_2^2s^2} \quad \text{with} \quad T_1' > T_2' > T_2
\]

**Fig. 14-24.** Root locus for \(kG = k/s^2\).

The technique of choosing adequate compensating networks to stabilize a given servo system will be studied later in detail. In the present chapter a typical example will be presented only in a qualitative manner in order to demonstrate the use of the root-locus method in visualizing the stabilizing effect of lead elements.

**14.3.6. Example. Roll Stabilization of an Airplane.** The block diagram of a roll-stabilized airplane is shown in Fig. 14-23, where \(\varphi\) is the angle of roll and \(\delta\) the aileron deflection. The latter is slaved to the controller output \(\eta\), so that the "servomotor stage" box actually implies an internal closed loop. The dynamics of the aircraft in the absence of dihedral effect is approximately described by the equation

\[
J \frac{d^2\varphi}{dt^2} + f \frac{d\varphi}{dt} = A\delta
\]
Fig. 14-25. Root locus for

\[ kG = \frac{k}{s^2} \frac{1}{s + 1/T} \]

Fig. 14-26. Root locus for

\[ kG = \frac{k}{s^2} \frac{s + 1/aT}{s + 1/T} \]

where \( a > 1 \).

Case 1: \( b < \frac{a^2}{4} \)

![Case 1 diagram](a)

Case 2: \( b > \frac{a^2}{4} \)

![Case 2 diagram](b)

Fig. 14-27. Root locus for

\[ kG = \frac{k}{s^2} \frac{1}{s + 1/bT^2} \]

where \( a > 1 \).
where $A\delta$ is the torque produced by a deflection $\delta$ of the ailerons, $J$ is the moment of inertia of the aircraft around its longitudinal axis, and $f\,d\varphi/dt$ is a damping moment. The latter is often considered negligible as a first approximation, which leads to the transfer function

$$\frac{\Phi}{\Delta}(s) = \frac{A}{J} \frac{1}{s^3}$$

If the servo motor were ideal and if the controller possessed a unit transfer function, the servo system would be oscillatory whatever the gain, since the root locus

![Root locus diagram](image)

**Fig. 14-28. Root locus for**

$$kG = \frac{k}{s^2} \frac{s^2 + (a/bT)s + 1/bT^2}{s + 1/T} \frac{1}{s + 1/T_1} \frac{1}{s + 1/T_2}$$

where $a > 1$. System is regular.

would include the imaginary axis (Fig. 14-24). Actually, lags are present: for example, the servomotor stage has a transfer function of the form

$$\frac{\Delta}{H}(s) = \frac{1}{1 + Ts}$$

The open-loop transfer function $kG(s)$ is

$$\frac{K}{s^2} \frac{1}{1 + Ts} = \frac{k}{s^2} \frac{1}{s + 1/T}$$

and the root-locus plot (Fig. 14-25) shows that the system is unstable for all values of $k$.\(^1\)

If now the first derivative of the error is introduced by the controller

$$\eta = \varepsilon + aT \frac{de}{dt}$$

\(^1\) This is a typical example of a structurally unstable system, as defined in Sec. 9.2.6.
the open-loop transfer function becomes

\[ \frac{K}{s^3} \frac{1 + aTs}{1 + Ts} \frac{k}{s^2} \frac{s + 1/aT}{s + 1/T} \]

where \(a\) is to be determined. It is seen from the root-locus plot (Fig. 14-26) that the system is stable for all values of \(k\) when \(a > 1\), but that the stability margin is always poor when \(a\) is small. Obviously, taking into account additional lags would lead to a system unstable for high values of \(k\), that is, to a regular system.

Case 1: \(b < \frac{a^2}{4}\)

![Diagram](a)

Case 2: \(b > \frac{a^2}{4}\)

![Diagram](b)

Fig. 14-29. Root loci for \(kG = \frac{k}{s^3} \frac{s^2 + (a/bT)s + 1/bT^2}{s + 1/T}\), where \(a < 1\). System is not regular.

If the second derivative is introduced, then \(kG(s)\) becomes

\[ \frac{K}{s^3} \frac{1 + aTs + bT^2s^2}{1 + Ts} = \frac{k}{s^2} \frac{s^2 + (a/bT)s + 1/bT^2}{s + 1/T} \]

Two cases may be defined.

Case 1. \(a > 1\). When \(b\) is small, an additional zero is introduced at a great distance on the negative real axis. The vertical asymptote disappears (Fig. 14-27a). When \(b\) becomes greater, the form of the root locus changes; and for \(b = a^2/4\) the additional zeros become complex (Fig. 14-27b). In all cases it is seen that the introduction of an additional zero has resulted in an increase in the stability margin. It should be noted that additional lags—represented, for example, by two factors \(1/(s + 1/T_1)\) and \(1/(s + 1/T_2)\)—cause instability for high \(k\) (Fig. 14-28), thus making the system regular.
Case 2. \( a < 1 \). When \( b \) is small, the additional zero lies far out on the negative real axis and the vertical asymptote disappears (Fig. 14-29a). As \( b \) becomes greater, the zeros become complex for \( b = a^2/4 \), which changes the shape of the locus (Fig. 14-29b). In all cases it is seen that the system is not regular, since increasing \( k \) always enhances the stability. Taking additional lags into account causes hunting to appear for high \( k \), thus making the system conditionally stable. This is verified by constructing the root locus when two factors \( 1/(s + 1/T_1) \) and \( 1/(s + 1/T_2) \) are added (Fig. 14-30), with \( T_1 \) and \( T_2 \) sufficiently small. In conclusion, the qualitative study just outlined shows that stabilization can be performed by the introduction of lead factors, the introduction of the second derivative of the error—that is, of two zeros in the open-loop transfer function—being necessary to guarantee satisfactory stability margins.

\[ kG = \frac{k \frac{s^2 + (a/b)T}{s^2 + 1/T}}{s + 1/T_1, s + 1/T_2} \]

where \( a < 1 \). System is conditionally stable.

\(^1\) If \( T_1 \) and \( T_2 \) were large, the system would be unstable, whatever the gain.
CHAPTER 15

THE STEADY STATE OF FEEDBACK
CONTROL SYSTEMS

Summary
1. Position error.
2. Generalization.
3. Application to automatic piloting.

It has been shown in the preceding chapters that the response of a feedback control system to a step input involves a steady-state error when there is no integration in the open-loop transfer function. When the open-loop transfer function has integration, however, there is no steady-state error.

It is the purpose of the present chapter to generalize these facts. The steady state ($t \to \infty$) of feedback control systems will be systematically studied for step and ramp inputs in relation to the low-frequency behavior ($s \to 0$) of the open-loop transfer function.

15.1. POSITION ERROR

15.1.1. Definition. Let a stable positional servo system be subjected to a step input. After the transient has died away, the steady state may or may not involve an error. It is said that there is (or is not) a steady-state error for a step position input, or more briefly a position error (Secs. 3.3.2 and 7.2.2). For example, it has been shown that the elementary second-order servo $KG(s) = K/s(1 + Ts)$ does not involve a position error.

It should be noted that the expression position error is extended by usage to feedback control systems which are not positional servomechanisms.

15.1.2. Feedback Control Systems without a Position Error. It has been seen previously by using the harmonic approach and the Hall chart (Sec. 13.2.2) that there is no position error if the open-loop transfer function involves an integration. This result will now be proved by three different methods.

First Method. (For the case of a second-order servo system.) If $KG(s)$ has integration, the differential equation of the system has the form

$$J \frac{d^2r}{dt^2} + f \frac{dr}{ds} = k(e - r) = k\varepsilon$$

where there is no $r$ term on the left-hand side of the equation. For the steady state we have $e = e_0$, and it is fairly obvious that the system is at
rest \((r = \text{const})\). If one had \(dr/dt \neq 0\), \(r\) would become infinite, which would be in contradiction to the given differential equation. Hence

\[
\frac{dr}{dt} = \frac{d^2r}{dt^2} = 0
\]

The differential equation then yields \(e = r\); that is, there is no position error.

If, on the contrary, there were no integration, one would have

\[
J \frac{d^2r}{dt^2} + f \frac{dr}{dt} + k'r = k(e - r)
\]

the steady state would be determined by equating the error torque and the restraining torque,

\[
k'r = k(e - r) \quad \text{whence} \quad \varepsilon = e_0 \frac{k'}{k + k'} \neq 0
\]

**Second Method.** The error \(\varepsilon\) corresponding to the input

\[
e = e_0 u(t) \quad \text{thus} \quad E(s) = \frac{e_0}{s}
\]

is given by

\[
\varepsilon(s) = \frac{1}{1 + KG(s)} E(s)
\]

The final-value theorem

\[
\varepsilon(\infty) = \lim_{s \to 0} s\varepsilon(s)
\]

shows that the condition for \(\varepsilon(\infty)\) to be zero is

\[
1 + KG(0) = \infty
\]

i.e., that there is an integration in the forward path.

**Third Method.** The presence of an integration means that a position error results in an output rate \(dr/dt\). In the steady state for a constant input, the response is constant. It follows that the error is zero; for if it were not, there would be a nonzero \(dr/dt\); that is, the steady state would not have been reached. In other words, the servo system keeps running as long as there is an error, its steady state being reached only when zero-error condition is achieved.

The latter proof assumes that there is no disturbance. If one attempts to investigate the steady-state condition in the presence of disturbances, one should refer to Sec. 15.1.3 and resort to the above first or second approach.

**15.1.3. Introduction of Disturbances.** It is important to note that the position error just studied is the steady-state error for a step input, i.e., for a constant disturbance applied at the input of the system. However, if a disturbance is applied elsewhere in the loop, it can be seen by applying the final-value theorem to the error-disturbance transfer function
Fig. 15-1. (a) Zero steady-state error for constant \( d \). (b) A steady-state error for constant \( d \).

[Sec. 13.1.7, Eq. (13-2)] or by making use of Note 1 of Sec. 13.1.7 that a position error is eliminated only if the open loop has an integration preceding the point where the disturbance is introduced (Fig. 15-1a and b).

15.2. GENERALIZATION

15.2.1. Definition of the Velocity Error. Let a position servo be subjected to a ramp input

\[ e(t) = e_0 tu(t) \]

The steady state may involve an error

\[ \varepsilon(\infty) = e(\infty) - r(\infty) \]

which is termed the steady-state error for a ramp (or velocity) input, or more briefly, the velocity error (Sec. 7.2.3).

15.2.2. Application to the Elementary Second-order Servo System. It can easily be shown that the elementary second-order servo system of Sec. 13.1.2, defined by \( KG(s) = K/s(1 + Ts) \), when subjected to the input \( e(t) = atu(t) \), has a velocity error (Fig. 15-2)

\[ \varepsilon(\infty) = \frac{a}{K_v} \]

which is (a) proportional to the slope of the ramp input and (b) inversely proportional to the open-loop velocity gain \( K_v \).

15.2.3. Application to Feedback Control Systems with One Integration. More generally, if the open-loop transfer function of a feedback system implies integration,

\[ KG = \frac{K_v}{s} \frac{1 + \cdots}{1 + \cdots} \]
there is a velocity error which is proportional to $1/K_v$. This can easily be shown by using any of the three methods used in Sec. 15.1.2 for evaluating the position error.

The open-loop velocity gain $K_v$ is, therefore, the velocity constant $C_v$ of the feedback control system. This result can also be proved by reference to the series expansion of $1 - H(s) = 1/[1 + KG(s)]$ (Sec. 7.2.4).

The results obtained in this section can be summarized in the form of Table 15-1.

**Table 15-1. Position and Velocity Error for Servo Systems with One Integration When Disturbance Is Applied at the Input**

<table>
<thead>
<tr>
<th>Number of integrations . . . . .</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-loop transfer function . .</td>
<td>$K_v \frac{1 + as}{s} \ldots$</td>
</tr>
<tr>
<td>Open-loop transfer locus . . . . . . .</td>
<td>$\frac{KG(j\omega)}{s}$</td>
</tr>
<tr>
<td>Position error . . . . . . . . . .</td>
<td>0</td>
</tr>
<tr>
<td>Velocity error . . . . . . . . .</td>
<td>$\frac{1}{K_v}$</td>
</tr>
</tbody>
</table>

**15.2.4. Application to Feedback Control Systems with Two Integrations.** By similar reasoning, the reader may easily see that feedback control systems with two integrations in the forward path have no velocity error. Rather, they have an acceleration error which is inversely proportional to the open-loop acceleration gain $K_a$. For that reason $K_a$ is the acceleration constant $C_a$ of the system.

If a ramp disturbance is applied at a point located in the forward path, the steady-state error is eliminated only if the two integrations precede the point at which the disturbance is introduced.

**15.2.5. Acceleration Error.** In the same way, for a step acceleration input

$$e(t) = \frac{1}{2}e_0 t^2 u(t)$$

one may consider an acceleration error, i.e., the steady-state error for a step-acceleration input,

$$\varepsilon(\infty) = e(\infty) - r(\infty)$$

On the other hand, there may exist three integrations in the forward path of a feedback control system, the presence of a third integration in the open-loop transfer function eliminating the acceleration error.

All the results obtained in this chapter are summarized in Table 15-2.
It should be pointed out that it is quite exceptional to have more than two integrations in the open-loop transfer function of a feedback control system. One reason is, generally speaking, that the greater the number of integrations, the less stable the system will be, and hence the harder to stabilize.

| Table 15-2. Steady-state Error for Servo Systems with 0, 1, 2, and 3 Integrations when Disturbance is Applied at the Input |
|---|---|---|---|
| Open-loop transfer locus $KG(j\omega)$ | $K$ | $K$ | $K$ |
| Position error . . . | $\alpha \frac{1}{K + 1}$ | 0 | 0 | 0 |
| Velocity error . . . | $\infty$ | $\alpha \frac{1}{K_v}$ | 0 | 0 |
| Acceleration error | $\infty$ | $\infty$ | $\alpha \frac{1}{K_a}$ | 0 |
| Number of integrations . . . . . | 0 | 1 | 2 | 3 |
| Open-loop transfer function $KG(s)$ | $\frac{1 + \cdots}{1 + \cdots}$ | $K_v \frac{1 + \cdots}{s + \cdots}$ | $K_a \frac{1 + \cdots}{s^2 + \cdots}$ | $K' \frac{1 + \cdots}{s^2 + \cdots}$ |

In conclusion, it is to be noted that important information concerning the steady state of a feedback control system is provided by a simple observation of the low-frequency region of the open-loop transfer locus.

15.2.6. Consequences. 1. If a feedback control system has a steady-state error (say, a position error), it can be eliminated by introducing an integration into the open-loop transfer function, that is, by multiplying it by a factor which has a zero pole. This is the integral-compensation technique, which will be developed in Chap. 18.

2. If it is impossible or undesirable to introduce integration, one may introduce a factor so that the open-loop gain adjustment is increased $N$ times: as a result, the position error will be divided by $N$. This is the generalized or undercompensated integral control, which will also be studied in Chap. 18.
15.3. APPLICATION TO AUTOMATIC PILOTING

15.3.1. Definitions. The preceding considerations concerning the effect of integrations on the performance of a servo system can be applied to the problem of automatic piloting. It was shown in Sec. 1.3.2 that an autopiloted aircraft is a feedback control system. It is now possible to proceed to more detailed analysis by differentiating among:

1. Position piloting with no integration involved in the autopilot
2. Velocity piloting with one integration in the autopilot
3. Acceleration or torque piloting with two integrations in the autopilot

1. Position Piloting. Ideally, position piloting consists of slaving the position \( \delta \) of the control surface to the error signal:

\[
\delta = Ke
\]  
(15-1)

Thus the block diagram of the controlled aircraft involves two superposed loops. Actually it is necessary to take into account:

a. The transfer function of the controls of the airplane, often represented by a second-order system, although nonlinearities are not always negligible

b. The addition of a compensating network for stabilization purposes, which generally adds to the error signal a term proportional to the time derivative of the error (derivative or phase-lead compensation; Sec. 18.2)

Hence the equation

\[
A \frac{d^2\delta}{dt^2} + B \frac{d\delta}{dt} + \delta = Ke + K' \frac{de}{dt}
\]  
(15-2)

2. Velocity Piloting. A constant error results in a constant rate of deflection for the control surface:

\[
\frac{d\delta}{dt} = Ke
\]  
(15-3)

Actually it is necessary to take into account:

a. The inertia of the control surface

b. A compensating network which provides a signal proportional to the first two derivatives of the error

Hence:

\[
A \frac{d^2\delta}{dt^2} + \frac{d\delta}{dt} = Ke + K' \frac{de}{dt} + K'' \frac{d^2e}{dt^2}
\]  
(15-4)

Open-loop control usually gives this relation, and as a consequence the block diagram of the aircraft involves only one loop.

3. Acceleration or Torque Piloting. In such autopilots a constant error would result in a constant torque applied to the control surface: \( \frac{d^2\delta}{dt^2} = Ke \).

No autopilot is based on this principle.

15.3.2. Transfer Functions. If longitudinal motion is concerned and if \( F(s) \) is the transfer function of the airplane when the input is the control-surface deflection and the output is the pitch angle, then:

a. For position piloting (Fig. 15-3),

\[
\Theta \frac{\Theta}{\delta} (s) = KF(s) \quad \text{ideal}
\]

\[
\Theta \frac{\Theta}{\delta} (s) = K + K' \frac{K'\delta}{As^2 + Bs + 1} F(s)
\]
Most autopilots today are of the position type. The position of the control surface is controlled by means of a servomechanism; as a consequence, the block diagram of the piloted aircraft involves two superposed loops (Fig. 1-47).

b. For velocity piloting (Fig. 15-4),

\[
\frac{\Theta}{\varepsilon}(s) = \frac{K}{s} F(s) \quad \text{ideal}
\]

\[
\frac{\Theta}{\varepsilon}(s) = \frac{K + K's + K''s^2}{s(As + 1)} F(s)
\]

In most cases\(^1\) the transfer function \(F(s)\) does not involve integration.

Therefore, a position-piloted aircraft is a servo system with an open-loop static gain: it has no integration. A velocity-piloted aircraft is a servo system with an open-loop velocity gain: it involves integration.

**Fig. 15-3. Position piloting.**

**Fig. 15-4. Velocity piloting.**

**15.3.3. Comparison of the Two Methods.** A general comparison between position and velocity piloting is a vast and complex subject about which general conclusions should be drawn with great care. However, the following arguments can be presented as applications of the theory of feedback control systems:

1. Velocity piloting demands less power than position piloting with similar aerodynamic load torques. This can be shown roughly by a reference to Eqs. (15-1) and (15-3). If a disturbance results in a sudden change in \(\varepsilon\), Eq. (15-1) theoretically requires that infinite power be involved, while Eq. (15-3) requires only finite values. Usual orders of magnitude are 15 to 50 watts for position piloting and 8 to 30 watts for velocity piloting.

2. In the case of position piloting, the inertia of the control surface may cause unwanted oscillations, implied by the presence of the second-degree denominator in the \(\Theta/\varepsilon\) transfer function, i.e., by the internal loop, if the corresponding resonant frequency, usually 3 to 10 cps, is not far enough from the airplane resonant frequency. If velocity piloting is implied, no such oscillations can arise, since the control-surface inertia does not produce a quadratic factor in the denominator of the transfer function.

3. A velocity-piloted aircraft involves no position error because of the integration in the open loop. However, it should be noted that the aerodynamic load on the

\(^1\) For the case of longitudinal motion, see Sec. 7.5.2. For the case of roll stabilization, \(F(s)\) incorporates integration unless there is a marked dihedral effect. Note that an integration in \(F(s)\) does not eliminate the effect of a constant disturbance under conditions shown in Fig. 15-1b.
control surface is in general not negligible and results in a static gain (see Sec. 15.1.2, first method, the case in which \( k'\tau \) is introduced). Therefore the low-frequency behavior of the system is actually that of a position pilot (Sec. 34.3.2, par. 3b).

4. The stabilization of a velocity-piloted aircraft is more difficult than that of a position-piloted aircraft. This is due to the 90° phase lag introduced by the integration. In order that this lag be compensated, it is necessary to introduce compensating networks that differentiate the error twice. This is possible only if an extremely low noise level is achieved in the error-sensing device, since each differentiation tends to emphasize the noise rather than the useful signal (Fig. 29-11). In order to avoid double differentiation, most velocity autopilots use rate gyros (i.e., angular-velocity sensing devices); as a result, the position error is no longer canceled.
CHAPTER 16

GRAPHICAL STABILITY CRITERIA FOR
FEEDBACK CONTROL SYSTEMS

Summary
1. Nyquist's criterion.
2. Various applications.
3. The left-hand criterion.
4. Practical stability. Phase and gain margins.

If $K_G(s)$ is the open-loop transfer function of a feedback control system (Fig. 13-2), the closed-loop system function has been shown to be

$$H(s) = \frac{K_G(s)}{1 + K_G(s)}$$  \hspace{1cm} (16-1)

According to the general result obtained in Chap. 9, the system is stable if all the zeros of the characteristic equation

$$1 + K_G(s) = 0$$  \hspace{1cm} (16-2)

have negative real parts.

If the mathematical expression for $K_G(s)$ is known, the application of Routh's criterion to Eq. (16-2) enables one to determine whether or not the system is stable and, if unstable, how many unstable roots the characteristic equation incorporates. Furthermore, the root-locus technique makes it possible to evaluate how far the roots of the characteristic equation are from the imaginary axis, and hence to study the degree of stability of the system.

However, it often happens that some components of the servo system are known from physical experimental data. In this case, the mathematical expression for $K_G(s)$ is not known, so it is not possible to resort to Routh's criterion or to the root-locus technique. This accounts for the interest of graphical stability criteria, which enable one to study the stability and the degree of stability of a feedback control system from the plot of its open-loop transfer locus.

16.1. NYQUIST'S CRITERION

16.1.1. General. The importance of the $K_G(j\omega)$ locus position with respect to the critical point $A(-1)$ in the stability of a feedback control

---

system has been shown in Chap. 13. Nyquist's criterion emphasizes this fact; it yields a relation between the stability of the system and the way in which the $KG(j\omega)$ locus encloses point $A$. Nyquist's criterion states that a closed-loop system is stable if its open-loop transfer locus encloses the $-1$ point a number of times equal to the number of the unstable poles of $KG(s)$. The number of turns should be counted counterclockwise.

Mathematically speaking, the criterion immediately results from the residue theorem. Practically, some pitfalls are involved in its application, and one must have a thorough understanding of the notions underlying the criterion. That is why we will begin with the mathematical theory.

16.1.2. Theory of the Criterion. 1. Fundamental Lemma. Let $F(s)$ be a function of the complex variable $s$: the magnitude and angle of $F(s)$ are functions of $s$. If point $s$ describes a closed curve $(C)$ in the $s$ plane (Fig. 16-1), point $F(s)$ describes a locus $(\Gamma)$ which is more or less complicated (Fig. 16-2) and the $(C)$ and $(\Gamma)$ loci correspond point for point. This being so, it can be proved that there is a relation between the number of revolutions of the $(\Gamma)$ locus around the origin and the number of poles and zeros of $F(s)$ which are inside the closed curve $(C)$. More precisely, when the $s$ point traces out the whole curve $(C)$ clockwise, the over-all change in phase of $F(s)$ is

$$\Delta \Phi = 2\pi (P - Z) \quad (16-3)$$

In this equation, positive phases are counted counterclockwise. $P$ and $Z$ are the numbers of poles and zeros (taking their order into account) of the $F(s)$ function which lie inside the curve $(C)$. In other words, the number $N$ of counterclockwise revolutions accomplished by the $(\Gamma)$ locus about the origin is given by

$$N = P - Z \quad , \quad (16-4)$$

2. Explanation for the Case in Which $F(s)$ is a Rational Function. Writing $F(s)$ in factored form to illustrate its zeros $z_1, z_2, \ldots$ and its poles $p_1, p_2, \ldots$, one can write

\[ F(s) = k \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots} \]

with \( z_i = z_j \) or \( p_i = p_j \) for zeros or poles of multiple order. Let \( Z_1, Z_2, \ldots, P_1, P_2 \) be the points \( z_1, z_2, \ldots, p_1, p_2, \ldots \) and \( M \) an ordinary point \( s \) of the \((C)\) curve. The phase shift of \( F(s) \) is then

\[ \Phi = \chi_1(s - z_1) + \chi_2(s - z_2) + \cdots - \chi_1(s - p_1) - \chi_2(s - p_2) - \cdots \]

that is to say

\[ \Phi = \chi_1(Ox, Z_1M) + \chi_2(Ox, Z_2M) + \cdots - \chi_1(Ox, P_1M) - \chi_2(Ox, P_2M) - \cdots \]

The increment \( \Delta \Phi \) of the phase shift of \( F(s) \) is equal to the sum of the increments of all its terms when \( M \) describes the whole \((C)\) curve clockwise.

In order to evaluate \( \Delta \Phi \), one may consider successively the different factors of \( F(s) \).

a. For the zeros. If a zero \( z_1 \) is exterior to \((C)\), Fig. 16-3 shows that

\[ \Delta(\chi_1 Ox, Z_1M) = 0 \]

If a zero \( z_2 \) lies within \((C)\), Fig. 16-4 shows that \( \Delta(\chi_1 Ox, Z_1M) = 2\pi \). Thus, if \( Z \) is the number of zeros within \((C)\), the value of the sum of the terms relative to the zeros will be

\[ \Delta(\chi_1 Ox, Z_1M) + \Delta(\chi_2 Ox, Z_2M) + \cdots = -2\pi Z \]

![Fig. 16-3.](image)

![Fig. 16-4.](image)

b. For the poles. Pole \( P_1 \) is exterior to \((C)\). Figure 16-5 shows that

\[ \Delta(\chi_1 Ox, P_1M) = 0 \]

If a pole \( P_2 \) lies within \((C)\), Fig. 16-6 shows that \( \Delta(\chi_1 Ox, P_2M) = -2\pi \) or

\[ -\Delta(\chi_1 Ox, P_2M) = 2\pi \]

Thus, if \( P \) is the number of poles within \((C)\), the value of the sum of the terms corresponding to the poles will be

\[ -\Delta(\chi_1 Ox, P_1M) - \Delta(\chi_2 Ox, P_2M) - \cdots = 2\pi P \]

![Fig. 16-5.](image)

![Fig. 16-6.](image)
In conclusion, if $Z$ and $P$ are the respective numbers of zeros and poles of $F(s)$ within (C) as point $M$ describes (C) clockwise, the phase shift of $F(s)$ is $\Delta \Phi = 2\pi (P - Z)$.

3. Application to $F(s) = 1 + KG$ and to Nyquist's Contour. This result will be used in the special case in which $F(s)$ is equal to $1 + KG(s)$. The closed curve (C), called Nyquist's contour, consists of (a) a line parallel to the $j$ axis at an infinitesimal distance to the right of it and (b) a semicircle of infinite radius situated in the right-hand plane (Fig. 16-7). Any pole or zero of $1 + KG(s)$ with a positive real part is, therefore, within this (C) contour. As point $M$ describes this contour clockwise, point $[1 + KG(s)]$ describes the transfer locus $1 + KG(j\omega)$ from low to high frequencies, as well as the symmetrical locus $1 + KG(-j\omega)$. Let $Z$ be the number of zeros of $1 + KG(s)$ with positive real parts and $P$ the number of poles of $1 + KG(s)$ with positive real parts, taking their order into account. The result found in Sec. 16.1.2, par. 2, shows that the phase-shift increment of $1 + KG(s)$ when $\omega$ increases from $-\infty$ to $+\infty$ is $2\pi (P - Z)$.

4. Equivalent Expression. The $KG$ and $1 + KG$ functions have the same poles. By calling the critical point point $A (-1, 0)$ the result found above is equivalent to the following: If point $s$ describes once the Nyquist contour clockwise in the $s$ plane, the number of counterclockwise circumvolutions of point $KG(s)$ around the critical point is equal to $P - Z$, or

$$N = P - Z$$

$Z$ being the number of zeros of $1 + KG(s)$ with positive real part and $P$ the number of poles of $KG(s)$ with positive real part. As point $s$ describes the Nyquist contour, point $KG(s)$ describes the whole Nyquist locus of $KG(s)$, that is, the $KG(j\omega)$ locus with $\omega$ varying from $-\infty$ to $+\infty$.

16.1.3. Nyquist's Criterion. For a feedback control system defined by its open-loop transfer function $KG(s)$, Eq. (16-4) yields the number $Z$ of unstable zeros for $1 + KG(s)$ as a function of (a) the number of unstable poles of $KG(s)$ and (b) the number of revolutions of the whole Nyquist locus [$KG(j\omega)$ locus from $\omega = -\infty$ to $\omega = +\infty$] around the critical point.

The general condition for stability, $^1 Z = 0$, becomes $N = P$. Thus, a feedback control system is stable if the number of counterclockwise revolutions around the critical point of its open-loop transfer locus, considered from $\omega = -\infty$ to $\omega = +\infty$, is equal to the number of unstable poles of its open-loop transfer function. This is a necessary and sufficient condition.

In particular, if $P = 0$, it is seen that a feedback control system with a

$^1$ The case of a zero of $1 + KG$ lying on the $j$ axis corresponds to the case in which the $KG$ locus passes through the critical point.
stable forward path is stable if the open-loop transfer locus does not enclose the critical point.

16.1.4. Practical Application. 1. Principle. The open-loop transfer locus $KG(j\omega)$ of the feedback control system, as well as the symmetrical locus $KG(-j\omega)$, is sketched, and the number of revolutions around the critical point ($-1$) are counted. Then condition (16-4) shows if the system is stable or unstable and, in case of instability, yields the number of unstable zeros of the characteristic equation $1 + KG(s) = 0$.

2. Although Nyquist’s criterion is directly derived from a mathematical theorem, its practical application is involved and the only way to avoid errors is to proceed methodically. Special attention must be paid to three points: the open-loop instabilities, the evaluation of the number of revolutions, and the event of $KG = \infty$ for $\omega = 0$.

a. Open-Loop Instabilities. In order to apply the criterion, one must know the number of unstable poles of $KG(s)$. This is no real difficulty, for $KG(s)$ is the product of transfer functions of elements which are given either mathematically or experimentally and hence are known to be stable or unstable.

b. Number of Revolutions. A safe way of finding this number is to come back to the phase shift by dividing the variation of the angle $\chi(Ax,AN)$ by $2\pi$ when point $N$ describes the $KG(j\omega)$ locus from $\omega = -\infty$ to $\omega = +\infty$, $A$ being the critical point and $Ox$ the real axis (Fig. 16-8a).

c. Infinite value of $KG(j\omega)$. The $KG(s)$ function often has a pole for $s = 0$, with the transfer locus $KG(j\omega)$ tending toward an asymptote of angle $-m\pi/2$, $m$ being the order of the pole. One must then know
what happens between the values of $\omega = 0^-$ and $0^+$: does $KG(0)$ have a positive infinity (Fig. 16-8b) value or a negative one (Fig. 16-8c)? There are different ways of proceeding. One can:

1. Suppose $s = \epsilon$ if the expression of $KG(s)$ is known, $\epsilon$ being a small real positive quantity. A real value is hence obtained for $KG(\epsilon)$, the sign of which is what one is looking for. (This method can be used when there is one integration.)

2. Physically analyze the static behavior. (This method can also be used when there is one integration.)

3. Consider the path followed by point $KG(s)$ as $s$ approaches zero. This is done by continually observing the phase variation of $KG(s)$ as point $s$ describes an infinitely small semi-circle around the origin in the right-hand plane (Fig. 16-9). In the authors’ opinion, this third method is to be preferred whenever the mathematical expression of $KG(s)$ is known. It is the most natural method when one has understood the mathematical meaning of Nyquist’s criterion.

3. First Example. For a second-order feedback control system

$$KG = \frac{K}{s(1 + Ts)}$$

the first method yields

$$KG(\epsilon) \cong \frac{K}{\epsilon} > 0$$

The second method yields the same result; for if a small positive constant error produced a very large positive output, the system would amplify the errors instead of correcting them.

The third method shows that, when point $s$ is near the origin,

$$KG \cong \frac{K}{s}$$

One has, therefore, Table 16-1.

<table>
<thead>
<tr>
<th>Initial phase shift</th>
<th>Final phase shift</th>
<th>Phase*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>$-\pi/2$</td>
<td>Increases</td>
</tr>
<tr>
<td>$1/s$</td>
<td>$\pi/2$</td>
<td>Decreases</td>
</tr>
<tr>
<td>$K/s$</td>
<td>$\pi/2$</td>
<td>Decreases</td>
</tr>
</tbody>
</table>

* Counted counterclockwise.

Hence, $KG(0)$ is infinite and positive (Fig. 16-8b).
4. Second Example. For a feedback control system with double integration,

\[ KG = \frac{K}{s^2(1 + Ts)} \]

The open-loop transfer locus is sketched in Fig. 16-10. When \( s \) approaches 0, one can write

\[ KG \approx \frac{K}{s^2} \]

whence Table 16-2.

<table>
<thead>
<tr>
<th></th>
<th>Initial phase shift</th>
<th>Final phase shift</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s )</td>
<td>(-\pi/2)</td>
<td>(+\pi/2)</td>
<td>Increases</td>
</tr>
<tr>
<td>( 1/s )</td>
<td>(\pi/2)</td>
<td>(-\pi/2)</td>
<td>Decreases</td>
</tr>
<tr>
<td>( 1/s^2 )</td>
<td>(\pi)</td>
<td>(-\pi)</td>
<td>Decreases</td>
</tr>
<tr>
<td>( +KG )</td>
<td>(\pi)</td>
<td>(-\pi)</td>
<td>Decreases</td>
</tr>
</tbody>
</table>

\( KG(j\omega) \) therefore describes, clockwise, a complete circle of infinite radius when \( s \) approaches 0 (Fig. 16-10).

16.1.5. Examples. 1. A second-order system is always stable (Fig. 16-11). Experience shows that any feedback control system can be made unstable by increasing the gain. Hence no real servo can be represented correctly by a second-order system.

2. A third-order system \( KG(s) = K/s(s^2 + 2zs + s) \) is stable for small values of \( K \), unstable for large values (Fig. 16-12a and b). For most systems this result corresponds with practical results: hunting appears when the servo is stiffened. Hence the fundamental dilemma: “soft” servo (bad accuracy) or hunting (bad stability). Compensating networks can help to resolve this difficulty (Sec. 18.1).

3. System with an unstable forward path, for instance a third-order system with an overcompensated integral controller (for the definition of these terms, see Sec. 18.3.2):

\[ KG(s) = \frac{s(s^2 + 2zs + 1) + s}{s^2} \]
The pole $s = 1$ causes instability in the forward path, so the condition for stability is, therefore, $N = 1$. Figure 16-13a and b shows the Nyquist locus for two values of the gain.

16.1.6. Application to the Inverse Transfer Locus. Nyquist's criterion can be applied, in theory, to the inverse transfer locus. However, it is advisable, in practice, not to proceed in such a manner, for the two following reasons:

1. If the criterion is to be applied to direct loci, it is necessary to know the number $P$ of unstable elements in the forward path. Likewise, it is necessary to know the exact number of all-pass elements if it is desired to apply the criterion to inverse transfer loci (Sec. 8.5.2). If it is easy to determine whether or not an element is stable, it is often more difficult to determine if it is a minimum-phase-shift system.

2. Direct loci have an infinite branch only when $KG$ has a pole for $s = 0$. For inverse loci there is always an infinite branch, as all transfer functions are equal to
zero for very high frequencies. Furthermore, the infinite branch no longer corresponds to the easily determined static condition, but rather to the behavior of the system for high frequencies, which is difficult to study experimentally and which is very often improperly represented by transfer functions that are valid only at low frequencies.

16.1.7. Nonunit Feedback (Fig. 16-14). One has

\[ H(s) = \frac{S}{E}(s) = \frac{K_1G_1}{1 + K_1K_2G_2G_2} \]

This happens when the feedback loop incorporates a compensating network. Such a system can be studied by making use of the remark of Fig. 18-4.

16.1.8. Final Remark. In order to avoid frequent misunderstandings, it should be borne in mind that the application of Nyquist’s criterion to a transfer locus \( X(j\omega) \) indicates whether or not the system with the transfer function \( X(s)/(1 + X(s)) \) is stable (Fig. 16-15). Thus, if \( X(s) \) is the transfer function of the system to be controlled and if the Nyquist criterion applied to the function \( KG = kX(s) \) shows that the feedback system is unstable for all values of the gain of \( X \), it is impossible to achieve stability by means of purely proportional control.

16.2. VARIOUS APPLICATIONS

16.2.1. Conditional Stability. The system studied as Example 3 in Sec. 16.1.5 was found to be stable only for values of the open-loop gain lying between two critical values. This situation can be dangerous, since low values of the open-loop gain cause instability. This is called conditional stability, and it occurs for all systems having an unstable forward path. When the feedback loop is closed, one must not forget that the system is unstable if the gain is too low. This may happen, for example, in the case in which the power stage is an electronic one, when the amplifiers have not finished warming up or, in the case of a stabilized missile, immediately after launching when the control surfaces are not effective enough because of the low speed of the missile.

16.2.2. Unstable Secondary Loop. A frequent case of feedback control systems with an unstable forward path is that in which there is a secondary loop, as in Fig. 16-16. The over-all stability conditions can be satisfied. However, it is preferable to avoid this, lest an accidental drop of the open-loop gain cause a catastrophe.

\(^1\) At least for positional servomechanisms.
16.2.3. Hunting. This phenomenon has already been referred to in Secs. 13.2.3 and 14.3.2. It is now easier to understand. For the user, hunting is essentially an undesirable oscillation which takes place in a feedback control system when he tries to stiffen it by increasing its open-loop gain, in order to improve the accuracy. The linear theory easily explains why the system becomes unstable when the open-loop gain is increased (Fig. 16-17). This increase is equivalent to amplifying the curves without changing their shape (Sec. 13.3.1). For a large $K$, the locus crosses the critical point and the system is hence unstable. But there are always some limitations, such as nonlinear saturation (Secs. 22.3 and 24.3.2), which prevent the unstable variables from increasing to infinity and so limit their magnitude to a fixed value. Hence the system tends toward a permanent oscillation with a fixed magnitude, whatever the initial conditions may be. The frequency is that for which the phase shift of $KG(j\omega)$ is $\pi$. Figure 16-17 shows that it is not exactly the resonance frequency of the system. As a rule, hunting is caused by unwanted time lags (see Sec. 13.4).\footnote{The "sinusoids" of a drunkard may be cited as an example of hunting for a closed-loop system; they are caused by the increasing brain response time on account of the alcoholic impregnation.}

One way to prevent it is to compensate these lags whenever possible (see Sec. 18.2). But it is better to be careful about the lags at the outset—prevention is better than cure.

In certain circumstances, it may be desirable that a servo system should hunt; for instance, its hunting oscillation may be used as a carrier for modulation or linearization purposes (Sec. 28.1).

Remark. It must be clearly understood that linear theory cannot account for hunting. As a matter of fact, if linear operation is assumed, the only possible oscillation with constant magnitude corresponds to the case of an oscillatory system with purely imaginary conjugate poles. This occurs when the $KG(j\omega)$ locus passes through the critical point. But (a) this is a purely mathematical case, a real system being either stable or unstable, and (b) the oscillation magnitude varies with the initial conditions, which is not the case for hunting.

16.2.4. Longitudinal Stabilization of an Airplane. Suppose it is desired that a rudder have a deflection proportional to an input voltage $v$. The first device one
thinks of, involving field control and feedback as shown in Fig. 16-18, is not stable with a purely proportional control

$$C = K(v - v_b)$$

where $C$ is the motor torque. Indeed, if the rudder is characterized by its inertia $J$, one has

$$C = J \frac{d^2\delta}{dt^2}$$

Hence, the open-loop transfer function (time lags represented by $T$) is

$$\frac{B(s)}{v - v_b} = \frac{K}{Js^2(1 + Ts)}$$

The locus is sketched in Fig. 16-19. Nyquist's criterion shows that the system is unstable:

$$P = 0 \quad \Delta \Phi = -4\pi \quad N = -2 \quad Z = 2$$

It can be stabilized by a lead network (Sec. 18.2) located in the controller. $K$ is replaced by

$$\frac{K}{\alpha} \frac{1 + \alpha \tau s}{1 + \tau s} \quad \text{with} \quad \alpha > 1$$

On the diagram, this means that the $KG(j\omega)$ locus is inflated as it approaches the critical point, so as to be on the other side (Fig. 16-20). The Nyquist criterion then shows (Fig. 16-21) that the system is stable:

$$P = 0 \quad \Delta \Phi = 0 \quad N = 0 \quad Z = 0$$
16.3. THE LEFT-HAND CRITERION

16.3.1. Statement. The following criterion for stability is a very convenient one to use. A servo system is stable if its open-loop transfer locus $KG(j\omega)$, traced out with increasing frequencies, always leaves the critical point $A(-1)$ at its left.

The left-hand criterion is extremely easy to apply. The reader can apply it to the above examples treated by the Nyquist criterion and see that the answer about the stability of the system is most readily obtained. Other examples will be dealt with in Sec. 16.4.5.

The application of the left-hand criterion involves no pitfalls when the function $KG(j\omega)$ has no poles or zeros with a positive real part, that is, when the system is known to be a stable and minimum-phase shift one in open-loop operation. If this condition is not fulfilled, the conditions for the validity of the criterion become somewhat involved\(^1\) and it is safer to use the Nyquist criterion.\(^2\)

Remark. If Nichols loci are used, the words right and left should be interchanged, because of the positive directions usually chosen. The criterion then becomes the right-hand criterion.

16.3.2. Interpretation of the Left-hand Criterion. The significance of the left-hand criterion can be established as follows: Consider in the $s = \alpha + j\omega$ plane the constant-$\alpha$ and the constant-$\omega$ loci. They are straight lines which are vertical and horizontal, respectively. The locus of the tip of the vector $KG(s) = KG(\alpha + j\omega)$ when $\alpha$ has a fixed value and $\omega$ varies is a curve in the $KG$ plane. If $\alpha$ is zero, the curve is the Nyquist locus. If $\alpha$ is different from zero, the locus is a curve "parallel" to the Nyquist locus (Fig. 16-22). The loci $KG(\alpha + j\omega)$ with fixed values of $\alpha(\alpha_1, \alpha_2, \alpha_3, \ldots)$ are said to be the maps in the $KG$ plane of the constant-$\alpha$ lines ($\alpha = \alpha_1, \alpha_2, \ldots$). Similarly, the constant-$\omega$ lines of the $s$ plane have maps in the $KG$ plane which are the loci of the tips of the vectors $KG(\alpha + j\omega)$ when $\omega$ is fixed and $\alpha$ varies. These loci are orthogonal to the previous ones if the scales are so chosen that one octave ($\omega_2/\omega_1 = 2$) is represented by the same length as 1 neper ($\alpha_2/\alpha_1 = 2$).

Thus it is possible, starting from the $KG(j\omega)$ locus, to extrapolate the $KG(\alpha_1 + j\omega)$ loci for values of $\alpha$ in the neighborhood of zero. This can be done by constructing from the $KG(j\omega)$ locus an orthogonal set of lines. The critical point $A(-1)$ is thus located at the intersection of a constant-$\alpha$ and a constant-$\omega$ curve in the $KG$ plane. This $\alpha$ and this $\omega$ characterize

\(^1\) See Probs. 34 and 35.

the dominant mode of the system $e^{-at}$ $\sin (\omega t + \varphi)$. The value of $\alpha$ indicates the damping, which is positive if $\alpha < 0$, that is, if the critical point $A$ lies at the left of the $KG(j\omega)$ locus. This is the left-hand criterion.


16.4. PRACTICAL CONCEPT OF STABILITY

16.4.1. Introduction. The study of stability is a very important part of the theory of linear systems, for nearly all systems (instrumentation, control, etc.) must be stable. But it must not be forgotten that a mere condition of stability, as studied in the previous sections, is by no means enough to make the performance of the system satisfactory; for example, badly damped transient oscillations of a stable system which approaches instability may be intolerable.

The fact that a linear system satisfies the stability condition of Sec. 9.1 means only that its characteristic equation has no zeros with a positive real part (Fig. 16-23). But some zeros may be troublesome, even though they have a negative real part. This will be the case if the modes converge too slowly (real negative zero, with too small a magnitude, Fig. 16-24) or if the oscillations are insufficiently damped ($\psi$ too small, Fig. 16-25). Hence the mathematical stability (no zero in the right-hand plane) is not necessarily a "good" stability (no zero "too near" the $j$ axis).

These considerations apply also to Nyquist's criterion, as it is only a particular application of the general condition of Sec. 9.1.2; this criterion evidences a mathematical stability only, not a "good" stability. The Nyquist criterion limit condition for stability is that the locus $KG(s)$ goes through the critical point $(-1)$. Hence a condition for a "good"
stability is that the locus does not approach too closely the critical point: the degree of stability can be reckoned by the distance of the $KG(j\omega)$ locus from the critical point. For regular systems, use is made of the gain and phase margins or of a specification of the resonance ratio $Q$.

16.4.2. Gain and Phase Margins. 1. Definition. The degree of stability is often defined by use of the phase margin $\Phi_m$ and the gain margin $G_m$, as defined in Fig. 16-26 for a regular system. The phase margin is the amount of phase lag allowable before the system becomes unstable, and the gain margin is the number of decibels which can be added to the gain before instability is produced. The gain and phase margins are shown in Fig. 16-27, using Nichols coordinates.

2. Significance of These Margins. The gain margin ensures that the system will be stable in spite of unexpected variations of the open-loop gain. Consider, for example, an aircraft autopilot which is set for given flight conditions (horizontal flight at altitude $Z$ and speed $v$). If the speed varies and becomes $1.4v$, the gain of the aircraft’s transfer function is multiplied by approximately the square of the ratio of the speeds, i.e., by 2. Under these conditions, the control system becomes unstable if the gain margin is not at least 6 db.

The phase margin ensures that the system will be stable in spite of the unwanted time lags which were not taken into account at the time of the setting. For example, if the phase margin is $50^\circ$ and if the lags cause a $10^\circ$ phase shift in the resonance region, they are said to have utilized one-fifth of the phase margin.

3. Orders of Magnitude. A reasonable degree of stability is generally considered as being provided by a phase margin of approximately $45$ to $50^\circ$ and a gain margin varying between 10 and 15 db. In regulator problems, somewhat lower values are generally admitted; for example, $30^\circ$ and 5 to 10 db.

16.4.3. Specification for the Resonance Ratio. The resonance ratio has been defined previously (see Sec. 8.1.1, par. 3, and Sec. 13.2.3). Specifying that the resonance ratio shall not exceed a value of the order
of 1.25 to 1.5 is another way of guaranteeing good stability, by making sure that the \( KG(j\omega) \) locus will not enter the \( MO/MA = 1.3 \) circle (Sec. 13.3.3). It can easily be seen that this condition is approximately equivalent to the condition for 10-db gain margin and 45° phase margin, as the \( \psi \) angle for \( Q = 1.3 \) is precisely 50° (Table 13-1).

16.4.4. Leonhard’s Criterion. The requirement that the region of operation for the system shall remain to the left of a parallel to the imaginary axis at a distance \( -\alpha \), or to the left of two oblique straight lines passing through the origin at an angle arcsin \( \gamma \) (Figs. 9-4 and 9-5), can be expressed in frequency-response terms. To do this, one must draw the locus of \( KG(s) \) as the point \( s \) traces out the limiting contour and count the number of times it encloses the critical point. This leads to drawing the loci

\[
\begin{align*}
KG(\alpha + j\omega) & \quad \text{for given } \alpha \\
KG[(1 \pm j \text{arcsin } \gamma)\lambda] & \quad \text{for given } \gamma
\end{align*}
\]

as \( \omega \) or \( \lambda \) varies. This has been done by Leonhard.\(^1\)

![Fig. 16-28. Nyquist locus for \( KG = K/s^4 \).](image)

![Fig. 16-29. Nyquist locus for](image)

\[ KG = K \frac{1}{s^2 + Ts} \]

16.4.5. Example of Stabilization: Roll Stabilization of an Airplane. Consider a roll-stabilized airplane under the assumptions stated in Sec. 14.3.6 (Fig. 14-23). As a first approximation the aircraft is described by the transfer function

\[
\frac{\Phi}{\Delta}(s) = \frac{A}{J} \frac{1}{s^2}
\]

and the servomotor stage by a lag factor

\[
\frac{\Delta}{H}(s) = \frac{1}{1 + Ts}
\]

The Nyquist locus of

\[ KG(s) = K \frac{1}{s^2 + Ts} \]

shows, by application of the left-hand criterion, that the noncompensated system is unstable, whatever \( K \) may be (Fig. 16-29).

If the first derivative of the error is introduced by the controller, the open-loop transfer function becomes

\[
\frac{K(1 + aTs)}{s^2(1 + Ts)}
\]

The left-hand criterion shows (Fig. 16-30) that the system is stable for all values of \( K \) when \( a > 1 \). But at high frequencies the \( KG \) locus approaches the origin along the negative real axis. This indicates a poor phase margin, especially when \( a \) is small.

If the second derivative is introduced, \( KG(s) \) becomes
\[
\frac{K}{s^2} \frac{1 + aTs}{1 + Ts} \frac{bT^2s^2}{1 + T_s}
\]

Two cases may be defined.

Case 1. \( a > 1 \) (Fig. 16-31). The system is always stable. The effect of the term \( bT^2s^2 \) is to make the locus tangent to the negative \( j \) axis at high frequencies, that is, to bulge it in the direction of positive phase shifts. This results in an increase of

![Fig. 16-30. Nyquist locus for \( KG = \frac{K}{s^2} \frac{1 + aTs}{1 + Ts} \), with \( a > 1 \).](image)

![Fig. 16-31. Nyquist locus for \( KG = \frac{K}{s^2} \frac{1 + aTs + bT^2s^2}{1 + Ts} \), with \( a > 1 \).](image)

System is regular.

Fig. 16-32. Nyquist locus for \( KG = \frac{K}{s^2} \frac{1 + aTs + bT^2s^2}{1 + Ts} \frac{1}{1 + T_s} \frac{1}{1 + T_2s} \), with \( a > 1 \).

Fig. 16-33. Nyquist locus for \( KG = \frac{K}{s^2} \frac{1 + aTs + bT^2s^2}{1 + Ts} \), with \( a < 1 \). System is not regular.

![Fig. 16-34. Nyquist locus for \( KG = \frac{K}{s^2} \frac{1 + aTs + bT^2s^2}{1 + Ts} \frac{1}{1 + T_s} \frac{1}{1 + T_2s} \), with \( a < 1 \).](image)

System is conditionally stable.

Note that the presence of additional lags \( 1/(1 + T_1s) \) and \( 1/(1 + T_2s) \) causes the locus to approach the origin along the positive \( j \) axis, thus resulting in an instability for high values of \( K \) (Fig. 16-32). The system is regular.

Case 2. \( a < 1 \). The system is stable only for high values of \( K \) (Fig. 16-33) and thus is not regular. Taking additional lags into account causes hunting to appear for high \( K \) (Fig. 16-34), thus making the system conditionally stable.

This discussion closely parallels the considerations outlined in Sec. 14.3.6 as an application of the root-locus method.
CHAPTER 17
PERFORMANCE CRITERIA

Summary
1. Ideal control systems and practical specifications.
2. A set of general-purpose criteria.
3. Application to the evaluation of the effect of time lags.
4. More elaborate criteria.
5. Practical use of performance criteria.

Accurate criteria are of great theoretical importance in evaluating the performance characteristics of a feedback control system. Their practical importance appears, however, only when the field has been developed to a satisfactory degree. Very good systems for performing simple tasks have been designed for years without the use of rational criteria. Today the technique of feedback control systems has become more elaborate and has been applied to new tasks, which makes the use of accurate performance criteria a necessity.

Similarly, in the early days of aviation the ambition of inventors was to build a machine that would be able to fly. Today, however, airplanes are designed with a view to fulfilling a certain need. The same transition has taken place in servo systems. The purpose of a stabilizer is not merely to stabilize, as was the case when, in the early 1930s, the first model planes were autopiloted, but to stabilize properly. Difficulties begin when one attempts to explain what is meant by "properly," that is, when criteria are to be expressed in explicit form. Thus, the necessity for explicit performance criteria in the automatic-control field indicates that this field has become technically mature.

The nature of the criteria used has varied to a great extent since the first explicit criteria were established in the early 1940s. This evolution is typical of that in automatic-control techniques. For example, the vogue of statistical criteria to take into account noise disturbances indicates that the increasing quality of servo systems has made noise the limiting factor in system accuracy.

It is for the above reasons that this chapter may be considered the most important of the entire book for the reader interested in the philosophy and historical development of automatic control.

17.1. IDEAL CONTROL SYSTEMS AND PRACTICAL SPECIFICATIONS

17.1.1. Ideal Control Systems. A system which slaves an output $r(t)$ to an input $e(t)$ is one for which
\[ r(t) = e(t) \quad (17-1) \]

whatever may be \((a)\) the variations of the input (control problem) and \((b)\) the disturbances (regulation problem). Problem \((a)\) is mainly one of speed of response, while problem \((b)\) is one of system stiffness.

If the deviation, or error, is defined by

\[ \epsilon(t) = e(t) - r(t) \]

the ideal feedback control system is one for which

\[ \epsilon(t) = 0 \quad (17-2) \]

whatever may be the control input and the disturbances (Fig. 17-1).

It is clear that neither relation \((17-1)\) nor \((17-2)\) can be satisfied by any real servomechanism. In fact, as was seen in Sec. 13.1.7, Eqs. \((13-1)\) and \((13-3)\), the error is the sum of two terms, one corresponding to its operation as a servo system, the other to its operation as a regulator, neither of which can be made equal to zero.

Neither \((1)\) the error due to changes in the command \(e(t)\) (Fig. 13-2), since

\[ \frac{KG(s)}{1 + KG(s)} = 1 \]

at all frequencies would require \(|KG(s)| = \infty\), whereas for all real mechanisms \(KG(j\infty) = 0\), nor \((2)\) the error due to disturbances (Fig. 13-9 or 13-10), since

\[ \frac{KG_2}{1 + KG_1KG_2} = 0 \]

would require \(|KG_1(s)KG_2(s)| = \infty\).

Having any of the two error terms zero leads to the same unrealizable condition, since all real mechanisms comply with \(KG(\infty) = 0\).

Thus, ideal servo systems are not realizable, zero-error condition being an ideal that cannot be achieved. A realistic approach then consists of trying to keep the error as small as possible. But, what does "as small as possible" mean exactly? In practice, the first approximation to it is obtained by the designer from the specifications stipulated.

**17.1.2. Some Usual Performance Specifications.** For any design problem, the specifications are the basis for the designer's work. Unfortunately, sponsors' specifications in the automatic-control field generally seem despairingly vague to the design engineer, who may literally have to fight to make the sponsor specify exactly what he means when he says that the system must "do the job."

\[ ^1 \text{The philosophical reason for this difficulty is, of course, that servo systems are intended to meet essentially unexpected inputs; thus, a quantitative statement of the specifications would imply that a preliminary statistical study of the control input and disturbances had been conducted. This, however, is the case in less than one per cent of the problems a designer meets.} \]
 Altogether, once the specifications have been stated, experience proves that they are amazingly similar, in spite of the great variety of automatic-control problems. If purely technological requirements such as correct operation in hot or cold weather or under certain accelerations are excepted, the performance specifications pertaining to the operation of the system as a controller usually include the following:

1. The sponsor will demand that the system be safe. It is obvious that a stabilizer should stabilize and never destabilize, that a regulator should not run away, etc. Furthermore, if the sponsor has had any experience with servo systems, he will know what hunting is, and consequently he will demand that hunting be prevented. This demand is often stated by requiring that the first overshoot of the step response be limited to a certain value.

2. The sponsor will always demand that the system possess a certain accuracy. To the question "under what conditions?" he generally answers, "in the steady-state condition," i.e., in a static condition or under constant-velocity operation.

3. Finally, the sponsor will demand that the system be sufficiently fast, i.e., that it be able to follow fast and rapidly varying inputs, or equivalently, that transient phenomena should not last too long.

The designer will then have to interpret the specifications in terms of feedback-control parameters.

17.2. A SET OF GENERAL-PURPOSE CRITERIA

17.2.1. General. Stated in feedback-control terms, the stipulated specifications and the corresponding performance criteria are equally uniform. In that sense, in spite of the infinite variety of automatic-control problems, it is possible to speak of general-purpose criteria. The qualification "general-purpose" does not imply that the corresponding criteria can be used blindly for solving any problem. Nevertheless, such general-purpose criteria can be of great use to the designer, provided he masters the fundamentals underlying them.

17.2.2. Statement. For most practical applications, a feedback control system can be considered "good" if it (a) is stable, (b) has a high open-loop gain, and (c) has a high resonance frequency. Some explanations concerning the significance and the limitations of these criteria are given below.

17.2.3. Stability. A feedback control system must, above all, be stable; furthermore, it must be more than just marginally stable. As shown previously, the effect of the critical point merely lying to the left of the \( K \omega \) locus is not necessarily a guarantee of sufficient stability. The latter is obtained only by ensuring that there is a reasonable separation between the locus and the critical point.

Two criteria are used for expressing the separation that is required to ensure sufficient stability. The first is a phase margin of about 50° with a gain margin of 10 to 15 dB. The second is a resonance ratio \( Q \) of about 1.3 (2.3 db).

The stability condition is peremptory: a feedback control system must
be stable. This condition must be fulfilled before going to conditions b and c of Sec. 17.2.2.

17.2.4. High Open-loop Gain. If $K$ is the value of $KG(s)$ for $s = 0$, the positional error is inversely proportional to $K$; that is, the response to an input $e(t) = eu(t)$ is such that $e_0 - r(+\infty) \cong e_0/K$. This expression shows why it is desirable to make the open-loop gain of a system very high.

If $KG(s)$ has the pole $s = 0$ (one integration), the positional error is equal to zero, but the velocity error is not; it is inversely proportional to the velocity constant $K_v$. It is desirable, therefore, to increase $K_v$.

On the other hand, considered as a regulator, a feedback control system is stiffer for disturbances if the open-loop gain is high. This can be seen from Eq. (13-4) of Sec. 13.1.7:

$$\frac{e}{D} = \frac{-K_3G_3}{1 + K_1K_2G_1G_2}$$

Therefore, the high-open-loop-gain criterion may be applied in most cases.

17.2.5. High Natural Frequency—Large Bandwidth. A high natural frequency, or a large bandwidth, is a measure of the system's speed of response and has been dealt with in detail in Part 1 (Sec. 6.4.5). The application of this criterion is less general than that of the high-open-loop-gain criterion, and some caution is necessary when using it. Two considerations temper the desire for a high natural frequency:

a. It must be remembered that the transfer functions of a system's elements are generally defined only for a limited bandwidth $(0,F)$. Therefore, if the bandwidth of the entire feedback control system is larger than $F$, the mathematical representations of the transfer functions may not be valid. For example, the transfer functions of an aircraft are seldom correct for frequencies greater than 3 cps. Consequently, if the design of an automatic pilot should result in a natural frequency of 3 cps for the piloted aircraft, there would be reason for carrying out a more accurate analysis of the phenomena.

b. The bandwidth must not be unduly increased because of noise effects (Sec. 12.3.4). This restriction, however, seldom occurs in servomechanisms.

17.2.6. Background Underlying the General-purpose Criteria. Summarizing the above, these general-purpose criteria are just a translation into servomechanism terms of the specifications usually stipulated:

a. Stability is a guarantee for safety. Good stability is a guarantee against hunting.

b. A high open-loop gain is an attempt at minimizing the position error or, if the latter is canceled by the infinite static gain, the velocity error.

c. A high resonant frequency is the condition for a fast transient and a satisfactory speed of response.

The major advantage of the criteria is that they are expressed in convenient terms from the designer's viewpoint. In fact, as follows
from Chaps. 13 to 15, the quantities they involve can be read from the \( KG(j\omega) \) plot, which thus constitutes the fundamental tool for the designer.

**17.2.7. Practical Application.** When, in the design of a feedback control system, a parameter \( \lambda \) is unspecified, the open-loop gain \( K \), corresponding to a \( Q \) of 1.3, is determined for each value of \( \lambda \) compatible with the stability. The choice of \( \lambda \) is the result of a compromise between a high open-loop gain and a large natural frequency. For instance, if the results shown in Table 17-1 are found, the value chosen would be \( \lambda = 0.2 \). A particularly instructive example will be studied in detail in the following chapter (lead-network adjustment, Sec. 18.2.6).

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( K_1 )</th>
<th>( \omega_R ) (rad/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.58</td>
<td>0.6</td>
</tr>
<tr>
<td>0.05</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>0.1</td>
<td>1.04</td>
<td>0.9</td>
</tr>
<tr>
<td>0.2</td>
<td>1.04</td>
<td>1.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.46</td>
<td>1.3</td>
</tr>
</tbody>
</table>

**17.2.8. Important Remark.** A thorough understanding must be retained at all times of the difference between (a) stability, (b) stability margin (phase and gain margins, or resonance ratio), and (c) high-open-loop-gain and large-bandwidth criteria:

a. The stability is a peremptory condition.

b. The stability margin ensures against such factors as small setting variations or time lags which have not been accounted for in the equations. As a performance criterion, its only value is that it generally guarantees a good damping.

c. The high-gain and high-natural-frequency criteria are true performance criteria, the first with respect to the steady state, the second with respect to the transient state.

**17.2.9. Second Important Remark.** As already pointed out, no criterion can be absolutely general; therefore, care must be taken when applying the above "general-purpose" criteria, since they may not be appropriate for some particular types of problems. As a rule, the above general-purpose criteria can be safely applied wherever the system is regular as defined in Secs. 8.3.9 and 14.3.2. But when nonconventional systems are under consideration, it is often advisable to study the problem more thoroughly.

For example, however unusual, there are systems whose speed of response becomes higher as their resonant frequency is made lower. The reason is that their speed of response is controlled by another mode of the system, an example being given in Prob. 39.

Even the criterion that a control system must be stable is subject to exceptions. For example, when stabilizing a missile just after its launch, there is no need to compensate for the very slow unstable mode, with a time constant of the order of 1 min, if the missile is to be captured by a radar beam 20 sec after launch.
In conclusion, it can be said that the above general-purpose criteria concentrate a great amount of experience and that applying them can save much time and thought. But engineers should be very prudent in sparing thought, especially when they are coping with nonconventional problems.

17.3. EVALUATING THE EFFECT OF TIME LAGS

17.3.1. General Method. The above criteria will now be applied to the evaluation of the effect of time lags. It has been shown in Sec. 13.4 that, in general, time lags cause a deterioration in the performance of a feedback control system. A method that is used for quantitatively specifying the effect of time lags (for example, in order to prescribe an upper limit to the tolerable lags) consists in comparing, in the viewpoint of the above criteria, the performances of the system with and without lags. This is accomplished by setting the gain to \( K_{1,2} \) for the various cases and measuring the increments of gain and of natural frequency.

Example. Without unwanted time lags, a system has an open-loop gain \( K_{1,2} = 0.46 \) and a natural angular frequency \( \omega_R = 0.8 \text{ rad/sec} \). Unwanted time lags, represented by a time constant \( T \), deteriorate the performance, as shown in Table 17-2. Thus, a 0.2-sec time lag causes a deterioration in performance (accuracy and speed) of about 10 per cent. For many applications, this amount of deterioration is the maximum due to the effects of unwanted time lags that may be accepted.

<table>
<thead>
<tr>
<th>( T ) (sec)</th>
<th>( K_{1,2} )</th>
<th>( \omega_R ) (rad/sec)</th>
<th>( \Delta K_{1,2}/K_{1,2} )</th>
<th>( \Delta \omega_R/\omega_R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.46</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>0.43</td>
<td>0.75</td>
<td>-6.7%</td>
<td>-7%</td>
</tr>
<tr>
<td>0.2</td>
<td>0.41</td>
<td>0.72</td>
<td>-11%</td>
<td>-10%</td>
</tr>
<tr>
<td>0.3</td>
<td>0.39</td>
<td>0.63</td>
<td>-15%</td>
<td>-18%</td>
</tr>
</tbody>
</table>

Remark. A quicker, but less accurate, method of studying the effects of time lags has been shown previously in Chap. 13, for example in Sec. 13.4.3. The method described here differs from the above in that it compares the performances of the feedback control system with and without time lags after the open-loop gains have been set at their optimum values. This is much more representative of a practical problem.

17.3.2. Usual Approximation. A first approximation for the value of tolerable time lags is that which causes a decrease in the phase margin of 5 to 10°. If these time lags are represented by a time constant \( T \), the following condition is obtained (\( \omega_R \) being the closed-loop resonant frequency of the system).

\[
\frac{1}{1 + \omega_R T} \leq 5 \text{ or } 10^\circ = \frac{5}{57} \text{ or } \frac{10}{57} \text{ rad}
\]

that is,

\[
\omega_R T \leq 0.1 \text{ or } 0.2
\]
For example, for a servo with a 1-cps natural frequency,

$$\omega_R = 2\pi \times 1 = 6.28 \text{ rad/sec}$$

The unwanted time lags must be compensated for as soon as $6.28 T = 0.2$; that is,

$$T = 30 \text{ msec}$$

17.3.3. Important Consequences. 1. It can now be easily understood why milliseconds are relentlessly tracked down in all components of fast feedback control systems. This has led to the development of particular techniques of fundamental practical importance such as fast-starting servomotors with time constants less than 10 msec. An example of a fast feedback control system is a stabilized supersonic guided missile with a natural frequency of about 2 to 3 cps.

2. It can also be understood why the human operator must be replaced by a fast automatic device as a component of a feedback system having a rather high natural frequency. In this respect, the piloting of the fastest modern aircraft is not far from the limit of human capabilities. Because of his complexity, the human operator is very difficult to study. He is sometimes considered as introducing a lag which varies considerably with his training (its value is about 0.10 sec for an operator without training and under certain specific experimental conditions). This subject will not be studied here, although it is very interesting.\(^1\)

17.4. MORE ELABORATE CRITERIA

17.4.1. The Integral-squared-error Criterion, or Hall-Sartorius Criterion. A criterion that is more accurate than the preceding ones was proposed for the first time by A. C. Hall\(^2\) and H. Sartorius,\(^3\) working independently. In this, the quality of a servo's performance is evaluated from the integral of the square of the error

$$I = \int -\infty^\infty \varepsilon^2(t) \ dt$$

Thus, the smaller the value of $I$, the better the feedback control system. In general, $I$ is computed either for a step input (Fig. 17-2) or for a unit-impulse input.

It is easy to see that the Hall-Sartorius criterion, when applied to a step input, guarantees (a) that there is no position error, since the presence of such an error would cause the integral $I$ to become infinite, and (b) that the transient is not too slow, since the values of the error at the

---


3 "Die zweckmäßige Festlegung der frei wählbaren Regelungskonstanten," dissertation at the Technische Hochschule, Stuttgart, Germany, 1944.
beginning of the transient are predominant in the expression of \( I \) (Fig. 17-3). As a result, the integral-squared-error criterion leads to adjustments that are, in general, \textit{less damped} than is usually required (for most servo applications). This criterion is often applied, especially to regulator problems.

\textbf{17.4.2. Other Criteria. The ITAE Criterion.} The fact that the squared-error criterion leads to adjustments that are underdamped for

\[ e(t) \]

\[ r(t) \]

\[ \varepsilon(t) \]

\[ \varepsilon^2(t) \]

\textbf{Fig. 17-2. The Hall-Sartori criterion.}

\[ r(t) \]

\[ r(t) \]

\textbf{Fig. 17-3. Shaded area is predominant in integral squared error.}

many practical purposes arises from the presence of the square under the integral sign, which results in small errors being neglected for the sake of large ones. That is why it has been proposed to minimize the integral of the absolute value of \( \varepsilon(t) \):

\[ I' = \int_0^\infty |\varepsilon(t)| \, dt \]

The use of the latter figure of merit leads to adjustments that correspond to amounts of damping which are adequate for most applications.

Finally, in order to give greater weight to errors occurring late in the transient, i.e., to penalize long-duration transients, it has been proposed to minimize the quantity
This criterion is generally known in the United States as the ITAE (for integral of time-multiplied absolute-value of error) criterion.\(^1\)

**17.4.3. The Necessity for Statistical Considerations.** It is quite apparent that none of the preceding criteria is sufficiently general or satisfactory. The basic reason for this is that fundamentally all inputs to a feedback control system, whether they are control inputs or disturbances, are of a random nature.

a. Consider a feedback control system which has been designed on the basis of the integral-squared-error criterion using a unit-step input. If, now, the input to this feedback control system is very different from that for which it has been designed, it is very likely that the performance of the system will be poor.

Many inputs are random by nature. Therefore, if a feedback control system is to be designed according to the integral-squared-error criterion, it is very important that the criterion be applied to the inputs which occur most frequently. To design a feedback control system, one is thus forced to apply successively the integral-squared-error criterion to various inputs \(E_1, E_2, \ldots\) and to average the results weighted by the probability of occurrence of each input. This is the reason for inputs being characterized by their statistical properties.

b. A second reason for introducing statistical considerations into the design of feedback control systems is the presence of wholly random disturbances which are defined either as noise or thermokinetic incertitude and which determine the limits of the accuracy that may be attained by a mechanism. For example, if the output of an infinitely accurate servo were observed through a microscope, small fluctuations caused by the noise would be seen.\(^2\) These phenomena are, admittedly, of much greater practical importance in electronics; for the latter is a field of extreme precision.

**17.4.4. Statistical Criteria.** Let \(S\) be a control system subjected to a control input \(e(t)\) and a disturbance \(d(t)\). It is desired that the output of the system be as closely as possible \(r(t) = e(t)\), that is, that the error \(\varepsilon(t) = e(t) - r(t)\) be minimum. But difficulties arise when it is desired to specify what is meant by minimizing the difference between two random functions of time. Most often, \(\varepsilon(t)\) is characterized by its mean-square value

\[
\overline{\varepsilon^2} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \varepsilon^2(t) \, dt
\]


\(^2\) Complements, as well as some indications of the usual sources of noise, are given in Sec. 29.1.
the best servo system being that for which $\overline{e^2}$ is the smallest. This is the root-mean-square-error criterion, or briefly the rms-error criterion, first developed by H. M. James, N. B. Nichols, and R. S. Phillips.\footnote{"Theory of Servomechanisms," McGraw-Hill, New York, 1947.}

The error $\varepsilon(t)$ can also be characterized by the integral

$$\overline{\varepsilon} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |\varepsilon(t)| \, dt$$

as was suggested by R. Oldenbourg and H. Sartorius.\footnote{"A Uniform Approach to the Optimum Adjustment of Control Loops," \textit{ASME Paper} 53-A-18, 1954.}

It must be well understood that the expressions $\overline{e^2}$ or $\overline{\varepsilon}$ are expressions that define the accuracy of a feedback control system. The position and velocity errors characterize the accuracy only for very particular inputs. A good definition of the accuracy must take into account the characteristics (frequency spectra) of the random inputs (control input and disturbances such as noise) to the system.

These criteria, especially the rms-error criterion, are very important; a great amount of literature has been devoted to them. We shall describe in Chap. 19 some applications of the rms-error criterion.

Before this is done, it should be stressed that the rms value of the error does not, by any means, characterize the error completely, and therefore use of the rms-error criterion may result in overlooking some very important aspects of the problem. For instance, $\overline{e^2}$ is independent of the distribution of $\varepsilon$ in the frequency spectrum. Now, since a system never operates independently, other systems connected with it will transmit $\varepsilon(t)$ in a manner that depends on their dominating frequencies and on the spectrum of $\varepsilon(t)$.

\textit{Example 1.} Let us consider a multiloop servomechanism, like the one in Fig. 1-45. The error of the internal loop will influence the performance of the over-all system by its frequency spectrum, and not only by its rms value. If the resonance of the over-all system is close to 1 cps, the performance of the system will depend greatly on whether the error of the internal loop has its prevailing frequency around 1 or around 10 cps, even though the rms value is the same in both cases.

\textit{Example 2.} When stabilizing a ship or an airplane intended for transportation of passengers, resonant frequencies of the order of magnitude of the duration of a human pace should be avoided because such frequencies are the most adverse to comfort. Thus, it is just as important to consider the frequency spectrum of the error as it is its rms value.

In other words, the rms criterion is not the complete solution to the problem. One figure alone cannot completely characterize an elaborate system.

17.4.5. \textbf{Applications of the RMS-error Criterion.} The rms-error criterion is the most commonly used and described of the statistical performance criteria. It can be applied in two ways to the design of linear feedback control systems whose inputs (command and disturbances) are characterized by their frequency spectra.
1. If the elements of a feedback control system have been chosen but the values of the parameters are not yet fixed, the rms-error value is calculated for different values of the parameters and the optimum adjustment is chosen. This is the more usual and realistic approach.

2. One can be more ambitious and attempt to determine the *form* of the transfer functions of the system that will minimize the rms error, and not only adjust parameters for a given form of the transfer functions. Such an approach, based on Wiener's ideas, involves a synthesis of the absolute optimum linear system in the sense of the rms-error criterion. Both approaches will be considered in Sec. 19.2.

17.5. PRACTICAL USE OF PERFORMANCE CRITERIA

The criteria which have been described in the preceding sections enable one to compare the performances of various feedback control systems and to choose the one that is best for the requirements which must be satisfied. Actually, these criteria are used twice in the design of feedback control systems:

1. In the initial study, in order that the fundamental elements of the system may be chosen
2. In the choice and the design of the compensating network

For the initial study, accurate criteria are useful (a) to interpret the specifications (speed, accuracy, etc.) in terms of feedback-control-system parameters (bandwidth, open-loop gain, etc.) and (b) to eliminate immediately from the design the elements which would be incompatible with the specifications, for example, power stages for which the response time would be larger than that required for the entire system.

Once the configuration of the system is fixed, the compensating network must be determined. The criteria, mostly the general-purpose ones, are helpful in choosing the network and fixing the values of its parameters. The technique of this operation will be explained in the following two chapters.
CHAPTER 18

COMPENSATION OF FEEDBACK CONTROL SYSTEMS

Summary

2. Lead compensation, or differential control.
3. Lag compensation, or integral control.
4. Combination of lead and lag compensation.
5. Compensating networks in the feedback loop.
6. Other compensation methods. Conclusion.

18.1. CONCEPT OF COMPENSATION. COMPENSATION NETWORKS

18.1.1. Review of the Stability-Accuracy Dilemma. The influence of the open-loop gain on the performance of a servo system has been studied in Chap. 13, and it has been shown that it is, in general, advisable to set the gain to such a value that \( Q = 1.3 \) when \( G(j\omega) \), that is, the frequency-dependent part of the open-loop transfer function, is given. Moreover, it has been shown that the obtained setting is the result of a compromise

![Diagram](image-url)

Fig. 18-1. Noncompensated servo system.

which seldom satisfies both the desired accuracy and stability requirements. As long as \( G(j\omega) \) is given, no improvement can be made; for any modification made would increase the accuracy of the system but would impair its stability, or conversely. An example of this dilemma is shown in Fig. 18-1, which displays the open-loop transfer locus of a feedback control system without integration, for three different settings of the gain.

18.1.2. The Concept of Compensation. The dilemma described above may be solved if it is possible to modify over particular regions the \( G(j\omega) \) transfer locus. Consider an open-loop transfer function \( F \) which corresponds to the full-line locus of Fig. 18-2 and which is coincident with

295
$K_1G$ at high frequencies and with $K_2G$ at low frequencies. This transfer function combines the advantages of $K_1G$ with respect to stability and those of $K_2G$ with respect to accuracy. To compensate a feedback control system is to substitute, then, a more favorable function $F(j\omega)$ for the initial open-loop transfer function $G(j\omega)$. The transfer locus is thus reshaped.

It can be seen that there are an infinite number of methods of reshaping the transfer locus, the two extreme limits consisting of (Fig. 18-3):

1. Starting from $K_2G$ and stabilizing by reshaping the locus in the resonance region.

2. Starting from $K_1G$ and increasing the gain by reshaping the locus in the low-frequency region.

In the first extreme, the phase of $KG$ is increased at each frequency in the zone which has to be reshaped. This is phase-lead compensation.

In the second extreme, an attempt is made to approach the conditions of a feedback control system with integration. At the limit the phase is $-90^\circ$ instead of $0^\circ$ for $\omega = 0$. The phase is decreased, whence the name of lag compensation, or integral compensation.

18.1.3. Compensating Networks. To transform the $G$ function into the $F$ function, an element, the transfer function of which is $F(s)/G(s)$, is inserted into the forward path. This produces the required compensation, an example of which has already been shown in Secs. 14.3.5 and 16.2.4. Such elements are called controllers, or compensating networks. In the forward path of the system, they are usually inserted into the low-power stage immediately following the sensing device and very rarely into the power stage. Occasionally the compensating elements are inserted into the feedback loop. In such a case, the concept of unity feedback may be

---

1 These elements are not inserted into the power stage because they might cause high power loss by dissipation. One of the aspects of cybernetics, the science of the clever control of systems by power amplification, can now be understood: the "intelligent" control of immense masses and powers by small, delicate elements.
retained by the transformation shown in Fig. 18-4. Finally, it may so happen that the compensating elements are inserted into a secondary loop.

So far as the realization of these elements is concerned, they may be either mechanical or electrical and, in addition, either active or passive. The majority of the compensating elements studied in this chapter will be of the passive electrical type. For the design of these compensating networks, two basic techniques are used; they correspond to the extreme cases shown in Fig. 18-3 and are (1) phase-lead compensation and (2) phase-lag compensation. The two techniques will now be studied in detail.

18.2. PHASE-LEAD COMPENSATION, OR DERIVATIVE CONTROL

18.2.1. Principle. Consider (Fig. 18-5) an unstable or marginally stable feedback control system, the open-loop transfer locus of which either lies to the left of the critical point or approaches that point too closely. In order that sufficient stability be obtained, the $KG(j\omega)$ locus must be reshaped in the resonance region in such a way that it remains outside the circle $MO/MA = 1.3$. This can be done by multiplying the $KG(j\omega)$ transfer function by a complex factor $J(j\omega)$ which has a positive phase in the resonance region—whence the expression, *phase-lead compensation*.

The loci shown in Fig. 18-6 can be obtained in the above manner. To do so, it is necessary to use a "differential" (or "derivative") element of the form $J(s) = 1 + Ts$. The transfer locus is a straight line of the type shown in Fig. 18-7, and the phase lead obtained at the frequency $\omega$ is $\arctan T\omega$. Since a pure differentiating element cannot be obtained by the use of a passive network [$J(\infty) = \infty$], use is made of an element with a transfer function approaching that of the differentiating element in the resonance region. It is called a *phase-lead* element, and its transfer function is generally of the form
where \( a \) and \( b \) are small enough to make \( as \) and \( bs^2 \) negligible in the resonance region of the feedback control system. The transfer locus is of the form shown in Fig. 18-8. The only region of this transfer locus that is of interest is that which has been marked by a full line and which is tangent to the half line of pure differentiating control.

Fig. 18-6.

Fig. 18-7. Derivative control.

In conclusion, it may be stated that a phase-lead element consists of a differentiating element in series with a second-order element having a very high natural frequency. The \( bs^2 \) term is very often negligible (see the shape of the locus then in Fig. 18-11). A phase-lead compensating element is then nothing more than a differentiating element with a time constant.

18.2.2. Realization. 1. Electrical Networks. One of the most common electrical networks is that shown in Fig. 18-9. Its transfer function is

\[
J(s) = \frac{R_2}{R_1 + R_2} \frac{1 + R_1Cs}{1 + [R_1R_2/(R_1 + R_2)]Cs}
\]
which can be written as
\[ J(s) = \frac{1 + \alpha \tau s}{\alpha s + 1} \]
where, for convenience,
\[ \tau = \frac{R_1 R_2}{R_1 + R_2} \frac{C}{\alpha} \]
which has the dimensions of a time
\[ a = 1 + \frac{R_1}{R_2} \]

The quantity \( \tau \) is termed the \textit{time constant} of the network and the quantity \( a \) the \textit{phase-lead factor}, or \textit{time-constant ratio}.

In general, an amplifier with a gain \( A \) is added (Fig. 18-10) to obtain \( J(s) = A(1 + \alpha \tau s)/a(1 + \tau s) \). The corresponding transfer locus is a semicircle located in the upper part of the complex plane (Fig. 18-11, for \( A = \alpha \)). The corresponding frequency responses are shown in Fig. 18-12. These figures show that there is a region of positive phase.

2. \textit{Numerical Values}. The value \( \Phi_m \) of the maximum phase lead is given by \( \Phi_m = \arcsin \left[ \left( a - 1 \right)/(a + 1) \right] \). It occurs at the angular frequency \( \omega_m \) given by \( \tau \omega_m = 1/a^{1/2} \). It can be seen that the phase-lead factor \( a \) characterizes the maximum of the phase lead, whereas the constant \( T \) characterizes the frequency at which it occurs. The relationship \( \Phi = f(a, \tau, \omega) \) is graphically displayed in Fig. 18-13, and it is very important that the reader be familiar with it.

Typical values for the lead factor \( a \) range from 3 to 20, the most usual values being 5, 7, and 10. The values of \( \tau \) vary with the problem under consideration, and this will subsequently be discussed.

3. \textit{Mechanical Networks}. An example of a mechanical compensating network is the \textit{gyropole}, a device invented by M. Gianoli and used as a sensing device in missile stabilization. The principle of control by means of the gyropole consists in slaving the position of the missile control surface not to the deviation of a standard gyro, but to the ordinate of a point \( M \) that is made a function of the position of the gyro by means of a spring, pivot, and dashpot, as shown in Fig. 18-14.

The manner of operation is as follows: The ordinate of the point \( N \) lags that of \( P \), which is the deviation of the standard gyro, because of the presence of dashpot fric-
tional forces. Now, because of the pivot, point M has a phase lead with respect to P. More precisely, it can be shown that \( \frac{x}{\alpha} \) is related to \( \alpha \) by a function of the form:

\[
\frac{x}{\alpha} = K \frac{1 + ars}{1 + rs + \tau a^2}
\]

In this relation, \( \tau a^2 \) is negligible for the frequencies which are of interest and \( \tau \) has the dimension of time. The phase-lead coefficient \( \alpha \) is proportional to \((l - l')n\) but the gain \( K \) is proportional to the reciprocal of \((l - l')n\). Thus, in this example, as in

![Amplitude](image1)

![Phase in degrees](image2)

**Fig. 18-12.** Amplitude and phase response of phase-lead controller (e.g., phase-lead network plus amplifier as in Fig. 18-10 with \( A = a \)): \((1 + ars)/(1 + rs)\). Abscissas are \( \tau \omega \).

all electrical lead networks, phase lead results in a loss of gain. The quantities \( a \) and \( \tau \) are set by adjusting the stiffness \( k \) and dashpot \( f \).

The gyropole is, therefore, an element which operates not only as a sensing device but also as a compensating network.

**18.2.3. Physical Interpretation.** In a regular servo the antihunt properties of phase lead can be explained by the very general damping effect of the introduction of a differential term. In fact, during hunting (Fig. 18-15) purely proportional control implies a controlling signal that is positive so long as the error is positive. Thus, because of its inertia, the system would overshoot. On the contrary, if the controller output
involves a term proportional to the time derivative of the error, the controlling signal will decrease and become negative before the system overshoots. This enables thwarting the natural oscillation and preventing overshoot.

Phase lead can thus be considered as prediction in time, for inputs which have more or less a prevailing frequency in the resonance-frequency region of the feedback control system:

\[
r(t + \Delta t) \cong r(t) + \frac{dr}{dt}
\]

\(^1\) See Chap. 12 on frequency spectra.
It can intuitively be said that, to avoid hunting, one must apply compensation sufficiently early. This is, of course, based on the assumption that the response of the system can be predicted for reasonably regular inputs. If the natural frequency were known, prediction would be possible for inputs regular enough to have an error-frequency spectrum with a maximum in the resonance-frequency region of the feedback control system.

A classic illustration of the above considerations is an analysis of the manner in which an experienced person, as compared with one who is inexperienced, drives an automobile. The driving of the inexperienced person consists of a continuous series of oscillations about the desired position. The more carefully he drives, the higher will be the frequencies of the oscillations about the desired position and the smaller the amplitudes of the oscillations. The effect of this is that the inexperienced person hunts for the desired position. The experienced person, on the other hand, takes into consideration the rate at which the desired position is being approached, and the type of control that he then exerts is a combination of position and rate control. He may thus be considered as introducing phase lead into his driving.

18.2.4. Adaptation. Defining a phase-lead controller requires that the two parameters $a$ and $\tau$ be determined. The general-purpose criteria outlined in the preceding chapter are applicable for this. The stability of the system is checked for different values of $a$ and $\tau$, and the open-loop gain is set to provide a resonance ratio of 1.3. The values of the optimum gain $K_{1,3}$ and of natural frequency $\omega_R/2\pi$ are then observed and the performances for the various cases compared. As two parameters must be determined, a study of the type just described would be tedious if
the starting values were unknown. Consequently, outlines of several methods for carrying out this analysis are given below.

**First Method.** This method is used for stabilizing unstable or marginally stable systems. It consists in first determining the desired phase lead \( \Phi_d \) and then, from the equation

\[
\Phi_m = \sin^{-1} \frac{a - 1}{a + 1}
\]

or, still better, from the chart of Fig. 18-13, the minimum value of \( a \) for which \( \Phi_m = \Phi_d \). The quantity \( a \) is then chosen to be a little larger than that just determined, an appropriate value being one that is 20 per cent larger.

To find \( \tau \), the maximum phase lead is assumed to occur at the closed-loop resonance frequency of the feedback control system

\[
\tau \omega_R = \frac{1}{a^{\frac{1}{4}}}
\]

In this manner, an initial set of values \( a \) and \( \tau \) is obtained. In general, the value chosen for \( a \) is not modified and that of \( \tau \) is varied about the first approximation \( \tau_1 \). One must, for example, complete Table 18-1:

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( K_1 )</th>
<th>( \omega_R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2( \tau_1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5( \tau_1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tau_1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2( \tau_1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5( \tau_1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Second Method.** This method is specifically used if it is desired to decrease the effects of unwanted time lags represented by a time constant \( 1/(1 + Ts) \). Let \( KG(s) \) be the open-loop transfer function when time lags are not taken into account and \( KG(s)/(1 + Ts) \) the transfer function when they are. For the resonance frequency of the closed-loop system the time lags decrease the phase margin by the quantity \( ar \), the phase lead replaces \( G(s)/(1 + Ts) \) by

\[
G(s) \frac{1}{1 + Ts} \frac{1 + ar}{1 + s}
\]

\(^1\) The phase lead generally increases this frequency. For a first approximation, it can be considered to be either the natural frequency of the system without phase lead or, better, a frequency that is, say, 30 per cent higher than the natural frequency.

\(^2\) For \( K \) so set that \( Q = 1.3 \), this resonance frequency is not the same for the \( KG \) adjustment as for the \( KG/(1 + Ts) \) adjustment. The resonance frequency corresponding to \( KG \) may be chosen for the first approximation.
If \( a \) and \( \tau \) are so chosen that \( a\tau = T \), this expression becomes

\[
G(s) = \frac{1}{1 + \tau s}
\]

Hence, the time constant \( T \) has been replaced by \( \tau = T/a \); that is, the time lags have been decreased by the factor \( a \). If it is then decided that the unwanted time lags must not decrease the phase margin by more than 5 or 10° (Sec. 17.3.2), \( \tau \) is determined from the condition \( \tau \omega_R < 0.1 \) or 0.2. The quantity \( a \) is then given by \( a = T/\tau \).

Starting with these values of \( a \) and \( \tau \), the best setting can be obtained by using the criteria outlined in the previous chapter.

18.2.5. Limitations of the Phase-lead Compensation. This second method could give the impression that any time lag can be compensated by means of phase lead. Actually, it is limited by the following considerations:

1. The introduction of phase lead results in an increase of the resonant frequency. For example, if phase lead is used with the second-order feedback control system

\[
KG = \frac{K}{s(1 + Ts)}
\]

where \( T = a\tau \), the time scale is changed by the factor \( a \) and the feedback control system becomes \( a \) times faster. By doing this, however, it may happen that the transfer functions which are valid for the usual applications must be replaced by some that are more correct for high frequencies, their denominator being generally of higher order. This is equivalent to the introduction of additional time lags which are not negligible at the new resonant frequency.

2. Very large values must be chosen for \( a \); that is, a large differentiating coefficient is needed if the time lag \( T \) which must be compensated for is very large. Under these conditions, noise considerations become of prime importance because of (a) bandwidth increase (see above) and (b) relative amplification of the parasitic oscillations by differentiation (Fig. 29-11). In practice, one seldom chooses \( a \) greater than 15, and never greater than 20.

3. Phase-lead networks, which can very well compensate for lags of the form \( 1/(1 + Ts) \), are much less effective for compensating inertial lags which are expressed by quadratic factors with a low damping ratio. The reason is that the rapid change of phase from 0 to \(-180°\) cannot be compensated for by a simple lead network with a maximum positive shift of less than 90°.

18.2.6. Example. Let \( K/s(1 + s)(1 + 0.25s) \) be the uncompensated open-loop transfer function of a feedback control system, expressed in terms of the nondimensional Laplace variable.

This expression could be the transfer function of a positional servomechanism with a Ward-Leonard type of control. For it, \( t_1 \) and \( t_2 = 0.25t_1 \) are respectively the mechanical and electrical time constants. Or, conversely, this expression could be the transfer function of a feedback control system with a motor having the transfer
Fig. 18-16. Example of phase-lead compensation.

function \( \frac{1}{s(1+s)(1+0.25s)} \) and an unwanted time lag \( 0.25T \). The results will, therefore, have to be compared with those that would be obtained for \( \frac{1}{s(1+s)} \) without unwanted lag. Figure 18-16 represents the setting of the open-loop gain on the Nichols chart for \( Q = 1.3 \), according to the technique explained in Sec. 13.33:

1. Without a compensating network, it is found that

\[
K_{1,3} = 1 \quad \omega_R = 0.9
\]
2. With a phase-lead network having \( a = 10 \), the locus is

\[
\frac{1}{s + 1 + 10a} \quad \frac{1}{s + 1 + 0.25a} \quad \frac{1}{1 + \tau_a s}
\]

and one finds for various values of \( a \) the values of \( K_{1,2} \) and \( \omega_R \) which are given in Table 18-2 and Fig. 18-17. It is seen from this figure that there is an optimum value of \( \tau \) in the region 0.05 to 0.1 for which the increase in gain is 10 db and the increase in bandwidth is approximately 300 per cent.

<table>
<thead>
<tr>
<th>( a )</th>
<th>( K_{1,2} ) (db)</th>
<th>( \omega_R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>5.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.05</td>
<td>10.5</td>
<td>2.0</td>
</tr>
<tr>
<td>0.07</td>
<td>10.5</td>
<td>2.3</td>
</tr>
<tr>
<td>0.1</td>
<td>9.0</td>
<td>2.5</td>
</tr>
<tr>
<td>0.2</td>
<td>4.0</td>
<td>2.6</td>
</tr>
<tr>
<td>0.35</td>
<td>-2.0</td>
<td>2.6</td>
</tr>
</tbody>
</table>

The approximation \( \tau_0 \omega_R \), \( \Phi \) being a maximum for the uncompensated resonant frequency, yields \( \tau_a = 0.35 \). This value is too large (the compensated resonance frequency is much larger than the uncompensated one). Furthermore, the maximum phase lead must now be used at a frequency that is larger than the resonance frequency. The approximation, which consists in taking \( \tau_a \) equal to the system time lag, yields \( \tau_a = 0.025 \) if the secondary time lag (0.25) is used and \( \tau_a = 0.1 \) if the main time lag (1.0) is used. It is thus seen that this approximation is unsatisfactory if the main lag is taken into account. In general, the optimum value of \( \tau_a \) is smaller than that just calculated. This, however, very much depends on the problem under consideration and can only be solved by means of a thorough investigation.

18.2.7. Phase-lead Compensation in Terms of Root Loci. In terms of the pole-zero configuration, the introduction of a lead network amounts to introducing a new zero and a new pole into the open-loop transfer function, the pole lying \( a \) times farther out than the zero on the negative real axis. As a result of this manipulation the asymptotic directions of the root locus are not changed, but the asymptotes are shifted to the left, which allows the system to continue to be stable for higher values of the open-loop gain.

For example, Fig. 18-18 shows the case of Sec. 18.2.6 after compensation by \( (1 + 0.7s)/(1 + 0.07s) \). This is to be compared with the root locus of the uncompensated system shown in Fig. 14-12. It will be noted that the presence of the zero, \( s = -1.2 \), causes the root locus to bend to the left after its breakaway from the real axis, just as the introduction of a sink tends to bend the streamlines toward it (see Sec. 9.3.7, The Hydraulic Analogy).
18.3. PHASE-LAG COMPENSATION, OR INTEGRAL CONTROL

18.3.1. Principle. Let it be assumed that the $K_{1,3}$ gain of a feedback control system without integration\(^1\) is insufficient for the required accuracy. According to the general theory, the position error can be eliminated if integration is introduced into the open-loop transfer function as an infinite factor for the zero frequency. The transfer locus is now a half line (Fig. 18-19) and the result is pure integral control or, more accurately, phase-lag compensation (Fig. 18-20).

The position error can also be decreased by a factor $b$ if a complex element having a value $b$ at low frequencies and a value 1 at high frequencies is introduced into the system. However, this element must not cause a phase shift in the

\(^1\) This hypothesis is used only to set the problem. Lag compensation may also be used for either eliminating or decreasing the velocity error of a system with one integration.
resonance region of such an amount that the gain has to be decreased. The resulting transfer locus in the useful frequency domain often has the shape shown in Fig. 18-21. This type of control is known as generalized or undercompensated integral control (Fig. 18-22).

**First Remark.** It has been shown (Sec. 18.1.2) that compensation can be realized by decreasing the magnitude in the resonance region. This is simply another way of describing integral control because, with the exception of the over-all open-loop gain, increasing the magnitude at low frequencies is equivalent to decreasing it in the resonance region. With the notations of Sec. 18.1.2, one starts from either the $K_1G$ locus or of $K_2G$ locus.\(^1\)

**Second Remark.** The transfer locus of pure phase lag (half line) or the generalized phase lag (semicircle) is located in the negative phase region of the complex plane, whence the expression phase-lag compensation. This type of compensation is in no way incompatible with phase-lead compensation, because the frequency regions which are affected by the compensations are widely separated: low frequencies for phase lag, resonance frequency for phase lead. Furthermore, in most feedback-control systems the two methods are complementary. This will subsequently be explained. In Fig. 18-23a and $b$ the resonant-frequency region of the feedback control system has been darkened on the transfer loci of the phase-lead and phase-lag controllers.

\(^1\) The type of compensation that results in a decrease of magnitude can be considered to be of either the phase-lead or phase-lag type. If the $\omega$ graduation is not specified, nothing can be stated.
18.3.2. Realization. One of the most usual realizations is that shown in Fig. 18-24. The transfer function is

\[ J(s) = \frac{1 + R_2C_s}{1 + (R_1 + R_2)C_s} \]

It can be written as

\[ J(s) = \frac{1 + \tau s}{1 + b\tau s} \]

with \( \tau = R_2C \quad b = 1 + \frac{R_1}{R_2} \)

The corresponding transfer locus is a semi-circle. It is tangent, in its low-frequency region, to the locus \( 1 + 1/(b - 1)\tau s \), which is the transfer locus of a pure integrals controller. The frequency-response curves are sketched in Fig. 18-26.

Phase-lag compensation is sometimes realized as follows: An \( RC \) network with a transfer function \( 1/(RC_s + 1) \) (Fig. 18-27) is introduced into the feedback loop of a regenerative feedback amplifier \( A \), as shown in Fig. 18-28. One then has

\[ \frac{X}{E}(s) = A \frac{1 + RC_s}{1 - Ah + RC_s} \]

If \( A \) and \( h \) are so chosen that \( Ah = 1 \), pure integral control is realized:

\[ \frac{X}{E}(s) = A \left(1 + \frac{1}{RC_s}\right) \]

As a matter of fact, this condition corresponds to a critical setting. If \( Ah \neq 1 \), the transfer function has no more integration. But, as long as the value of \( Ah \) is close to unity, the magnitude is large at low frequencies. It can easily be verified that under-compensated integral control as defined above is obtained for a setting such that \( Ah < 1 \). Conversely, for \( Ah > 1 \) one obtains what is sometimes called overcompensated integral control. This is seldom used, because it produces open-loop instability.
18.3.3. Practical Adaptation. Consider the example already used for the study of phase-lead compensation and consider the manner in which it can now be improved by use of integral control. The loci have been drawn in Fig. 18-30 for \( b = 10 \), a typical value, and the values \( \tau_i = 0, 1, 3, \) and \( 5 \). It is seen that, for \( \tau_i \) smaller than \( 5 \), the solution is of no interest, as \( K \) and \( \omega_n \) become smaller as \( \tau_i \) increases. Another solution arises for \( \tau_i \) larger than \( 5 \). An adaptation is then quickly obtained; it corresponds to \( K = 20 \) db, \( \omega_K = 0.9 \). If \( \tau_i \) is further increased, \( K \) and \( \omega_R \) do not appreciably change, but the phase margin increases.

Because it is undesirable to choose \( \tau_i \) too large, since large time constants are difficult to realize, especially with electrical networks, \( \tau_i \) is taken equal to \( 7 \). The initial \( K_{1.3} \) being \( 0 \) db, the integral control has boosted the gain, i.e., the velocity constant, by \( 10 \).

This result is valid for regular feedback control systems which are compensated by uncompensated integral control, as \( (1 + \tau_i \omega) / (1 + \tau_i \omega) \). If \( \omega_K \tau_i \) is large enough—that is, if the compensation network can be considered as a pure integral controller—the network does not modify the resonance frequency and multiplies the open-loop gain \( K_{1.3} \) by \( b \).
Fig. 18-30. Example of integral compensation.

\[ G(s) = \frac{1}{s(1+s)(1+0.25s)} \frac{1+\tau_i s}{1+10\tau_i s} \]

for \( \tau_i = 0; 1; 3; 5 \). Loci are scaled in rad/sec.
The choice of the lower limit of $\tau_i$ results from the consideration of the undesirable secondary effect of the phase shift, which is approximately equal to $(\pi/2) - \arctan \tau_i \omega_R$ for the considered frequency region (darkened in Fig. 18-31). If, for example, the phase shift for the resonance angular frequency $\omega_R$ is assigned to be less than, or equal to, $\varphi$ radians, one should take $\tau_i \geq 1/\omega_R \varphi$.

**18.3.4. Physical Interpretation.**

The phase-lag controller may be considered as a low-pass filter controlling the power stage. It minimizes the effect of small deviations that fluctuate rapidly about a zero average value, and it especially takes into account the permanent deviations, or systematic errors, because of the amplification at low frequencies. This explains its effect on the position and the velocity errors.

**18.4. COMBINATION OF LEAD AND LAG COMPENSATION**

**18.4.1. Interest and Principle.**

By considering the practical setting of the phase-lead and phase-lag networks, it is seen that, near resonance, the undesirable secondary effect of each network is in opposition to the useful effect of the other. Thus, the undesirable secondary effect of the integral control is a phase shift which is opposite to the main effect of the phase-lead network. It is therefore understandable why the combination of the two networks is highly favorable, provided both are properly set.

From the previous example (loci sketched for $\tau_a = 0.1$, $\tau_i = 3$, and $b = a = 10$) it can be seen (Fig. 18-32) that (a) $\tau_i = 3$ without phase lead is not satisfactory and (b) $\tau_i = 3$ combined with a $\tau_a = 0.1$ phase lead yields $K_{1,1} = 29$ db, $\omega_R = 2.5$. It is seen that $K_{1,1}$ is improved because of the integration and $\omega_R$ because of the phase lead.

**18.4.2. Synthesis of Two Cascaded Networks.**

In this section, reference is made to Chap. 10, where it has been shown that the product of two transfer functions computed for an open output circuit is not correctly represented by two cascaded dipoles. The approximation is valid only if the input impedance of the second network is large with respect to the output impedance of the first one. Thus, with the two networks shown in Fig. 18-33 one has for the lead network the output impedance $Z = R_2$, and for the lag network the input impedance

$$Z_1' = R_1' + R_2' + \frac{1}{C's}$$

One can choose for example, $R_1' + R_2' > 10R_2$. Furthermore, the load impedance of the second network must be larger than $Z_0' = R_2' + 1/C's$.

† As the respective values of the time constants already impose $R_1' + R_2' > R_2$, it is advantageous to put the lead network before the lag network, as shown in Fig. 18-33. See, however, Sec. 29.1.7, Example 1.
Fig. 18-32. Example of combination of phase-lead and phase-lag compensation.
Other solutions are theoretically possible. Thus, for example, an electronic amplifier, a buffer amplifier, can be inserted between the two networks. Synthesis of the two cascaded networks can be performed by using the method outlined in Sec. 10.2.2.

18.4.3. Other Networks. Numerous other methods are possible. For example, the network shown in Fig. 18-34 can be applied directly. At low frequencies it behaves like a lag network, and at high frequencies like a lead network:

$$\frac{R}{E}(s) = \frac{(1 + R_1C_1s)(1 + R_2C_2s)}{(1 + R_1C_1s)(1 + R_2C_2s) + R_1C_2s}$$

This makes it particularly interesting for many applications, provided the constants have been properly adjusted.

18.5. GENERALIZATION

18.5.1. More Complex Compensating Networks in the Forward Path. More elaborate networks can be inserted in the forward path in order to solve special problems. The problem of synthesizing such elaborate networks having a prescribed transfer function is a classical problem of circuit synthesis. Note that the use of active networks incorporating electronic amplifiers often makes the solution of such problems much easier.

18.5.2. Compensating Networks in Feedback Paths. Compensating networks can be introduced into feedback paths (Fig. 18-35). The most important case is that in which a derivative term is introduced into a feedback path, thus making the feedback transfer function of the form

\[ F(s) = 1 + \lambda s \]

The most typical application of this technique is the well-known tachometric feedback. The derivative term is generated by a tachometric dynamo, specially devised to produce a voltage proportional to angular velocity (Sec. 29.4.1). The stabilizing effect of a tachometric feedback can be shown by applying it to a second-order servo (Sec. 13.1.2). The uncompensated open-loop and closed-loop transfer functions are respectively

\[ KG = \frac{C}{Js^2 + f} \quad H = \frac{1}{1 + (f/C)s + (J/C)s^2} \]

After tachometric feedback has been introduced, the system can be considered as one with two superposed loops (Fig. 18-36). The resulting transfer function is

\[ H = \frac{KG}{1 + (1 + \lambda s)KG} = \frac{1}{1 + (f/C + \lambda)s + (J/C)s^2} \]

which shows the damping effect of the \( \lambda \) term. This effect is somewhat similar to that of a cascade-type lead controller in the forward path. More generally the introduction of two derivatives in the feedback loop \( F = 1 + \lambda s + \mu s^2 \) gives

\[ H = \frac{1}{1 + (f/C + \lambda)s + (J/C + \mu)s^2} \]

Both the damping and the natural frequency are modified.

Compensating networks in feedback paths can be studied by taking into account the equivalence of the two block diagrams shown in Fig. 18-4. This makes their study and design identical with that of conventional compensating networks in the forward path. The interest of inserting
compensating networks in the feedback loop can be briefly discussed as follows:

a. From the energy viewpoint, the idea of inserting compensating networks in the feedback path is, generally speaking, excellent, because the corresponding energy is taken from the power stage and, therefore, the compensating network is easier to design for a given purpose than if it were passive.

b. From the control viewpoint, on the contrary, any imperfections of such network are most adverse to the system's accuracy, since they have an immediate influence on the crucial operation of error sensing without any possibility of being compensated by the existence of feedback.

18.5.3. Other Types of Compensating Networks. Other types of compensating networks have been devised. They cannot all be discussed here. We shall mention only two of them and refer the reader to bibliographical references for others.

First, the idea of placing the compensating network at the output of the system, in the form of friction dampers mounted on the output shaft, has been conceived. This arrangement has the obvious disadvantage of power consumption, but nonnegligible advantages can be cited.¹

Second, it has been suggested that compensation be effected by introducing into the forward path a function of time that leads not the error but the command itself. This is known as introducing an anticipating path. The example shown in Fig. 18-37 corresponds to the equation

\[ W + FE + FR \]

and is equivalent to the block diagram shown in Fig. 18-38. The introduction of a tachometric anticipating path can be equivalent to that of pure derivative control.² When the control input can be foreseen, the use of anticipating paths enables one to

¹ Prob. 31.
² Naslin, op. cit., p. 175.
eliminate the error of a linear servo system completely. Unfortunately, this method is not applicable in the presence of random inputs.

To sum up, the possible varieties of compensating networks are innumerable. Particular types can present advantages for specific types of problems. However, the fundamental ideas are quite uniform. That is why the authors considered it preferable, when writing the present chapter, to present a thorough discussion of the fundamental concepts, rather than a collection of more or less unrelated technical notes.
CHAPTER 19

SERVO-SYSTEM SYNTHESIS

Summary

2. Statistical approach.
3. Conclusion.

It has been shown in the preceding chapter how servo systems can be compensated; that is, if some components are given, compensating networks can be designed in order to improve the performance of the system so that it will meet the specifications stipulated. The method has been outlined with emphasis being placed on the two fundamental ideas of lead and integral compensation. It is, in practice, extremely satisfactory. The designer trained in this technique soon becomes so familiar with it that he can "think in the KG plane." Further advantage of this approach is the facility given by graphical techniques (Chap. 8) and the ability to treat at the same time and in the same manner systems that are characterized by their differential equations as well as systems that are determined from experimental data (Secs. 8.4 and 8.6).

Nevertheless, attempts at still more systematic approaches have been made in two directions. First, an attempt has been made to start the design procedure by specifying the requirements and then conducting the design in a straightforward manner in order to force the system to meet the requirements. This is somewhat different from the compensating procedure, outlined in Chap. 18, which consists in applying usually successful techniques to a given system and adjusting the parameters in order to obtain satisfactory performance. Such an approach, based on considerations of pole-zero configuration, was first presented by J. G. Truxal\(^1\) as an application of Guillemin's procedure for filter synthesis.

Second, attempts have been made to produce design methods for servo systems consistent with the fact that inputs in automatic-control problems are of essentially random nature. As a consequence, any rational design procedure should start from the statistical properties of the inputs (command and disturbances). This differs from the previous techniques in that the problem is taken from its own physical origin, the existence of a command that is to be executed in spite of unwanted disturbances also present. It thus enables one to solve the problem specifically presented, and not only to design a system that is satisfactory for certain specific input conditions. In particular, such an approach makes possible an

attempt at optimizing the system, i.e., designing the best possible system for the job. Such an approach, first used for radar problems, was systematized by James, Nichols, and Phillips and was generalized by some authors (G. Newton, M. Pélégrin) who applied the ideas of N. Wiener and Y. W. Lee to feedback control systems.

It is the purpose of this chapter to outline these two different trends and then to proceed to a brief discussion outlining the different methods of design for linear feedback control systems.

19.1. SYNTHESIS BY MEANS OF POLE-ZERO CONFIGURATION

19.1.1. General Outline of the Method. When a servo system is to be designed, the specifications generally concern the transfer function of the closed-loop system. Therefore, it is natural to start by building a system with a transfer function $H(s)$ that will meet the specifications of the problem, and then design the compensating network that will make the closed-loop function equal to $H(s)$. The steps involved in such a procedure are:

1. Find a desired $H(s)$ that enables the system to meet the specifications. The desired open-loop transfer function is then

$$KG(s) = \frac{H}{1 - H}$$

2. Design the compensating network that will make the open-loop transfer function equal to $KG$. If the open-loop transfer function of the noncompensated system is $K_0G_0(s)$, the transfer function of the compensating network is $KG/K_0G_0$.

19.1.2. Finding a Suitable Closed-loop System Function: First Approximation. As explained in Sec. 17.1.2, the specifications for $H(s)$ generally involve conditions concerning (a) the steady state; and (b) the transient. Conditions (a) specify maximum permissible values for position, velocity, and sometimes acceleration errors; conditions (b) specify that the transient must be sufficiently fast and well damped.

As a first approximation, the system that is to be designed may be represented by its dominant mode (Sec. 9.3.3). In other words, the first attempt can be to meet the specifications by means of a second-order system. It should be remembered that a second-order system with a transfer function

$$H_2(s) = \frac{k}{s^2 + 2\xi \omega_n s + \omega_n^2} = \frac{K}{1 + (2\xi / \omega_n)s + s^2 / \omega_n^2}$$

has a zero position error if its static gain $K$ is unity. Furthermore, its velocity and acceleration constants are given by

\[ \frac{1}{C_v} = \frac{2\xi}{\omega_n} \quad \frac{1}{C_a} = \frac{1 - 4\xi^2}{\omega_n^2} \]

So far as the transient is concerned, the speed of response is characterized by the natural frequency $\omega_p = \omega_n(1 - \xi^2)^{\frac{2}{3}}$ (Fig. 6-17) or the response time (Fig. 6-21); and the amount of damping is characterized by the damping ratio $\xi$, which is directly related to the overshoot (Fig. 6-16). These relations enable one to choose the parameters $k$, $\xi$, and $\omega_n$ to meet the specifications. However, it often occurs that it is not possible to find such a set of values for $k$, $\xi$, and $\omega_n$.

**Example.** Let the specifications be (a) for the steady state, zero position error and a velocity constant of at least 3 sec⁻¹; (b) for the transient, an overshoot that should not exceed 20 per cent and a natural frequency that should be about 2.5 to 3 rad/sec.

The conditions for the steady state require

\[ K = 1 \quad k = \omega_n^2 \quad \frac{2\xi}{\omega_n} \leq 0.33 \]

The conditions for the transient require

\[ \xi \geq 0.45 \quad (\text{from Fig. 6-16}) \]
\[ 2.5 < \omega_n(1 - \xi^2)^{\frac{2}{3}} < 3.0 \]

If one chooses, for example (Fig. 19-1),

\[ \xi = 0.5 \quad \omega_n = 3.2 \text{ rad/sec} \quad k = \omega_n^2 = 10.2 \text{ sec}^{-1} \]

it is seen that the specifications are met, since

\[ C_v = \frac{\omega_n}{2\xi} = 3.2 \text{ sec}^{-1} \]

**19.1.3. Second Approximation for the System Function.** When it is impossible to find $k$, $\xi$, and $\omega_n$, it means that no second-order system will meet the specifications. This is usually due to incompatibility between the steady-state specifications, essentially a large velocity constant, and the transient specifications, especially a reasonable damping.

The procedure suggested by J. Truxal then consists in multiplying $H_2(s)$ by a complementary function $H_c(s)$ the zeros and poles of which are systematically chosen in order for the transfer function

\[ H(s) = H_2(s)H_c(s) = H_2(s) \frac{(s - z_1)(s - z_2) \cdots}{(s - p_1)(s - p_2) \cdots} \]

![Fig. 19-1. Second-order system (with $\omega_n = 3.2$, $\xi = 0.5$).](image-url)
to meet the specifications. For this purpose, two basic ideas can serve as helpful guides. One is dipole compensation, the second is compensation by addition of zeros. In terms of the open-loop transfer function, these ideas correspond to integral and lead compensation, respectively.

1. Compensation by Addition of a Zero. Let

\[ H_c(s) = s - z_c \]

The velocity constant of the compensated system is given by Eq. (9-4) as

\[ \frac{1}{C_v} = \frac{2\tau}{\omega_n} + \frac{1}{z_c} \]

which shows that the addition of the negative real zero \( z_c \) results in an increase of \( C_v \).

In practice, \( H_c(s) \) also involves poles

\[ H_c(s) = \frac{s - z_c}{p_c} \quad \text{or} \quad H_c(s) = \frac{s - z_c}{(s - p_c')(s - p_c'')} \]

Generally, \(|p_c'|\) and \(|p_c''|\) are chosen sufficiently large for their influence on the transient performance to be negligible, so that the effect of introducing \( H_c(s) \) is essentially to increase the velocity constant.

**Example.** Let the specifications be (a) for the steady state, zero position error and a velocity constant of at least 5 sec\(^{-1}\); (b) for the transient, overshoot < 15 per cent, \( \omega_n = 2 \) to 3 rad/sec.

These specifications cannot be met by a second-order transfer function, since

\[ \frac{\omega_n}{2\tau} > 5 \quad \zeta > 0.5 \quad 2 < \omega_n(1 - \zeta^2)^{1/2} < 3 \]

are not compatible.

Let the configuration of Fig. 19-2 be chosen for \( H_2(s) \)

![Fig. 19-2. Compensation by addition of a zero. (Not drawn to scale.)](image)

and find

\[ H_r = \frac{s - z_c}{(s - p_c')(s - p_c'')} \]

The velocity constant of \( H_2(s) \) was given by

\[ \frac{1}{C_v} = \frac{2\tau}{\omega_n} = 0.4 \ \text{sec} \]

The velocity constant of the new system \( H_2(s)H_c(s) \) is given by

\[ \frac{1}{C'} = 0.4 + \frac{1}{z_c} - \frac{1}{p_c'} - \frac{1}{p_c''} \]

The condition \( \frac{1}{C'} < 0.2 \) can then be met by choosing (Fig. 19-2):

\[ \frac{1}{z_c} = -0.4 \quad z_c = -2.5 \]
which would provide zero velocity error in the absence of \( p'_c \) and \( p''_c \), and

\[
\left| \frac{1}{p'_c} + \frac{1}{p''_c} \right| < 0.2
\]

for example, \( |p'_c| > 10 \) \( |p''_c| > 10 \)

If \( |p'_c| \) is taken as 10 and \( |p''_c| \) as 30 in order that they be sufficiently separated, the specifications will be met by

\[
H(s) = \frac{k}{s^2 + 3.6s + 9(s + 10)(s + 30)} \quad s + 2.5
\]

where \( k \) is chosen to make the static gain unity: \( 2.5k = 9 \times 10 \times 30 \), whence

\[
H(s) = \frac{1,080(s + 2.5)}{(s^2 + 3.6s + 9)(s + 10)(s + 30)}
\]

2. **Compensation by Addition of a Dipole.** A greater velocity constant can also be attained, without appreciably affecting the transient response, by choosing

\[
H_c(s) = \frac{s - z_c}{s - p_c}
\]

where the zero \( z_c \) and the pole \( p_c \) lie close together on the negative real axis near the origin. The above manipulation is said to add a dipole to \( H_2(s) \) (Fig. 19-3). In fact, the new velocity constant will be given by

\[
\frac{1}{C_v} = \frac{2T}{\omega_n} + \frac{1}{z_c} - \frac{1}{p_c}
\]

![Diagram of a complex plane with points labeled as \( p_c \), \( z_c \), and \( H(s) \) with the real and imaginary axes marked.](image)

Fig. 19-3. Compensation by addition of a dipole.

If \( |z_c| \) and \( |p_c| \) are chosen small enough, a noticeable increase in \( C_v \) can be achieved. So far as the transient is concerned, the essential modification consists of the addition of the term \( A \exp(p_c t) \), which is very small if \( Z_c \) and \( P_c \) lie close to each other, since the modulus of \( Z_c P_c \) is in the numerator of the expression of \( A \) (see Secs. 9.3.1 and 9.3.5, Example 2).

**Example.** Let the specifications be the same as those of par. 1. If a dipole is added with the zero lying at \( z_c = -0.5 \), the inverse velocity constant will be

\[
\frac{1}{C_v} = \frac{2T}{\omega_n} + \frac{1}{z_c} - \frac{1}{p_c} = 0.4 - 2 - \frac{1}{p_c}
\]

Thus, the requirement that \( C_v \) be equal to 5 will be met if \( p_c = -0.55 \). Acceptable values for \( p_c \) and \( z_c \) are

\[
z_c = -0.50 \quad p_c = -0.53
\]
3. **Generalization.** More generally, the complementary function \( H_c(s) \) can involve different zeros and poles in the left half plane, which may be considered as complex. A typical pattern is shown in Fig. 19-4. The location of the additional zeros and poles depends on the problem under consideration. For a given problem, different configurations may produce approximately the same result. The concepts of compensation by the addition of a zero or a dipole can often serve as a useful guide; it was seen in Chap. 18 in exactly the same manner that an infinite number of compensating networks is conceivable, the basic ideas of lead and integral compensation often serving as a useful guide.

**19.1.4. Finding the Open-loop Transfer Function.** The next step in the procedure consists in finding the open-loop transfer function \( KG \). Its expression is readily given by

\[
H = \frac{KG}{1 + KG} \quad KG = \frac{H}{1 - H}
\]

but its poles and zeros are not so easily obtained.

Let \( H(s) \) be written as \( P(s)/Q(s) \), where \( P \) and \( Q \) are polynomials in \( s \). Then

\[
KG(s) = \frac{P(s)}{Q(s) - P(s)}
\]

The zeros of \( KG(s) \) are the zeros of \( H(s) \), and the poles of \( KG(s) \) are the roots of the equation

\[
Q(s) - P(s) = 0
\]

The degree of this equation is the degree \( n \) of the polynomial \( Q(s) \); so that the problem of finding its roots may appear to be a difficult one, especially if the roots are complex. Fortunately, in practice, it can be somewhat simplified. First, most servo systems comply with the zero-position-error condition. Therefore (Sec. 15.1), the open-loop transfer function has integration; that is, \( s = 0 \) is a root of the equation \( Q = P \). As a consequence, the number of poles to be found is only \( n - 1 \). Second, it is possible to find the real roots of the equation \( Q(s) = P(s) \) graphically, by plotting the curves \( Q(\alpha) \) and \( P(\alpha) \) as functions of the real quantity \( \alpha \) and finding their intersections. Thus, the number of roots to be found analytically is reduced by the number of real roots.

It is possible to go even further, and specify that all roots of \( Q = P \) must be real and negative. Such a condition is not only helpful for finding the roots graphically, but it is also advisable from the viewpoint of the realizability of the transfer functions under consideration. If this
condition is fulfilled, all the poles of \( KG(s) \) can be obtained by finding the intersections of the \( Q(\alpha) \) and \( P(\alpha) \) curves.

In practice, the \( P(\alpha) \) curve is simple, since only rarely does \( H(s) \) have a large number of zeros. It is a horizontal line if no zero is present, a straight line if one zero is present, and a parabola if two are present. The \( Q(\alpha) \) curve is of higher degree, in practical problems often being a quartic. Its plotting is facilitated by knowing such zeros as were determined in the first step of the procedure. The zeros—i.e., the poles of \( H(s) \)—can, if necessary, be readjusted at the present stage of the procedure in order to guarantee that all intersections be real or, more simply, to facilitate the graphical constructions.

![Diagram](image.png)

Fig. 19-5. Determining the poles of \( KG \) for \( H(s) \):

\[
\frac{1,080(s + 2.5)}{(s^2 + 3.6s + 9)(s + 10)(s + 30)}
\]

**Example.** Consider the first example outlined in Sec. 19.1.3. A satisfactory system function was found to be

\[
P(s) = \frac{1,080(s + 2.5)}{(s^2 + 3.6s + 9)(s + 10)(s + 30)}
\]

\[
Q(s) = (s^2 + 3.6s + 9)(s + 10)(s + 30)
\]

A sketch of \( P(\alpha) \) and \( Q(\alpha) \) (Fig. 19-5) shows that the straight line \( P(\alpha) \) intersects the quartic \( Q(\alpha) \) four times, including the \( \alpha = 0 \) intersection. Therefore, all the poles of \( KG(s) \) are real and can be found graphically. If this were not the case, the values of the poles of \( H_d(s) \), \( p'_e = -10 \) and \( p''_e = -30 \), could be readjusted while still satisfying the condition

\[
\begin{vmatrix}
1 & 1 & 1 \\
\end{vmatrix} < 0.2
\]

in order to assure the existence of four real intersections.

A more accurate plot or an analytical determination shows that the intersections lie at \( s = 0, -0.85, -15.4, \) and \(-27.0 \). Thus

\[
Q(s) - P(s) = s(s + 0.85)(s + 15.4)(s + 27)
\]

and the open-loop transfer function is

\[
KG(s) = \frac{1,080(s + 2.5)}{s(s + 0.85)(s + 15.4)(s + 27)}
\]
19.1.5. Designing the Compensating Network. At this stage of the procedure, an expression for \( KG(s) \) that will enable the system to meet the specification has been found. The problem now consists in multiplying the open-loop transfer function of the noncompensated system, \( K_0G_0(s) \) by the transfer function \( J(s) \) of a suitable compensating network in order that

\[
K_0G_0(s)J(s) = KG(s)
\]

The transfer function of the compensating network is thus

\[
J(s) = \frac{KG(s)}{K_0G_0(s)}
\]

which shows clearly the essence of the method: Replace the transfer function of the noncompensated system \( K_0G_0(s) \) by the desired transfer function \( KG(s) \) obtained directly from the specifications.

The last step is to design a network with the transfer function \( J(s) \) and adjust it for impedance matching (Sec. 10.2).

Example. Consider now the example outlined in Secs. 19.1.3 and 19.1.4. A desirable open-loop transfer function has been found to be

\[
KG(s) = \frac{1,080(s + 2.5)}{s(s + 0.85)(s + 15.4)(s + 27)}
\]

In addition, assume that the noncompensated system is the one considered in the previous chapter, where integral and lead compensation were applied to it (see Secs. 18.2.6 and 18.3.3):

\[
K_0G_0(s) = \frac{K_0}{s(s + s)(1 + 0.25s)} = \frac{k_0}{s(s + 1)(s + 4)}
\]

The transfer function of the required compensating network is

\[
J(s) = \frac{KG(s)}{K_0G_0(s)} = \frac{1,080}{k_0} \frac{(s + 2.5)(s + 1)(s + 4)}{(s + 0.85)(s + 15.4)(s + 27)}
\]

It is the product of three homographic functions in \( s \), one of the lead type and two of the integral type

\[
\frac{s + 2.5}{s + 0.85} \quad \frac{s + 1}{s + 15.4} \quad \frac{s + 4}{s + 27}
\]

The problem of synthesizing such a network has been developed in Sec. 10.2.2.

Note. In this example the procedure can be somewhat simplified by taking into account the fact that the pole \( s = -0.85 \) added by the compensating network is not very different from the pole \( s = -1 \) already present in the noncompensated system. The procedure and the design would be simplified if the pole \( s = -1 \) were incorporated into the open-loop transfer function. This can be done by readjusting \( Q(s) \) so that the curves \( Q(\alpha) \) and \( P(\alpha) \) will intersect at \( \alpha = -1 \).

Recall that \( Q(s) = (s^2 + 3.6s + 9)(s + 10)(s + 30) \) and \( P(s) = 1,080(s + 2.5) \). In the present case it is easier to readjust the damping ratio of the quadratic factor then the real poles. If \( \xi \) is that damping ratio, then

\[
Q(s) = (s^2 + 6\xi s + 9)(s + 10)(s + 30)
\]
Equating \( P(-1) \) and \( Q(-1) \) yields \( \zeta = 0.63 \). As far as the performance of the system is concerned, this value is practically equivalent to the previously chosen value of \( \zeta = 0.60 \). The next steps are as follows: The equation \( Q(s) = P(s) \) becomes

\[
(s^2 + 3.79s + 9)(s + 10)(s + 30) = 1,080(s + 2.5)
\]

Two of its roots are known \( (s = 0, -1) \); the other two are most readily found by equating to zero the remaining quadratic factor

\[
s = -8.2 \quad s = -34.8
\]

Hence,

\[
K(s) = \frac{1,080(s + 2.5)}{s(s + 1)(s + 14.8)(s + 28.2)} \quad J(s) = \frac{1,080 s + 2.5}{K_0 s + 14.8 s + 28.2}
\]

This time the transfer function of the compensating network is the product of only two homographic functions.

The network can be synthesized as outlined in Sec. 10.2.2 by a cascade of two double lattice networks whose real characteristic impedance is to be made equal to the load impedance.

**Practical Application.** The Guillemin and Truxal approach is a very elegant and powerful tool for the designer. It can be used for a great number of problems, provided all the fixed components in the system are characterized by their differential equations. At first sight, one might think that any kind of specification could be met by starting from any fixed components.\(^1\) In practice, the synthesis procedure has to take physical limitations into account. If one is led to unrealizable or unrealistic compensating-network characteristics, it is advisable to start again from the beginning.

Thus the practical synthesis procedure is not as straightforward as the basic theory. It has been seen, in the very simple example chosen, how readjustments of the prescribed \( H(s) \) can occur in the course of the design, and in fact readjustments are the rule rather than the exception. This is, moreover, true of all technical fields. That is, all methods of design, however straightforward their idea, involve in practice a process of successive approximations.

**19.2. STATISTICAL APPROACH TO SERVO-SYSTEM SYNTHESIS**

**19.2.1. Statistical Criteria.** The reason for interest in statistical criteria has already been pointed out: Usual specifications and criteria concern the response of the control system to typical inputs, whereas feedback control systems are actually subjected to random inputs. Therefore, a statistical approach is the only consistent one if it is wished to evaluate the actual operation of the system.

Among statistical criteria, the most commonly used is the rms-error criterion, which consists in minimizing

\[
\bar{e}^2 = \lim_{T \to \infty} \frac{1}{T} \int_{-T}^{T} e^2(t) \, dt
\]

\(^1\) Similarly, the theory of phase-lead compensation gives the impression that any lag can be compensated for. The actual limitations for this are discussed in Sec. 18.2.5.
As pointed out in Sec. 17.4.4, $\bar{\varepsilon}^2$ is a figure of merit that does not completely characterize the performance of the system. In spite of this, a great deal of interest has been shown in the rms-error criterion. One reason, besides the actual interest of the criterion, may be the fact that the rms criterion lends itself easily to calculations.

**19.2.2. Computation of the RMS Error.** Random inputs are characterized by their frequency spectrum, generally obtained from experimental data by means of the autocorrelation function (Sec. 12.2.5). It has been shown in Chap. 12 that, if $H(s)$ is the error-transfer function of a system subjected to a random input with a frequency spectrum $\Phi(\omega)$, the mean-square value of $\varepsilon(t)$ can be written

$$
\bar{\varepsilon}^2(t) = \int_{-\infty}^{+\infty} \Phi(\omega)|H(j\omega)|^2 d\omega
$$

(19-1)

In a more general way, for two uncorrelated inputs with spectra $\Phi_1$ and $\Phi_2$ one has

$$
\varepsilon(s) = H_1(s)E_1(s) + H_2(s)E_2(s)
$$

and

$$
\bar{\varepsilon}^2(t) = \int_{-\infty}^{+\infty} (\Phi_1|H_1|^2 + \Phi_2|H_2|^2) d\omega
$$

(19-2)

In case of a linear servo system with unity feedback, if $e_1(t)$ is the control input and $e_2(t)$ a disturbance (Fig. 19-6), the error-command transfer function is

$$
H_1(s) = \frac{\varepsilon(s)}{E_1(s)} = \frac{1}{1 + KG(s)}
$$

The error-disturbance transfer function of the system $H_2(s)$ depends on how the disturbance is introduced into the system. For the particular case where the block diagram of the system is that of Fig. 19-7, it can be written as

$$
H_2(s) = \frac{\varepsilon(s)}{E_2(s)} = \frac{-G(s)}{1 + F(s)G(s)}
$$

Hence the complete expression of $\varepsilon(s)$ is

$$
\varepsilon(s) = \frac{1}{1 + FG} E_1(s) - \frac{G}{1 + FG} E_2(s)
$$

(19-3)

**Note.** It is important to understand why the rms-error criterion represents a method of optimizing a servo system for different inputs, a
greater weight being given to the more probable inputs as was stated in Sec. 17.4.2. In fact, Eq. (19-1) expresses a weighting of $H(j\omega)$ by $\Phi(\omega)$.

19.2.3. Practical Application of the RMS-error Criterion. If the frequency spectra of inputs $E_1$ and $E_2$ are known, $\varepsilon(s)$ is given by Eq. (19-3), Eq. (19-2) yielding the mean-square error, which is the performance criterion of the system. In practice, this expression is obtained by means of integrations in the complex plane. The resulting expression will be found in the work of James, Nichols, and Phillips.\footnote{Op. cit., pp. 369–370.}

The criterion is usually applied when all the elements of the feedback control system, except for the values of yet undetermined parameters, are chosen. Mathematically, this means that the $H_1$ and $H_2$ functions are determined except for the values of some parameters which can be chosen to give a minimum mean-square error.

Remark with Regard to the Regulator Case. If there is no information about the frequency spectrum of the possible disturbances, the mean-square-error criterion may be applied by assuming that they are uniformly distributed along the frequency spectrum:

$$
\Phi_2(\omega) = \text{const} \quad \text{(white noise)}
$$

Thus, the function to minimize is

$$
\int_0^\infty |H_2(j\omega)|^2 \, d\omega
$$

This can easily be done with the use of a planimeter. As a very rough approximation, the method may even be considered equivalent to the minimization of $|H_2(j\omega)|_{\text{max}}$, as all the $|H_2(j\omega)|$ curves have rather similar shapes, being equal to zero for $\omega = 0$ and $\omega = \infty$.

This result can be directly found by means of the following physical reasoning: The very slow disturbances need not be considered, for the regulator has sufficient time in which to react. This is expressed by $H_2(0) = 0$. As to the very rapid disturbances, they need not be considered either, for the inertia of the system that is being controlled filters them out even in the absence of regulation. This is expressed by $H_2(\infty) = 0$. One must, therefore, pay the greatest attention to the disturbances which may cause a dangerously large $H_2(j\omega)$, hence the concept of minimization of the peak value of $|H_2(j\omega)|$.

19.2.4. Wiener's Application of the RMS-error Criterion. N. Wiener\footnote{Different publications, especially “Extrapolation, Interpolation, and Smoothing of Stationary Time Series,” Wiley, New York, 1949.} suggested that the rms-error criterion be applied in a somewhat different way. Instead of starting from a transfer function $H(s)$ which is already roughly outlined with just a certain number of parameters being left undetermined, one can be more ambitious and state the problem as follows: Knowing the stochastic properties of the inputs (e.g., command and noise), determine the form of the transfer function of a system that minimizes $\overline{\varepsilon^2}$. If the system is later realized, it can be said that a total synthesis of the optimum linear system has been obtained.

Much literature has been devoted to Wiener's ideas and to the methods derived from them. We shall first outline briefly the general idea of
Wiener’s approach, then we shall show how this approach can be applied to servo systems.

19.2.5. Wiener’s Optimization Method. It is supposed that the input \( e(t) \) is a stationary stochastic function with a known autocorrelation function \( \varphi_e(\tau) \). It is desired to find the linear system which permits one to obtain a response \( r(t) \) equal to a desired function of \( e(t) \), say \( r_d(t) \), with the rms error, given by

\[
\overline{\varepsilon^2} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} [r(t) - r_d(t)]^2 \, dt
\]

as small as possible. The system is characterized by its impulse response \( h(t) \), which is the unknown function of the problem. The transfer function of the system is the Laplace transform of \( h(t) \) (Sec. 7.3.1).

The minimizing of \( \overline{\varepsilon^2} \) is accomplished by an application of the calculus of variations, leading to the following integral equation (Wiener-Hopf equation):

\[
\int_{-\infty}^{+\infty} h(\sigma) \varphi_e(\tau - \sigma) \, d\sigma - \varphi_{er}(\tau) = 0 \quad \text{for all } \tau \geq 0
\]

where \( \varphi_{er}(\tau) \) is the cross-correlation function of the input \( e(t) \) and the desired output \( r_d(t) \). It can be shown that this equation has a solution which, after generalized Fourier transformation of its terms, has the form

\[
H(\lambda) = \frac{1}{\Phi_+^+(\lambda)} \int_{0}^{\infty} \psi(t)e^{-\lambda t} \, dt
\]

in which:

a. \( \lambda \) is the complex frequency \( \sigma + j\omega \).

b. \( H(\lambda) = \int_{0}^{\infty} h(t)e^{-\lambda t} \, dt \) is the Fourier transform of the impulse response \( h(t) \), giving the transfer function of the system (Sec. 7.3.1).

c. \( \Phi_+^+(\lambda) \) is the part of the spectrum \( \Phi_e(\lambda) \) whose poles and zeros are situated in the right-hand half plane

\[
\Phi_e(\lambda) = \Phi_+^+(\lambda)\Phi_e^-(\lambda) = \int_{-\infty}^{+\infty} \varphi_e(\tau)e^{-\lambda \tau} \, d\tau
\]

d. \( \psi(t) \) is defined by the integral

\[
\psi(t) = \int_{C} \Psi(\lambda)e^{\lambda t} \, d\lambda
\]

taken along a path that encloses the poles of \( \Psi(\lambda) = \Phi_{er}(\lambda)/\Phi_+^-(\lambda) \).

19.2.6. The “Theory of Prediction.” The preceding theory and its results can be applied to the case in which \( r_d(t) = e(t - \alpha) \). Then \( \psi(t) \) becomes

\[
\psi(t) = \int_{C} \Phi_+^+(\lambda)e^{\lambda(t-\alpha)} \, d\lambda
\]

If \( \alpha < 0 \), one is thus enabled to solve the problem of prediction (\( \alpha \) sec in advance) of a stationary stochastic function, or more correctly, the problem of the most probable extrapolation.
Discussion. Without entering into details and mathematics, we will show that even in theories involving complex mathematical notions, common sense never loses its value. First of all, the prediction theory applies only for stationary inputs; it is a mathematical expression of the fact that same causes produce the same effects.

Furthermore, it is possible to define the degree of predictability of a stochastic function of time \( f(t) \). A function can be well predicted only if its second derivative remains finite. It is more predictable if its first derivative is also finite; on the other hand, functions that involve discontinuities are not predictable at all. The above results are not at all surprising.

Finally, these results can be expressed by examining, in the neighborhood of \( \tau = 0 \), the slope of the autocorrelation function \( \varphi(\tau) \) of the function \( f(t) \) which is to be predicted. The function \( f(t) \) is predictable if \( (d\varphi/d\tau)(0) \) is zero, and is not predictable in the contrary case (Fig. 19-8). This result has the merit of underlining the following significance of the autocorrelation function: a function is predictable if its correlation between \( f(t) \) and \( f(t + \alpha) \) is sufficiently strong; that is, if \( \varphi(\alpha) - \varphi(0) \) vanishes as \( \alpha \to 0 \)—or, faster, as \( \alpha \to 0 \)† (see Sec. 12.2.5, par. 3).

19.2.7. Application to Feedback Control Systems. Wiener's method has been applied to electrical networks and to servo systems. If the servo system shown in Fig. 19-7 is considered, it enables one to determine

![Fig. 19-8. Typical aspects of autocorrelation functions in the neighborhood of \( \tau = 0 \): Case a, predictability; Case b, nonpredictability.](image)

the optimum transfer function \( F(s) \), when the transfer function \( G(s) \) of the fixed components and the random characteristics of the command \( e(t) \) and the disturbance \( d(t) \) are known. It can be shown that the calculations become very simple.

The output of the system is given by Eq. (13-1), which can be written in terms of the given portion of the open-loop transfer function \( G(\lambda) \) and of the unknown closed-loop function \( H(\lambda) = FG/(1 + FG) \)

\[
R = HE + G(1 - H)D
\]

The Wiener-Hopf equation takes the form

\[
\int_{-\infty}^{+\infty} h(\tau)F_1(\tau - \sigma) \, d\tau + G_1(\sigma) = 0 \tag{19-4}
\]

In this expression \( F_1 \) and \( G_1 \) are functions of the autocorrelation functions \( \varphi_e \) and \( \varphi_d \) (of the input and the disturbance) and \( \varphi_{ed} \) and \( \varphi_{dr} \) (between input and disturbance and between disturbance and desired output). Now, the desired output of a servo system is \( r_d(t) = e(t) \). Taking this

† The reader will find a more fully developed outline of this discussion in R. Cohen, "Some Analytical and Practical Aspects of Wiener's Theory of Prediction," *Tech. Rept. 69*, Radiation Laboratory of Electronics, MIT, 1948.
into account, it can be shown that the Fourier transform of the Wiener-Hopf equation becomes

\[ Q(\lambda) = P(\lambda)[H(\lambda) - 1] \]  

(19-5)

where \( Q(\lambda) \) is a function of \( \lambda \) that should have all its poles in the right-hand half plane and \( P(\lambda) \) is given by

\[ P(\lambda) = \Phi_e(\lambda) - G(\lambda)\Phi_{ed}(\lambda) - G^*(\lambda)\Phi_{ed}(\lambda) + G(\lambda)G^*(\lambda)\Phi_d(\lambda) \]

where the \( \Phi \)s are the Fourier transforms of the correlation functions and \( G^* \) is the complex conjugate of \( G \).

If, moreover, no correlation exists between the input and the disturbance, the expression of \( P(\lambda) \) is reduced to

\[ P(\lambda) = \Phi_e(\lambda) + |G(\lambda)|^2\Phi_d(\lambda) \]  

(19-6)

where \( \Phi_e \) and \( \Phi_d \) are the input and disturbance spectra, respectively.

From these equations it is possible to obtain the desired transfer function \( H(\lambda) \) submitted to complementary physical conditions such as the following:

a. \( H(\lambda) \) must have no poles in the left half plane.

b. \( H(\lambda) \) must approach zero faster than \( 1/|\lambda| \) as \( |\lambda| \) becomes infinite.

c. The product \( P(\lambda)[H(\lambda) - 1] \) must have all its poles in the right-hand half plane.²

Thus, actual synthesis has been obtained by a straightforward procedure which leads to an optimum solution. Note that no integrations in the complex plane are involved, the only computation being that of \( \Phi_e + |G|^2\Phi_d \) by means of a sum of rational fractions.³

This technique can be extended to systems incorporating certain types of nonlinearities.

19.2.8. Examples. Two examples are given below. The first is purely numerical; the second consists in applying the Wiener method to a servo system whose compensation has already been performed by conventional techniques (Secs. 18.2.6, 18.3.3, and 18.4.1) and by the Guillemin-Truxal approach (Sec. 19.1.5).

1. Assume

\[ G(\lambda) = \frac{1}{\lambda^2} \quad \Phi_e(\lambda) = \frac{(13\sqrt{16})^{1/2}}{\lambda^2 + 13^{1/2}} \quad \Phi_d(\lambda) = 468^{1/2} \]

Using the expression of Eq. (19-6), the Wiener-Hopf equation (19-5) is

\[ Q(\lambda) = \frac{(\lambda + 2)(\lambda - 2)(\lambda + 3)(\lambda - 3)}{\lambda^4[\lambda + (3\sqrt{13})^{1/2}]^2[\lambda - (3\sqrt{13})^{1/2}]} [H(\lambda) - 1] \]

¹ The method can be extended to the case in which this condition is not fulfilled.

² This technique is that developed by one of the authors (Pélegrin, op. cit., pp. 46–56). Since the date of this publication (1952) a method giving an explicit expression of \( H(\lambda) \) has been derived by Newton, Gould, and Kaiser, "Analytical Design of Linear Feedback Controls," pp. 149–165, Wiley, New York, 1957.

³ The use of a series of orthonormal functions enables one to solve this problem quickly, more especially as the Fourier coefficients of the expression can be given by a computer (for instance, a correlator).
To satisfy conditions a and c, one is led to write

\[ H(\lambda) - 1 = k \frac{\lambda + (\frac{8}{13})^{34}}{(\lambda + 2)(\lambda + 3)} \]

\( H - 1 \) can be written in the form

\[ H(\lambda) - 1 = k \frac{\lambda + (\frac{8}{13})^{34}(\lambda + u)}{(\lambda + 2)(\lambda + 3)} \]

where \( u \) is to be determined by the condition that \( H(\lambda) \) should comply with condition b. It is thus found that \( u = 5 - (36/13)^{34} \); whence

\[ H(\lambda) = \frac{6 - (\frac{8}{13})^{34}[5 - (\frac{36}{13})^{34}]}{(\lambda + 2)(\lambda + 3)} \approx \frac{0.45}{(\lambda + 2)(\lambda + 3)} \]

2. Consider the system shown in Fig. 19-9 in which \( G(\lambda) = 1/\lambda(1 + \lambda) \) and the black box \( X \) represents the compensating network to be determined. Assume the following frequency spectra:

\[ \Phi_s(\lambda) = \frac{\Omega^2}{\pi} \frac{2k}{(2k)^2 + \lambda^2} \quad \Phi_d(\lambda) = n \]

Writing, for the sake of simplicity, \( \Omega^2/\pi = a \) and \( 2k = b \), the expression of \( P(\lambda) \) is

\[ P(\lambda) = \frac{-ab\lambda^4 + (ab + n)\lambda^2 - bn}{(b - \lambda)(b + \lambda)\lambda^2(1 - \lambda)(1 + \lambda)} \]

It is seen that \( P(\lambda) \) has a double pole at the origin.

*Fig. 19-9.*

a. Case 1. If the numerator has equal roots, the method can be applied by multiplying the denominator by \( (\lambda + u) \) and determining \( u \) from condition b. It is thus found that

\[ F(\lambda) = K \frac{\lambda(1 - 0.25\lambda)}{\lambda + b} \quad \text{with} \quad K = \left(\frac{bn}{a}\right)^{34} \left[ b + 1 - \left(\frac{bn}{a}\right)^{34} \right] - b \]

b. Case 2. For the general case in which the numerator does not have multiple roots, the double pole at the origin can be considered as the limiting case of two poles (one with a positive and one with a negative real part) as the latter simultaneously approach zero. Thus \( G(\lambda) \) is written

\[ G(\lambda) = \frac{1}{(\lambda + \epsilon)(1 + \lambda)} \quad G^*(\lambda) = \frac{1}{(\epsilon - \lambda)(1 - \lambda)} \]

If now \( H - 1 \) is written as

\[ H(\lambda) - 1 = -\frac{(\lambda + b)(\lambda + 1)(\lambda + \epsilon)}{(\lambda + u)(\lambda + v)(\lambda + w)} \]

where \( u \) and \( v \) are the roots of

\[-ab\lambda^4 + \lambda^2(n + ab) - nb^2 = 0\]
that have negative real parts, the solution is the limit as $\epsilon$ approaches zero of the expression

$$F(\lambda) = \frac{(1 + 0.25\lambda)((uv + vw + u\bar{w} - b)\lambda + uw)}{\lambda + b}$$

where $w = b + 1 - (u + v)$.

19.2.9. Comments on Optimization Methods. Optimization methods based on statistical considerations, and especially those based on Wiener's works, have been very much in vogue in recent years. This is probably due, at least partially, to the fact that the idea of optimizing a system is, intellectually, an extremely satisfactory one.\footnote{A valuable discussion of optimization methods can be found in H. Sartorius, "Das Optimierungsproblem in der Regelungstechnik," Regelungstechnik, 1(4): 74 (1953). Some authors have attempted to go further and design self-optimizing systems. In this connection, see E. Burt, "Self-optimizing Systems," in G. Müller (ed.), "Regelungstechnik: Moderne Theorien und ihre Verwendbarkeit," pp. 305-308, Oldenbourg, Munich, 1957; and A. Batkov and V. Solodovnikov, "Metod opredeleniya optimal'nykh kharakteristik odnovo klassa samonastravajuschikhksja sistem," ibid., pp. 308-322; and Automatika i Telemekhanika, 18(5): 377-391 (1957).}

Nevertheless, these methods are not so frequently used by design engineers as one might expect by judging from the great amount of literature devoted to them. One reason is that their application demands that complete statistical data concerning the inputs have been obtained. This represents a very large amount of work and causes the corresponding methods to be applicable only in the case of very elaborate or large-scale projects.

Furthermore, engineers must not overemphasize the practical value of optimization, which is but a modern, more systematic variation of the very old engineering practice of compromise. Optimizing equations is not equivalent to optimizing real physical systems; and people who have carried optimization methods through to physical realization will confess that the best thing to do, once the optimum has been calculated, is adjust the system to a value differing slightly from the optimum. In fact, the existence of an optimum indicates that the variables considered when obtaining it are stationary in its neighborhood, whereas variables that have not been taken into account are not. As a result, it is generally advisable to take the latter into account, thus setting the system slightly off optimum. The elements which have not been taken into account, because of the assumptions made while determining the optimum, often play a very important part, and spectacular examples are easy to quote.\footnote{One of the most striking is the lack of resemblance between real airplanes and the perfect aerodynamic body, classically determined from variational calculus. In the servo field, it is somewhat disappointing for the engineer who has worked out the application of Wiener's method to learn that many nonlinear systems may have better performance than the optimum linear one.}

Actually, the interest of optimization methods is not so much practical as philosophical. Such methods have the merit of showing upper limitations of the possible improvements of certain types of systems. As such, they are of great interest in helping the research engineer see clearly in
which direction his efforts for improving the system will be most beneficial. The great interest of the work "Analytical Design of Linear Feedback Controls"\(^1\) lies in showing what maximum performance can be expected from linear systems. Such a work is the best guide for selecting research topics in the linear field.

19.3. CONCLUSION ON SERVO-SYSTEM SYNTHESIS

The methods developed in the last two chapters make the synthesis of linear servo systems possible. Synthesis involves methods that are systematic enough for the design engineer to evaluate quantitatively his deviation from the stipulated specifications and decide what he should do to correct it. At the present time, this is quite possible, so far as linear systems are concerned, for the engineer who masters the methods outlined above. The engineer should master them all in order to be able to apply them simultaneously to a given problem, since these methods often complement one another:

1. Compensation by means of the harmonic approach has the advantage of remaining applicable when some components of the system are characterized not by their differential equations, but by experimental data.\(^2\) Philosophically, compensation is a synthesis procedure that starts from what exists (the fixed components) in order to decide what should be done.

2. The Guillemin-Truxal approach is applicable only when the differential equations of all the components are known. Philosophically, it starts from what should be, in order to treat what exists accordingly.

3. Statistical approaches are more satisfactory from the intellectual viewpoint, but they are applicable only for projects which deserve that a very considerable effort be spent on them. The practical results stemming from use of them do not, in practice, differ appreciably from those obtained by the more conventional approaches (1) and (2). Thus, the prime interest of statistical optimization would seem to be, essentially, to show the theoretical limits of possible improvements.

Summarizing, engineers of today are provided with very powerful tools for the synthesis of linear feedback control systems. So far as such systems are concerned, it is not an exaggeration to state that an optimum set of methods has been achieved. Although minor improvements will, of course, occur in the future, it can be said that, when starting from given fixed components, one is able today to construct the best possible system. Thus, the problem of obtaining better linear servo systems is no longer a problem of finding better methods for synthesis; rather, it is a problem of producing better servo components.


\(^2\) As will be seen in Part 3, the harmonic approach also has the advantage that it can be extended with a reasonable degree of accuracy to many nonlinear systems (Chap. 24).
CHAPTER 20
SAMPLED-DATA SYSTEMS

Summary

1. Linear sampled-data systems.
2. z transforms.
3. Frequency analysis of sampled-data systems.

A certain number of servo systems operate on sampled data. This means that, at one or several points of such a system, the information is fed, not continuously, but in the form of signals (samples) occurring at regular intervals of time. The time interval T between two consecutive instants at which information is fed into the system is called the period of sampling T.

Examples of such systems can be found in many regulating devices where the sensing element operates at discrete instants. Amplitude- or phase-modulation systems, including pulsed detectors and guidance systems associated with radar scanning, are sampled-data systems. A class of sampled-data systems which has increasingly extensive application is that of servo systems actuated through digital components, since digital computers work on samples of time functions. Servo systems operating on sampled data have been extensively studied in recent years, and several new works on the subject are to be published in the near future. The present chapter, which is restricted to the fundamental concepts that make the study of linear control systems with periodic data sampling possible, follows the existing literature on the subject, to which the reader is referred for more complete information. In particular, the following publications present interesting aspects of the question.

20.1. LINEAR SAMPLED-DATA SYSTEMS

20.1.1. Analysis of an Elementary Sampling System. When a continuous signal \( e(t) \) is the input of a linear system, the output \( r(t) \) can be completely defined by its Laplace transform \( R(s) \):

\[
R(s) = F(s)E(s)
\]  

(20-1)

where \( E(s) \) is the Laplace transform of the input \( e(t) \) and \( F(s) \) is the transfer function of the system. Now assume that the function \( e(t) \) is known only at periodic instants \( 0, T, 2T, 3T, \ldots \), termed sampling instants, and that the input \( e_1(t) \) of the system is equal to \( e(t) \) at the sampling instants but is maintained constant between these instants (Fig. 20-1).

\[
e_1(t) = e(0) \quad 0 \leq t < T
\]
\[
e_1(t) = e(T) \quad T \leq t < 2T
\]
\[
e_1(t) = e(2T) \quad 2T \leq t < 3T
\]

and so on. The Laplace transform \( E_1(s) \) of the input \( e_1(t) \) is (Sec. 4.2.1)

\[
E_1(s) = \frac{e(0)}{s} (1 - e^{-Ts}) + \frac{e(T)e^{-Ts}}{s} (1 - e^{-Ts})
\]

\[
+ \frac{e(2T)e^{-2Ts}}{s} (1 - e^{-Ts}) + \cdots
\]

\[
E_1(s) = \frac{1 - e^{-Ts}}{s} [e(0) + e(T)e^{-Ts} + e(2T)e^{-2Ts} + \cdots]
\]

By defining

\[
E^*(s) = e(0) + e(T)e^{-Ts} + e(2T)e^{-2Ts} + e(3T)e^{-3Ts} + \cdots
\]

\[
= \sum_{n=0}^{\infty} e(nT)e^{-nTs} \quad (20-2)
\]

one obtains

\[
E_1(s) = \frac{1 - e^{-Ts}}{s} E^*(s) \quad (20-3)
\]

Hence the Laplace transform of the response is

\[
R(s) = \frac{1 - e^{-Ts}}{s} F(s)E^*(s) \quad (20-4)
\]
The expression \( E^*(s) \), defined by Eq. (20-2), can be looked upon (Secs. 4.2.1 and 4.3.2) as the Laplace transform of a function of time \( e^*(t) \) that consists of a sequence of impulses occurring at the instants \( 0, T, 2T, \ldots \) with respective magnitudes \( e(0), e(T), e(2T), \ldots \), respectively:

\[
e^*(t) = e(0)\delta(t) + e(T)\delta(t-T) + e(2T)\delta(t-2T) + \cdots
\]

where \( \delta(t) \) or \( u_1(t) \) is the unit-impulse function with

\[
\mathcal{L}\delta(t) = 1 \quad \mathcal{L}\delta(t-T) = e^{-Ts} \quad \text{etc.}
\]

The function \( e^*(t) \) is shown in Fig. 20-2, where the impulses are represented by arrows with lengths equal to their respective magnitudes. It is termed the sampled time function associated with the continuous time function \( e(t) \) and the sampling period \( T \). The function \( E^*(s) \) is the associated sampled Laplace transform. Note that the functions \( e^*(t) \) and \( E^*(s) \) depend only on the values of \( e(t) \) at the sampling instants, and not on the values of \( e(t) \) during the intervals between sampling instants.

Referring to the above example, it can be seen that the actual input to the system, \( e_1(t) \), can be obtained from \( e(t) \) by performing two successive operations (Fig. 20-3):

1. The operation that gives \( e^*(t) \) from \( e(t) \). This operation is called sampling, and the corresponding operator is a sampler.

2. The operation that gives \( e_1(t) \) out of \( e^*(t) \). In terms of Laplace transforms this operation is a multiplication by the transfer function \( (1 - e^{-Ts})/s \) and is performed by what is called a clamping, or holding, system. Consequently, by taking Eq. (20-4) into account, it is seen that the output \( r(t) \) can be obtained from the input \( e(t) \) by performing three successive operations:
a. Sampling
b. Holding, that is, multiplication by \((1 - e^{-T_s})/s\)
c. Multiplication by \(F(s)\)

Operations \(b\) and \(c\) can be combined into a multiplication by the transfer function \((1 - e^{-T_s})F(s)/s\), (Fig. 20-4). The problem of studying the performance of linear sampled-data systems can, therefore, be solved in a general manner by studying the response \(r(t)\) of linear systems to sampled functions of the form \(e^\ast(t)\) (Fig. 20-5). The response \(r(t)\) is a continuous time function that is completely characterized by its Laplace transform:

\[
R(s) = H(s)E^\ast(s) = H(s) \sum_{n=0} e(nT) e^{-nT_s} \tag{20-5}
\]

For the particular case considered above as an introductory example, the function \(H(s)\) is

\[
H(s) = F(s) \frac{1 - e^{-T_s}}{s}
\]

**20.1.2. Analysis of the Sampler.** As shown previously, the physical situation defined in the preceding example can be analyzed by considering an ideal operation, termed **sampling**, which converts a continuous signal into a sequence of regularly spaced impulses, the height (or area) of an impulse representing the value of the time function at the associated sampling instant. Thus, sampling can be considered as **modulation** of the impulse train \(i(t)\), shown in Fig. 20-6, by the time function \(e(t)\). Thus,

\[
e^\ast(t) = e(t)i(t)
\]

where \(i(t) = u_1(t) + u_1(t - T) + u_1(t - 2T) + \cdots\)

Sampling is an ideal operation. It may be physically realized to a satisfactory approximation by using a pulse generator which generates impulses whose duration is sufficiently small compared with the time constants of the system.
Note 1. Equation (20-5) shows that the sampling device cannot be described by an ordinary transfer function relating the transforms of input and output, as was the case in Eq. (20-1) for an ordinary linear system. Equation (20-5) indicates, however, that the principle of superposition is still applicable.

Note 2. Sampling is considered as modulation by a unit-impulse train. If this impulse train is multiplied by a constant factor, the factor should be introduced into the gain of the linear element.

20.1.3. Laplace Transform of a Sampled Time Function. If one is interested in the values of the output \( r(t) \) (Fig. 20-7) at the sampling instants \( 0, T, 2T, \ldots \) and not in a complete knowledge of \( r(t) \), a simple relation between the sampled input and the sampled output can be derived. Applying the superposition principle to the linear system \( H(s) \), it is seen that the response \( r(t) \) is obtained by summing the responses of the system to each impulse. Hence, terming \( h(t) \) the unit-impulse response of the system [recall that \( H(s) = \mathcal{L}h(t) \), see Sec. 7.3.1]:

\[
\begin{align*}
  r(t) &= e(0)h(t) & 0 \leq t < T \\
  r(t) &= e(0)h(t) + e(T)h(t - T) & T \leq t < 2T \\
  r(t) &= e(0)h(t) + e(T)h(t - T) + e(2T)h(t - 2T) & 2T \leq t < 3T \\
  \vdots & & \vdots
\end{align*}
\]

etc., and at the sampling instants:

\[
\begin{align*}
  r(0) &= e(0)h(0) \\
  r(T) &= e(0)h(T) + e(T)h(0) \\
  r(2T) &= e(0)h(2T) + e(T)h(T) + e(2T)h(0) \\
  \vdots \\
  r(nT) &= \sum_{p=0}^{n} e(pT)h[(n - p)T]
\end{align*}
\]

Now from Eq. (20-2),

\[
R^*(s) = r(0) + r(T)e^{-Ts} + r(2T)e^{-2Ts} + \cdots = \sum_{n=0}^{\infty} r(nT)e^{-nTs}
\]

Hence,

\[
R^*(s) = \sum_{n=0}^{\infty} \sum_{p=0}^{\infty} e(pT)h[(n - p)T]e^{-nTs}
\]

It can be verified from this expression that

\[
R^*(s) = [e(0) + e(T)e^{-Ts} + \cdots]h(0) + h(T)e^{-Ts} + \cdots = E^*(s)H^*(s)
\]

where \( H^*(s) \) is the sampled Laplace transform associated with the unit-impulse response \( h(t) \) or, equivalently, with the transfer function \( H(s) \). In other words, the sampled transform associated with the response of a linear system to a sampled input is simply the sampled transform associated with input multiplied by the sampled transform associated with the transfer function of the system,

\[
R^*(s) = E^*(s)H^*(s)
\]

(20-6)
Equation (20-6) is simpler than Eq. (20-5), but the inverse Laplace transform of \( R^*(s) \) is coincident with the system response only at the sampling instants. No information is provided concerning the behavior of \( r(t) \) during the intervals between sampling instants.

![Fig. 20-8. Cascade system with \( H_1 \) and \( H_2 \) in tandem.](image)

![Fig. 20-9. Cascade system with \( H_1 \) and \( H_2 \) separated by sampler.](image)

20.1.4. Different Possible Cascade Combinations. If the sampler is followed by two cascaded linear systems with transfer functions \( H_1(s) \) and \( H_2(s) \) (Fig. 20-8), these systems may be combined in one linear system with the transfer function \( H_1(s)H_2(s) \). The (ordinary) Laplace transform of the output \( r_2(t) \) is

\[
R_2(s) = E^*(s)H_1(s)H_2(s)
\]

and its sampled transform is

\[
R^*_2(s) = E^*(s)[H_1(s)H_2(s)]^*
\]

If another sampler operating at the same sampling instants 0, \( T \), \( 2T \), \( \ldots \) is inserted between the two systems \( H_1(s) \) and \( H_2(s) \) (Fig. 20-9), it is no longer possible to combine \( H_1 \) and \( H_2 \), since the input of \( H_2(s) \) is \( r_1^*(t) \) and not \( r_1(t) \). The Laplace transform \( R'_2(s) \) of the output \( r'_2(t) \) is [Eq. (20-5)]

\[
R'_2(s) = H_2(s)R^*_1(s)
\]

or, taking into account \( R^*_1(s) = E^*(s)H^*_1(s) \),

\[
R'_2(s) = E^*(s)H^*_1(s)H_2(s)
\]

Hence the sampled transform \( R^*_2(s) \) associated with \( r'_2(t) \) is

\[
R^*_2(s) = H^*_1(s)R^*_1(s)
\]

Finally

\[
R^*_2(s) = E^*(s)H^*_1(s)H^*_2(s)
\]

The factor \( H^*_1(s)H^*_2(s) \) differs from the factor \( [H_1(s)H_2(s)]^* \), often denoted by \( H_1H^*_2(s) \), which appears in Eq. (20-7b). As a consequence of Eq. (20-8b) it is seen that, when the information is carried through cascaded systems in terms of sampled data (this is the case when digital components are present) the sampled transforms associated with the cascaded systems simply multiply in the usual manner.

20.1.5. Application to Feedback Systems. Consider a feedback system in which the error signal is sampled [Fig. 20-10, in which \( G_1(s) \) and \( G_2(s) \) are linear transfer functions]. Equations (20-5) and (20-7b) yield

\[
R(s) = \mathcal{E}^*(s)G_1(s) \quad R'^*(s) = \mathcal{E}^*(s)G_1G^*_2(s)
\]

Furthermore, the fundamental equation \( \varepsilon(t) = e(t) - r'(t) \) obviously yields

\[
\mathcal{E}^*(t) = e^*(t) - r'^*(t) \quad \mathcal{E}^*(s) = E^*(s) - R'^*(s)
\]
Hence, by eliminating $E^*(s)$,

$$R(s) = \frac{G_1(s)}{1 + G_1G_2^*(s)} E^*(s) \quad (20-9a)$$

The sampled transform of the output is

$$R^*(s) = \left[ G_1(s) \frac{E^*(s)}{1 + G_1G_2^*(s)} \right]^*$$

that is, after Eq. (20-6),

$$R^*(s) = \frac{G_1^*(s)}{1 + G_1G_2^*(s)} E^*(s) \quad (20-9b)$$

It can be found in a similar manner that the sampled transform of the error is

$$E^*(s) = \frac{1}{1 + G_1G_2^*(s)} E^*(s)$$

These results can be extended to the case in which the sampling device operates at other points in the loop.

### 20.2. $z$ TRANSFORMS

#### 20.2.1. Definition. It is seen in Eq. (20-2) that only the factor $e^{-Ts}$ depends on $s$ in the sampled transform $E^*(s)$. By letting

$$e^{Ts} = z \quad (20-10)$$

$E^*(s)$ can be written as a function of $z$. The expression thus obtained

$$E(z) = e(0) + \frac{e(T)}{z} + \frac{e(2T)}{z^2} + \cdots = \sum_{n=0}^{\infty} e(nT)z^{-n} \quad (20-11)$$

is called the $z$ transform of $e(t)$.$\dagger$ The $z$ transform does not differ from the sampled transform of which it is merely a simpler form. It must be recalled that $z$ transforms are associated with time functions defined only at periodic sampling instants, the period of which must be specified. All

---

1 Some authors define $z$ as $\exp(-Ts)$, which is equivalent to changing $z$ into $1/z$.

$\dagger$ The function $E(z)$ has not the same form as the function $E(s)$. Strictly speaking, a different notation should therefore be used for $z$ transforms. Generally, no misunderstanding is possible, since the letter $E$ indicates that $E(s)$ is associated with $E(s)$ and $e(t)$, and the argument $z$ implies sampling.
the equations derived in Sec. 20.1 can be rewritten by replacing sampled transforms by $z$ transforms.

20.2.2. Fundamental Properties of $z$ Transforms. Any time function $f(t)$ has an associate $z$ transform $F(z)$ for a given sampling period $T$. All time functions which have the same values as $f(t)$ at the associated sampling instants have the same $z$ transform.

When a continuous function $e(t)$ is fed into a sampler, and then filtered by a linear system of transfer function $H(s)$, the output $r(t)$ is a continuous function whose Laplace transform $R(s)$ is given by

$$R(s) = H(s) \sum_{n=0}^{\infty} e(nT)e^{-nTs}$$

The $z$ transform $R(z)$ of $r(t)$ is given by

$$R(z) = H(z)E(z) \quad \text{(20-12)}$$

The rule of computation is the same as for linear systems with continuous inputs. The function $H(z)$ is termed the $z$ transfer function of the linear system $H(s)$. It is the $z$ transform associated with the unit-impulse response of the system $h(t) = \mathcal{L}^{-1}H(s)$. Note that Eq. (20-12) gives no information concerning what happens to $r(t)$ between the sampling instants. If it is desired to determine $r(t)$ completely, one must resort to ordinary Laplace transforms.

20.2.3. Deriving Simple $z$ Transforms. 1. $z$ Transform of a Unit Step $u(t)$ (Fig. 20-11). From Eq. (20-11)

$$U(z) = \sum_{n=0}^{\infty} z^{-n} = 1 + \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \cdots$$

This series is convergent for $|z| > 1$, that is, $|s| > 0$, and yields:

$$U(z) = \frac{1}{1 - 1/z} = \frac{z}{z - 1}$$

2. $z$ Transform of an Exponential $f(t) = \exp(-at)$. From Eq. (20-11)

$$F(z) = \sum_{n=0}^{\infty} e^{-anT}z^{-n} = 1 + \frac{e^{-aT}}{z} + \left(\frac{e^{-aT}}{z}\right)^2 + \cdots$$

This series is convergent for $|z| > \exp(-aT)$ and yields

$$F(z) = \frac{z}{z - e^{-aT}}$$

3. $z$ Transfer Function for a First-order System $H(s) = 1/(1 + Ts)$. The unit-impulse response is $h(t) = e^{-t/T}/T$, hence the $z$ transfer function is

$$H(z) = \frac{1}{\tau z - e^{-T/\tau}}$$
4. z Transfer Function for a Holding Device $H(s) = (1 - e^{-Ts})/s$. The unit-impulse response is $u(t) - u(t - T)$ (Fig. 4-2). The $z$ transfer function associated with $1/s$ is $z/(z - 1)$. The $z$ transfer function associated with $e^{-Ts}/s$ is obtained from Eq. (20-11),

$$0 + z^{-1} + z^{-2} + z^{-3} + \ldots = \frac{1}{z^2 - 1} = \frac{1}{z - 1}$$

Hence,

$$H(z) = \frac{z}{z - 1} - \frac{1}{z - 1} = 1$$

This expression is identical with the $z$ transform of a unit impulse. This can be understood by noting that $h(t)$ has the same value (zero) as the unit-impulse function $\delta(t)$ or $u_1(t)$ for $t = nT, n$ being any positive integer.

5. $z$ Transfer Function for Holding Device in Tandem with a First-order System

$$H(s) = \frac{1 - e^{-Ts}}{s + a}$$

The $z$ transfer functions associated with

$$\frac{1}{s(s + a)} = \frac{1/a}{s} - \frac{1}{s + a} \quad \text{and} \quad \frac{e^{-Ts}}{s + a}$$

are

$$H_1(z) = \frac{1}{a} \frac{z}{z - 1} - \frac{1}{a} \frac{z}{z - e^{-aT}} \quad \text{and} \quad H_2(z) = \frac{1}{z} H_1(z)$$

respectively. Hence,

$$H(z) = H_1(z) - H_2(z) = \frac{1}{a} \left(1 - \frac{z - 1}{z - e^{-aT}}\right) = \frac{1 - e^{-aT}}{a} \frac{1}{z - e^{-aT}}$$

20.2.4. Table of $z$ Transforms and its Use for Obtaining the Time Response at the Sampling Instants. A table can be computed for $z$ transforms as well as for Laplace transforms associated with common time functions (Table 20-1). This property results from the fact that the superposition principle applies to $z$ transforms as well as to $s$ transforms and time functions. If now the $z$ transform of the input to the system is also a polynomial fraction in $z$, it is possible to perform the partial-fraction expansion for $R(z) = H(z)E(z)$. The associated time functions can then be found with the help of the table. Summing them yields a certain time function $r_s(t)$ which has the same value as the actual response at the sampling instants $0, T, 2T, \ldots$ but differs from it within the sampling intervals:

$$r_s(t) \neq r(t) \quad \text{but} \quad r_s(t)i(t) = r(t)i(t)$$

Therefore, only the values of $r_s(t)$ at the sampling instants are relevant. Among the infinity of time functions that are equal to $r(t)$ at the sampling instants, $r_s(t)$ has the particularity of being, like $r(t)$, a sum of terms

1 This can be performed either algebraically or graphically, in a manner similar to the pole-zero procedure outlined in Sec. 9.3.5.

† The subscript $z$ is used in order to recall that this particular time function is obtained by means of the $z$-transform procedure.
### Table 20-1. Laplace and z Transforms (z = e^{T})

<table>
<thead>
<tr>
<th>Time function</th>
<th>Laplace transform</th>
<th>z transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit impulse ( u(t) ) or ( \delta(t) )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( \delta(t - nT) )</td>
<td>( e^{-nTz} )</td>
<td>( z^{-n} )</td>
</tr>
<tr>
<td>Unit step ( u(t) )</td>
<td>( \frac{1}{s} )</td>
<td>( \frac{z}{z - 1} )</td>
</tr>
<tr>
<td>( tu(t) )</td>
<td>( \frac{1}{s^2} )</td>
<td>( \frac{Tz}{(z - 1)^2} )</td>
</tr>
<tr>
<td>( t^2 )</td>
<td>1</td>
<td>1 ( \frac{z(z + 1)}{(z - 1)^3} )</td>
</tr>
<tr>
<td>( e^{-at} )</td>
<td>( \frac{1}{s + a} )</td>
<td>( \frac{z}{z - e^{-aT}} )</td>
</tr>
<tr>
<td>( at/T )</td>
<td>( \frac{s}{s - (\ln a)/T} )</td>
<td>( \frac{z}{z - a} )</td>
</tr>
<tr>
<td>( e^{-aTf(t)} )</td>
<td>( F(s + a) )</td>
<td>( F(e^{aTz}) )</td>
</tr>
<tr>
<td>( \sin \omega t )</td>
<td>( \frac{s}{s^2 + \omega^2} )</td>
<td>( \frac{z \sin \omega T}{z^2 - 2z \cos \omega T + 1} )</td>
</tr>
<tr>
<td>( \cos \omega t )</td>
<td>( \frac{s}{s^2 + \omega^2} )</td>
<td>( \frac{z^2 - z \cos \omega T}{z^2 - 2z \cos \omega T + 1} )</td>
</tr>
<tr>
<td>( e^{-at} \cos \omega t )</td>
<td>( \frac{s + a}{(s + a)^2 + \omega^2} )</td>
<td>( \frac{z^2 - ze^{-aT} \cos \omega T}{z^2 - 2ze^{-aT} \cos \omega T + e^{-2aT}} )</td>
</tr>
<tr>
<td>( e^{-at} \cos \frac{\pi t}{T} ) (Exponential alternately positive and negative)</td>
<td></td>
<td>( \frac{z}{z + e^{-aT}} )</td>
</tr>
<tr>
<td>( e^{-at} \sin \omega t )</td>
<td>( \frac{s}{(s + a)^2 + \omega^2} )</td>
<td>( \frac{z \sin \omega Te^{-aT}}{z^2 - 2ze^{-aT} \cos \omega T + e^{-2aT}} )</td>
</tr>
</tbody>
</table>

Typically encountered in the solution of linear differential equations, such as

\[(at + b)^n e^{-at} e^{-at} \sin (\omega t + \varphi) \text{ etc.}\]

As can be seen in the table, this results from the fact that \( R(z) \) is a polynomial fraction in \( z \). It is possible to obtain directly the sampled values of \( r(t) \) without computing \( r_s(t) \). Thus, assume that an expansion
of \( R(z) \) in powers of \( 1/z \) has been performed,

\[
R(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \cdots
\]

Referring to Eq. (20-11), it can be seen that the coefficient of \( z^{-n} \) in this expansion is equal to the value of the time function at the \( n \)th sampling instant.

20.2.5. Examples. 1. Response of a First-order System to a Sampled Step Input (Fig. 20-12). Assume that a first-order system, of transfer function \( H(s) = 1/(1 + \tau s) \), is subjected to the information supplied by an ideal sampler. This sampler yields impulses occurring at periodic sampling instants \( 0, T, 2T, 3T, \ldots \), the magnitude of each impulse being the value of the input \( e(t) \) to the sampler at the associated instants, \( 0, T, 2T, 3T, \ldots \). What is the response \( r(t) \) of the linear system \( H(s) \) to a unit-step input \( e(t) = u(t) \) to the sampler?

Using Eq. (20-5), one can derive \( R(s) \):

\[
R(s) = H(s) \sum_{n=0}^{\infty} e(nT)e^{-nT}\tau
\]

Now

\[
e(0) = e(T) = e(2T) = \cdots = e(nT) = 1
\]

Therefore,

\[
R(s) = \frac{1}{1 + \tau s} + \frac{1}{1 + \tau s} e^{-sT} + \frac{1}{1 + \tau s} e^{-2sT} + \cdots
\]

\[
r(t) = h(t) + h(t - T) + h(t - 2T) + h(t - 3T)
\]

where \( h(t) = e^{-t/T} \tau \) is the unit-impulse response of the system. The response \( r(t) \) can be drawn by graphical superpositions. In particular, at the sampling instants it is found that

\[
\begin{align*}
    r(0) &= \frac{1}{\tau} \\
    r(T) &= \frac{1}{\tau} (1 + e^{-T/\tau}) \\
    r(2T) &= \frac{1}{\tau} (1 + e^{-T/\tau} + e^{-2T/\tau}) \\
    r(nT) &= \frac{1}{\tau} (1 + e^{-T/\tau} + \cdots + e^{-nT/\tau})
\end{align*}
\]

These values, \( r(0), r(T), r(2T), \ldots \) can be obtained directly by making use of \( z \) transforms. Starting from the transforms found in Table 20-1,

\[
E(z) = \frac{z}{z - 1} \quad H(z) = \frac{1}{\tau} \frac{z}{z - \exp(-T/\tau)}
\]

and applying Eq. (20-12), it is found that

\[
R(z) = \frac{1}{\tau (1 - \lambda)} \left( \frac{z}{z - 1} - \frac{\lambda z}{z - \lambda} \right)
\]

This expression can be expanded into partial fractions. Letting, for the sake of simplicity, \( \lambda = e^{-T/\tau} \), the equation yields

\[
R(z) = \frac{1}{\tau (1 - \lambda)} \left( \frac{z}{z - 1} - \frac{\lambda z}{z - \lambda} \right)
\]
The sampled time function associated with \( R(z) \), as then obtained from Table 20-1, is

\[
r_s(t) = \frac{1}{\tau(1 - \lambda)} [u(t) - \lambda \exp(-t/\tau)] = \frac{1 - \lambda \exp(-t/\tau)}{\tau(1 - \lambda)}
\]

Hence, for \( t = nT \)

\[
r(nT) = r_s(nT) = \frac{1 - \lambda^{1+n}}{\tau(1 - \lambda)}
\]

which can be verified as identical with the value obtained from conventional Laplace transform computation (Fig. 20-13)

\[
r(nT) = \frac{1}{\tau} \left( 1 + e^{-\tau/\tau} + \cdots + e^{-n\tau/\tau} \right)
\]

Instead of expanding \( R(z) \) into partial fractions, it is also possible to perform an expansion with respect to the powers of \( 1/z \). This yields successively, writing \( \lambda \) instead of \( \exp(-T/\tau) \) for simplicity,

\[
R(z) = \frac{1}{\tau z} \frac{z}{1 - z} = \frac{1}{\tau} \left( \frac{1}{1 - 1/z} \frac{1}{1 - \lambda/z} \right)
\]

\[
R(z) = \frac{1}{\tau} \left( 1 + \frac{1}{z} + \frac{1}{z^2} + \cdots \right) \left( 1 + \frac{\lambda}{z} + \frac{\lambda^2}{z^2} + \cdots \right)
\]

\[
R(z) = \frac{1}{\tau} z^0 + \frac{1}{\tau} (1 + \lambda) z^{-1} + \frac{1}{\tau} (1 + \lambda + \lambda^2) z^{-2} + \cdots + \frac{1}{\tau} (1 + \lambda + \cdots + \lambda^n) z^{-n} + \cdots
\]

The value of \( r(t) \) for \( t = nT \) is given by the coefficient of the \( n \)th term in the series expansion

\[
r(0) = \frac{1}{\tau} \quad r(T) = \frac{1}{\tau} (1 + \lambda) \quad \ldots \quad r(nT) = \frac{1}{\tau} (1 + \lambda + \cdots + \lambda^n)
\]

Assume, now, that the sampler is followed by a holding device. The holding device and the linear component \( H(s) \) can be lumped into a single linear system with a transfer function

\[
\frac{1 - \exp(-Ts)}{1 + \tau s}
\]

\[
H(s) = \frac{1 - \exp(-t/\tau)}{1 + \tau s}
\]

and an associated \( z \) transfer function \((1 - \lambda)/(z - \lambda)\), where \( \lambda = e^{-T/\tau} \). The \( z \) transform of the response is

\[
R(z) = \frac{1 - \lambda}{z - \lambda} = \frac{z}{z - 1} - \frac{\lambda}{z - \lambda} \quad r_s(t) = 1 - e^{-t/\tau}
\]
Thus \( r_s(t) \) is found to be the response of the first-order system \( 1/(1 + \tau s) \) to a continuous unit-step input. In fact, a unit-step function is not altered by the operations of sampling and holding.

2. Response of a First-order System Following Sampling and Holding Devices to a Ramp Input (Fig. 20-14). The \( z \) transform of the response is obtained from Eq. (20-12) by means of

\[
E(z) = \frac{Tz}{(z - 1)^2} \quad H(z) = \frac{1 - \lambda}{z - \lambda} \quad \lambda = e^{-\tau/\tau}
\]

Identification of \( R(z) \) with

\[
R(z) = \frac{Az}{z - 1} + \frac{BTz}{(z - 1)^2} + \frac{Cz}{z - \lambda}
\]

yields

\[
A = \frac{-T}{1 - \lambda} \quad B = 1 \quad C = \frac{T}{1 - \lambda}
\]

Hence

\[
r_s(t) = A + Bt + C e^{-\tau t}
\]

\[
r(nT) = A + BnT + C\lambda^n = nT - \frac{T}{1 - \lambda} (1 - \lambda^n)
\]

For large values of \( t \)

\[
r^*(t) \cong (Bt + A)^* \cong \left( t - \frac{T}{1 - \lambda} \right)^*
\]

This means that the response at sampling instants tends, as \( \exp (-nT/\tau) \) becomes negligible, to become a straight line parallel to the input and delayed by the time

![Fig. 20-14.](image)

![Fig. 20-15.](image)

\( T/(1 - \lambda) \). If the sampling period \( T \) is small compared to the time constant \( \tau \) of the system, then

\[
\lambda = e^{-\tau/\tau} = 1 - \frac{T}{\tau} + \frac{1}{2} \left( \frac{T}{\tau} \right)^2 - \cdots \cong 1 - \frac{T}{\tau}
\]

and the delay \( T/(1 - \lambda) \) is approximately \( \tau \); that is, the system behaves approximately like an ordinary first-order system (Sec. 5.1.4).

3. Response of a First-order Servo with Sampling and Holding Devices (Fig. 20-15). From Eq. (20-96)

\[
R(z) = \frac{KG(z)}{1 + KG(z)} E(z)
\]

with

\[
G(s) = \frac{1 - \exp (-T_s)}{s} \quad \frac{1}{1 + \tau s} \quad G(z) = \frac{1 - \lambda}{z - \lambda} \quad \lambda = e^{-\tau/\tau}
\]
The closed-loop $z$ transfer function is

$$\frac{KG(z)}{1 + KG(z)} = \frac{K(1 - \lambda)}{z + K(1 - \lambda) - \lambda}$$

The $z$ transform of the response to a step input is

$$R(z) = \frac{K(1 - \lambda)}{z + K(1 - \lambda) - \lambda - \lambda z^{-1}}$$

$$= \frac{K}{1 + K} \left[ \frac{z}{z - 1} - \frac{z}{z + K(1 - \lambda) - \lambda} \right]$$

(a) If $K(1 - \lambda) - \lambda < 0$ that is $K < \frac{1}{\lambda - 1}$

the expression of the response is

$$r(t) = \frac{K}{1 + K} - \frac{K}{K + 1} \exp \left\{ \frac{t}{T} \ln |\lambda - K(1 - \lambda)| \right\}$$

The forced response is $K/(1 + K)$; that is, there is a steady-state error $1/(1 + K)$. The transient is unstable if

$$\lambda - K(1 - \lambda) > 1 \quad \text{that is} \quad K < -1$$

(b) If $K > 1/(\lambda - 1)$, then

$$r(t) = \frac{K}{1 + K} - \frac{K}{1 + K} \exp \left[ \frac{t}{T} \ln |\lambda - K(1 - \lambda)| \right] \cos \frac{\pi}{T} t$$

the transient response is oscillatory (and stable, since the condition $K > -1$ is satisfied). This possibility for a first-order servo to exhibit oscillations is a consequence of the sampling.

![Fig. 20-16.](image)

4. Response of a First-order Servo with Integration (Fig. 20-16). The open-loop transfer functions are

$$G(s) = \frac{1 - e^{-\tau s}}{s} \quad G(z) = \frac{Tz}{(z - 1)^2} \left( 1 - \frac{1}{z} \right) = \frac{T}{z - 1}$$

whence the closed-loop $z$ transfer function is

$$\frac{KG(z)}{1 + KG(z)} = \frac{KT}{z - 1 + KT}$$

and the $z$ transform of the step-input response is

$$R(z) = \frac{KT}{z - 1 + KT} \frac{z}{z - 1} = \frac{z}{z - 1} - \frac{z}{z - (1 - KT)}$$

When $1 - KT > 0$, that is, $K < 1/T$, the $z$ response is

$$r_z(t) = 1 - \exp \left[ \frac{t}{T} \ln (1 - KT) \right]$$
There is no steady-state error. The response is stable if \( 1 - KT < 1 \), that is, \( K > 0 \). For \( K > 1/T \), the response is oscillatory:

\[
    r(t) = 1 - \exp \left[ \frac{t}{T} \ln (KT - 1) \right] \cos \frac{\pi t}{T}
\]

This oscillatory response is stable if \( KT - 1 < 1 \), that is, for \( K < 2/T \); it is unstable if \( K > 2/T \). The possibility of this servo being unstable is due to the sampling.

**Fig. 20-17.**

5. **z Transfer Function of a Second-order Servo with Sampled Data** (Fig. 20-17).

This time, again letting \( \lambda = \exp (-T/r) \),

\[
    G(s) = \frac{1}{s} - \frac{r}{1 + rs} \quad G(z) = \frac{z}{z - 1} - \frac{z}{z - \lambda} = \frac{z(1 - \lambda)}{(z - 1)(z - \lambda)}
\]

The closed-loop \( z \) transfer function is

\[
    \frac{KG(z)}{1 + KG(z)} = \frac{Kz(1 - \lambda)}{z^2 + [K(1 - \lambda) - (1 + \lambda)]z + \lambda}
\]

For example, the \( z \) transform of the unit-step response will be

\[
    R(z) = \frac{z}{z - 1} \frac{Kz(1 - \lambda)}{z^2 + [K(1 - \lambda) - (1 + \lambda)]z + \lambda}
\]

It is seen that mathematical expression of \( r(t) \) quickly becomes extremely complicated as the complexity of the system under consideration increases.

**20.2.6. Stability of Sampled-data Linear System.** 1. **Condition of Stability in the z Plane.** Consider a linear system where sampling operations are involved. Because of the linearity, the response \( r(t) \) of this system to an impulse input is a sum of linear terms such as

\[
    (at + b)^n, e^{-at}, e^{-at} \sin (\omega t + \varphi), \ldots
\]

Now the \( z \) transform of \( r(t) \) is \( R(z) = H(z)E(z) = H(z) \times 1 = H(z) \), which is a polynomial fraction in \( z \). From the \( z \)-transform form, a certain \( r_z(t) \) can be associated with \( H(z) \), such that \( r_z(t) \equiv r(t) \). Like \( r(t) \), \( r_z(t) \) is a sum of linear terms of the form

\[
    (a't + b')^n, e^{-a't}, e^{-a't} \sin (\omega't + \varphi'), \ldots
\]

As a result, \( r(t) \) and \( r_z(t) \) are simultaneously stable, or unstable. Thus, the stability of \( r(t) \)—that is, the stability of the system—can be determined from the stability of \( r_z(t) \), i.e., the stability of \( H(z) \).

The stability of the response \( r_z(t) \) is determined by the exponential terms it involves. Real exponentials \( e^{-at} \) are associated with poles \( e^{-aT} \) of \( H(z) \); oscillatory exponentials \( e^{-at} \cos (\omega t + \varphi) \) are associated with poles \( e^{-aT} e^{i\omega T} \). Therefore, for every pole \( z_i \) of \( H(z) \), the condition of stability \( \alpha_i > 0 \) can be written \( |z_i| < 1 \) (Fig. 20-18).
This result can be obtained directly by noting that the $z$ transform is derived from the $s$ transform by letting $e^{sT} = z$. In the $s$ plane, the condition of stability is that the poles of the transfer function lie in the left half plane; in the $z$ plane, the associated poles should lie inside the unit circle centered at the origin.

2. **Application to Feedback Sampled-data Systems.** In the case of a servo system with unity feedback incorporating a sampled-error device (Fig. 20-19), the $z$ transform of the closed-loop system is

$$H(z) = \frac{KG(z)}{1 + KG(z)}$$

The system is stable if all the zeros of $1 + KG(z)$ have a magnitude less than unity. If the mathematical expression for $KG(z)$ is known, the equation $1 + KG(z) = 0$ can be solved as an algebraic equation. Criteria analogous to Routh’s criterion also exist. Finally, it is possible to go back to the Routh criterion itself by performing the transformation $z = (w + 1)/(w - 1)$, which is equivalent to mapping the outside of the $z$ unit circle in the right-hand half of the $w$ plane. All these algebraic methods, however, become very cumbersome as soon as the complexity of the system increases.

3. **Example.** Consider the system described in Sec. 20.2.5, par. 5:

$$KG(s) = \frac{K}{s(ta + 1)} \quad KG(z) = \frac{Kz(1 - \lambda)}{(z - 1)(z - \lambda)} \quad \lambda = e^{-T/r}$$

$$1 + KG(z) = z^2 + \frac{[K(1 - \lambda) - (1 + \lambda)]z + \lambda}{(z - 1)(z - \lambda)}$$

The system is stable if the roots of

$$Q(z) = z^2 + [K(1 - \lambda) - (1 + \lambda)]z + \lambda = 0$$

have a magnitude less than unity. If the roots are imaginary, they have equal magnitude. The condition for stability is that their product $Q(0) = \lambda$ be less than unity.
This condition is always satisfied. Therefore, the system cannot be unstable if the zeros of \(Q(z)\) are imaginary. When the roots are real, stability demands that they lie between \(-1\) and \(+1\). Hence the conditions
\[
Q(-1) = 2(1 + \lambda) - K(1 - \lambda) > 0 \quad Q(+1) = K(1 - \lambda) > 0
\]
or
\[
0 < K < 2 \frac{1 + \lambda}{1 - \lambda} = \frac{2}{\tanh (T/\tau)}
\]
For example, if the sampling frequency is 4 rad/sec and if \(\tau\) is 1 sec, the condition is \(K < 3.05\). This condition shows that a second-order servo, which is stable regardless of the value of the open-loop gain, can be made unstable by the introduction of a sampler if the sampling frequency is sufficiently low (in this case, if \(T > \tau\)). The tendency of a sampled system to oscillate when its sampling frequency is not high enough is general and can be understood easily. This fact is, indeed, analogous to the effect of a phase lag in the system which compensates for past rather than present errors. This explains the tendency of the system to overcorrect the errors which have accumulated during the interval between samples: as a result, stability can be maintained only by limiting the gain.

20.2.7. Some Other Properties of \(z\) Transforms. 1. Initial- and Final-value Theorems. The following equations relating the initial and final values of a time function to the associated transform can be written:

Initial value:
\[
f(0) = \lim_{z \to \infty} \frac{z - 1}{z} F(z)
\]
Final value:
\[
f(nT) = \lim_{n \to \infty} (z - 1) F(z)
\]
The latter relationship may be used to study the steady-state error of a feedback system and analyze the effect of integration in the loop as with continuous systems (Chap. 15).

2. Frequency Response in Terms of \(z\) Transforms. Consider a sampled system \(H(z)\) subjected to a sinusoidal input \(e(t)\) and suppose we are interested only in the values of the response \(r(t)\) at the sampling instants \(0, T, 2T, \ldots\). Let
\[
e(t) = e^{j\omega t} \quad E(z) = \frac{z}{z - e^{j\omega T}}
\]
and write, for simplicity, \(\chi\) instead of \(\exp(j\omega T)\). Then
\[
R(z) = H(z) \frac{z}{z - \chi} = z \left( \frac{A}{z - \chi} + \frac{a_1}{z - \alpha_1} + \frac{a_2}{z - \alpha_2} + \cdots \right)
\]
where \(A, a_1, a_2, \ldots\) are the residues of \(R(z)\) associated with the pole \(z = \chi\) and the poles \(\alpha_1, \alpha_2, \ldots\) of \(H(z)\), respectively. The \(Az/(z - \chi)\) term is the steady-state solution; the other terms constitute the transient. The residue \(A\) is equal to the value of \(H(z)\) for \(z = \chi\). Hence, the steady state is
\[
R(z) = \frac{zH(\chi)}{z - \chi} \quad r_s(t) = H(\chi)e^{j\omega t}
\]
In other words, $H(\chi)$ defines the amplitude ratio and the phase angle for the envelope of the impulse train.

3. *Alternative Method in the Time Domain.* The response of a linear sampled-data system is given in terms of $z$ transforms by $R(z) = H(z)E(z)$, where $H(z)$ is a polynomial fraction in $z$, or $1/z$, of the type

$$H(z) = \frac{a_0 + a_1z^{-1} + a_2z^{-2} + \cdots}{b_0 + b_1z^{-1} + b_2z^{-2} + \cdots}$$

Hence,

$$b_0R(z) + b_1 \frac{R(z)}{z} + b_2 \frac{R(z)}{z^2} + \cdots = a_0E(z) + a_1 \frac{E(z)}{z} + \cdots$$

Since $z^{-1} = \exp(-sT)$, it can be seen as a consequence of the lag theorem (Sec. 4.2.1) that

$$b_0r(t)i(t) + b_1r(t)i(t - T) + b_2r(t)i(t - 2T) + \cdots = a_0e(t)i(t) + a_1e(t)i(t - T) + a_2e(t)i(t - 2T) + \cdots$$

as

$$i(t) = u_1(t) + u_1(t - T) + u_1(t - 2T) + \cdots$$

$$u_1(t)[bo(0) + u_1(t - T)[b_0e(T) + b_1e(0)] + u_1(t - 2T)[b_0e(2T) + b_1e(0)] + \cdots = u_1(t)[a_0e(0)] + u_1(t - T)[a_0e(T)] + a_1e(0)] + u_1(t - 2T)[a_1e(2T) + a_1e(T) + a_2e(0)] + \cdots$$

Hence,

$$b_0e(0) = a_0e(0)$$

$$b_0e(T) + b_1e(0) = a_0e(T) + a_1e(0)$$

This set of equations gives the values of the output $r(t)$ at the different sampling instants 0, $T$, $2T$, \ldots

This result can be expressed in a general form: Inverse transformation of a linear $z$ equation yields a linear difference equation in the time domain that can be solved step by step. The analogy with continuous linear systems is evident (Table 20-2).

<table>
<thead>
<tr>
<th>Table 20-2.</th>
</tr>
</thead>
</table>
| Ordinary linear systems | Linear differential equation $\leftrightarrow$ algebraic $s$ equation
| Sampled linear systems | Linear difference equation $\leftrightarrow$ algebraic $z$ equation

20.3. **FREQUENCY ANALYSIS OF SAMPLED-DATA SYSTEMS**

20.3.1. *Relationship between Sampled Laplace Transform and Ordinary Laplace Transform.* In the preceding section, methods have been developed for studying the behavior of a sampled-data system at sampling instants. The present section will be concerned with the frequency analysis of the complete time response of the system.

Let us assume that $F(s)$ is a polynomial fraction in $s$. This may mean either that $F(s)$ is the transfer function of a linear system or that $F(s)$ is
the Laplace transform of a time function of appropriate type. \( F(s) \) can be expanded into a sum of partial fractions of the form \( A/(s - a) \), where \( A \) is the residue associated with the pole \( a \) (\( A \) and \( a \) can be real or complex):
\[
F(s) = \frac{A}{s - a} + \frac{B}{s - b} + \frac{C}{s - c} + \cdots
\]

The sampled transform \( F(z) = F^*(s) \) associated with \( F(s) \) is, therefore, a sum of terms of the form
\[
F^*_a(z) = F^*_a(s) = \frac{Az}{z - e^{-(s-a)T}} = \frac{Ae^{eT}}{e^{eT} - e^{aT}} = \frac{A}{1 - e^{-(s-a)T}} + \frac{B}{1 - e^{-(s-b)T}} + \frac{C}{1 - e^{-(s-c)T}} + \cdots
\]

Every term of this sum can be, in turn, expanded into a sum of simple fractions of the form \( \alpha_n/(s - s_n) \), where the \( s_n \) are the poles of \( F^*_a(s) \) and the \( \alpha_n \) are the residues corresponding to these poles. The poles of \( F^*_a(s) \) are given by
\[
e^{-(s-a)T} = 1 \quad \text{that is} \quad -(s-a)T = 2\pi n j
\]
where \( n \) is an integer that can be positive, negative, or zero. If the "sampling angular frequency" \( 2\pi/T \) is denoted as \( \Omega \), the poles of \( F^*_a(s) \) are
\[
s_n = a - jn \Omega \quad \text{with} \quad n = 0, +1, -1, +2, -2, \ldots
\]
and their number is infinite.

The residue associated with the pole \( s_n \) is
\[
\alpha_n = \left. \frac{(s - a + jn \Omega)F^*(s)}{1 - \exp[-(s-a)T]} \right|_{s=a-jn \Omega}
\]
\[
+ \left. \frac{B(s - a + jn \Omega)}{1 - \exp[-(s-b)T]} \right|_{s=a-jn \Omega} + \cdots
\]

In this expression all the terms are zero except the first one, which is equal to
\[
A \left. \frac{\partial (s - a + jn \Omega)/\partial s}{\partial [1 - \exp[-(s-a)T]]/\partial s} \right|_{s=a-jn \Omega} = \frac{A}{T}
\]
Hence,
\[
F^*(s) = \sum_{n = -\infty}^{\infty} \frac{A}{T} \frac{1}{(s - a + jn \Omega)} + \frac{B}{1 + e^{-(s-b)T}} + \frac{C}{1 + e^{-(s-c)T}} + \cdots
\]

In a similar manner, all the partial fractions associated with the poles \( b, c, \ldots \) can be expanded in series, so that
\[ F^*(s) = \frac{1}{T} \sum_{n=\infty}^{+\infty} \frac{A}{s - a + jn\Omega} + \frac{B}{(s - b + jn\Omega)} + \frac{C}{(s - c + jn\Omega)} + \ldots \]

Finally,

\[ F^*(s) = \frac{1}{T} \sum_{n=\infty}^{+\infty} F(s + jn\Omega) \]  

(20-13)

Equation (20-13) relates the sampled Laplace transform to the ordinary Laplace transform. It may be recalled that Eq. (20-2), which is rewritten here as

\[ F^*(s) = \sum_{n=0}^{\infty} f(nT)e^{-nTs} \]

relates the sampled Laplace transform to the *time function*.

This result can be obtained in a more condensed and abstract way by considering the convolution integral (Sec. 4.4.5). Indeed if \( f(t) \) is the inverse Laplace transform of \( F(s) \),

\[ f^*(t) = f(t) \delta(t) \quad F^*(s) = F(s) * I(s) = F(s)^* \frac{1}{1 - e^{sT}} \]

where the asterisk means convolution in the \( s \) domain. Evaluation of the convolution integral applying the residue theorem yields Eq. (20-13).

*Note.* The proof given cannot be considered as rigorous because of the convergence problems associated with infinite series. In fact, \( F^*(s) \) may differ from

\[ \frac{1}{T} \sum_{n=\infty}^{+\infty} F(s + jn\Omega) \]

by an entire function of \( s \) (in particular a constant). For example, if \( F(s) = 1/s \), one should have

\[ F^*(s) = \frac{1}{T} \sum_{n=\infty}^{+\infty} \frac{1}{s + jn\Omega} + \frac{1}{2} \]

Equation (20-13) is, however, rigorous when \( F(s) \) approaches zero for infinite \( s \) as quickly as \( 1/s^m \) with \( m \geq 2 \).

**20.3.2. Consequence.** It results from Eq. (20-13) that \( F^*(s) \) is a periodic function of \( s \) with a period \( j\Omega \): for any value of \( s_1 \), \( F^*(s_1 + j\Omega) \) is equal to \( F^*(s_1) \). In the \( s \) plane \( F^*(s) \) repeats itself at the corresponding points \( s_1, s_1 + j\Omega, s_1 + 2j\Omega, \ldots \) in successive strips parallel to the real axis (Fig. 20-21). Furthermore, \( F^*(s) \) is symmetrical with respect to
the real axis. Therefore, $F^*(s)$ is completely determined by its values in the half strip delimited by the horizontal lines $\omega = 0$ and $\omega = \Omega/2$.

20.3.3. Sampled-data Filtering. Consider the behavior of $F^*(s)$ along the imaginary axis which is associated with the harmonic approach. If the amplitude spectrum $|E(j\omega)|$ of the input $e(t)$ (not sampled) is all contained within the band $(-\Omega/2, +\Omega/2)$, the amplitude spectrum

![Diagram](attachment:image1)

**Fig. 20-21.** Periodicity of $F^*(s)$. Knowledge of $F^*(s)$ in the half strip $(0,\Omega/2)$ fully determines the function.

![Diagram](attachment:image2)

**Fig. 20-22.**

$E^*(j\omega)$ of the output to a theoretical sampler consists of a primary component identical with $|E(j\omega)|/T$ within the band $(-\Omega/2, +\Omega/2)$ plus complementary components due to the periodicity at $\Omega$ (Fig. 20-22). Thus sampling appears as a particular case of modulation (amplitude modulation by train of impulses) which introduces high-frequency components in a periodical manner.

The former situation occurs when $\Omega/2$ is large enough compared with the frequency components of the signal, that is, if $T$ is sufficiently small
compared with the time scale of input variations, the sampling frequency \( \Omega \) being at least twice the highest input frequency \( \omega_1 \). If this is the case, all information after sampling can be gained by an ideal low-pass filter that should have a constant amplitude ratio \( \omega < \omega_1 \) and zero amplitude ratio for \( \omega > \Omega/2 \) (Fig. 20-23).

When the amplitude spectrum of the input extends beyond \( \Omega/2 \), the situation is more complicated as the sampled signal spectrum becomes an entangled combination of the overlapping primary and secondary components. In particular, within the interval \((-\Omega/2, +\Omega/2)\), \( |E^*(j\omega)| \) differs from \( |E(j\omega)|/T \). If the highest input frequency \( \omega_1 \) is slightly greater than \( \Omega/2 \), the information that could be gained by linear filtering is contained in a frequency band somewhat less than \( \Omega/2 \). As seen previously, in most control systems\(^1\) a filtering of high frequencies is performed by the components themselves between the sampler and the output: motor, amplifier, and smoothing circuits. The components often include a holding system, where the value of the sampling pulse is held constant until the arrival of the next pulse. From an information-filtering standpoint, the optimum transfer function for the combination of all these components is that of an ideal low-pass filter with a cutoff frequency cutting off in the neighborhood of \( \Omega/2 \), as shown in Fig. 20-23. It is obvious that an ideal cutoff at frequency \( \Omega/2 \) would, in practice, introduce phase lags in the vicinity of \( \Omega/2 \); these phase lags generally impair the stability of the system (when \( \Omega/2 \) is included in the passband of the system, or even if it is not very far from it). The design of smoothing circuits for sampled data in servo problems therefore involves a typical compromise between stability and filtering requirements, this compromise being more in evidence in proportion as the sampling frequency is lower.

The filtering effect obtained by the holding device can be described by the frequency characteristic associated with the device. It can be shown that the transfer function of the holding device \( H(s) = (1 - e^{-Ts})/s \) yields

\[
|H(j\omega)| = \frac{2\pi}{\Omega} \left| \frac{\sin (\pi \omega/\Omega)}{\pi \omega/\Omega} \right|
\]

\(^1\) This is not the case for all-digital computers, where each component operates on samples.
The function $|H(j\omega)|$ is sketched in Fig. 20-24. It is apparent that this device does not cut off sharply (the gain is down only 36 per cent at $\Omega/2$. Furthermore, the phase shifts rather rapidly).

20.3.4. Sampled Transfer Loci. Knowing the transfer locus of a linear system $H(j\omega)$, Eq. (20-13) enables the locus of the function $H^*(j\omega)$ to be drawn. This locus characterizes the behavior of the system for sampled inputs and may be called the sampled transfer locus of the system. The sampled transfer locus can be obtained graphically from the ordinary transfer locus, either from the $H(j\omega)$ locus itself (Fig. 20-25) or from the separate plots of its real and imaginary parts $U(\omega)$ and $V(\omega)$ (Fig. 20-26).

In practice, because of the filtering out of high frequencies performed by most transfer functions, the summation converges rapidly and can practically be limited to $n = 0, 1, 2, -1, -2,$ or even, in many cases, to $n = 0$ and $-1$. This is seen graphically by the fact that the vectors (or phasors) $H[j(\omega + n\Omega)]$ are negligible when $n$ exceeds 2 or 3. When it is permissible to restrict oneself to $n = 0$ and $-1$, a sketch of $H^*(j\omega)$ is very easily drawn and lends itself to qualitative discussions. The construction summarized in Fig. 20-27 is based on the equation

$$TH^*(j\omega) \cong H(j\omega) + H[j(\omega - \Omega)] = H(j\omega) + \bar{H}[j(\Omega - \omega)]$$

where $\bar{H}$ is the complex conjugate of $H$.

The following properties of sampled transfer loci are evident:

a. At zero frequency the sampled locus starts from the real axis, since

$$TH^*(j0) = U(0) + 2 \sum_{n=1}^{\infty} U(n\Omega)$$

where $U = \text{Re } H$. However, if $H(s)$ has integration, the sampled locus starts from infinity.

b. At the frequency $\omega = \Omega/2$ the sampled transfer function is real and the sampled locus crosses the real axis, since

$$H^* \left( j \frac{\Omega}{2} \right) = \ell^* \left( \Omega \frac{\Omega}{2} \right)$$
where

\[ TU^*(\omega) = \sum_{n=-\infty}^{\infty} U(\omega + n\Omega) \]

c. As \( \omega \) increases above \( \Omega/2 \), the \( H^*(j\omega) \) point traces out a curve symmetric with the part of the locus that corresponds to \( 0 < \omega < \Omega/2 \), and for \( \omega = \Omega \) the starting point \( \omega = 0 \) is reached. Then for \( \omega > \Omega \) the original portion is traced out again, so that the sampled locus appears as a closed curve which \textit{periodically repeats itself} with the period \( \omega = \Omega \).

Fig. 20-26. Obtaining sampled transfer locus from separate plots of \( U(\omega) \) and \( V(\omega) \).
d. When the sampling frequency is high, $|H(\Omega/2)|$ and $|H(\Omega)|$ are usually very small. Hence

$$TH^*(j\omega) \cong H(j\omega)$$

and the sampled locus does not differ greatly from the ordinary transfer locus. Figures 20-27 and 20-28 show two typical aspects of sketches of sampled transfer loci. They correspond to different shapes of the transfer locus and to different locations of $\Omega/2$ and $\Omega$ on its graduation.

20.3.5. Generalization of Nyquist’s Criterion. Nyquist’s criterion can be applied to sampled open-loop transfer loci. This can be proved in exactly the same manner as in Sec. 16.1 for ordinary open-loop transfer loci. The only difference is that, in the case of sampled loci, integration is performed in the $s$ plane along a closed path (Fig. 20-29) consisting of:

1. The imaginary axis from $-\Omega/2$ to $\Omega/2$
2. The horizontal line $\text{Im} \ s = \Omega/2$ in the right-half plane
3. A portion of the circle of “infinite” radius centered at the origin
4. The horizontal $\text{Im} \ s = -\Omega/2$ back to the imaginary axis

It results from the property stated in Sec. 16.1.2 that the phase variation of $1 + KG^*(s)$ when the $s$ point traces out the contour is equal to
$2\pi(P - Z)$, where $P$ and $Z$ are the numbers of the poles and zeros of $1 + KG^*(s)$ that lie inside the contour, i.e., in the horizontal $(-\Omega/2, \Omega/2)$ strip. Note that $P$ is also the number of the poles of $KG(s)$ that lie in the strip. Now it is easy to evaluate the phase variation of $1 + KG^*(s)$. The contribution of the infinite circle (3) is zero, because of the behavior of $KG^*(s)$ for infinite $|s|$. Furthermore, the contributions of the horizontal portions (2) and (4) are opposite because of the periodicity of $1 + KG^*(j\omega)$ with the period $\Omega$; therefore they cancel each other. Thus the over-all phase variation merely consists of the contribution (1) of the portion of the imaginary axis. It is the phase variation of the $1 + KG^*(j\omega)$ point as $\omega$ is varied from $-\Omega/2$ to $+\Omega/2$, that is, $2\pi$ times the number of counterclockwise revolutions accomplished by the $KG^*(j\omega)$ locus around the $-1$ point when $\omega$ is varied from $-\Omega/2$ to $+\Omega/2$ or, equivalently, from 0 to $\Omega$. If $N$ is this number of revolutions, then $P - Z = N$. The condition for stability $Z = 0$ becomes $N = P$. In other words, the $KG^*(j\omega)$ locus traced out in the direction of increasing $\omega$ from 0 to $\Omega$ should encircle the critical $(-1)$ point a number of times equal to the number of unstable poles of the open-loop system. In particular, for an open-loop stable system, the $KG^*(j\omega)$ locus should not encircle the $(-1)$ point.

Once the sampled locus has been drawn, the practical application of the criterion is identical with that of the Nyquist criterion for an ordinary transfer locus. The reader is, therefore, referred to Sec. 16.1.4. In particular, difficulties arise when the function $KG(s)$ has poles on the imaginary axis, the most frequent case being that in which $KG(s)$ has integration, i.e., a pole at the origin. In that case the integration contour should include small semicircles with the poles at their left (Fig. 20-30). Each of these circles will contribute by $-m\pi$ to the phase variation of $1 + KG^*(j\omega)$, $m$ being the order of multiplicity of the corresponding pole. Thus the presence of one integration will cause the $KG^*(j\omega)$ locus to be infinite for zero $\omega$ and to “jump” from $0^-$ to $0^+$ by turning clockwise by $\pi$ on a half circle of infinite radius. The situation is identical with that of the ordinary Nyquist criterion (Fig. 16-9).

The application of Nyquist's criterion to sampled loci allows the stability of sampled-data servo systems to be easily determined. In many practical cases a sketch of the sampled locus, drawn as explained in Sec. 20.3.4, is

![Fig. 20-29.](image-url)
sufficient, the stability being determined by the position of the \( \omega = \Omega/2 \) point with respect to the critical \(-1\) point.

**20.3.6. Generalization of Stability Margins.** The concept of a resonance ratio and the use of Hall’s chart (Sec. 13.2.1) can be extended to sampled-data servos, the tangency of the sampled locus \( KG^*(j\omega) \) to a certain \( Q \) circle ensuring a certain damping. However, this gives a guarantee only so far as the damping of the output \( r_s(t) \) is concerned; that is, there is no control of the behavior of the actual output \( r(t) \) between the sampling instants. Simultaneous consideration of the ordinary and sampled loci, \( KG(j\omega) \) and \( KG^*(j\omega) \), respectively, allows conclusions to be drawn concerning \( r(t) \) between sampling instants.

**20.3.7. Extension of the Root-locus Method.** The root-locus method can be extended to sampled servos. In the \( s \) plane, one has to consider the locus of the zeros of \( 1 + KG^*(s) \) when \( K \) varies. It must be borne in mind that the poles and zeros of \( G^*(s) \) are infinite in number; the root locus is periodic with the same period \( \Omega \) as the function \( G^*(s) \). An approximation can be made in plotting the root locus by considering only a few of the infinite number of poles and zeros of \( G^*(s) \).

The root locus can also be considered in the \( z \) plane, where the number of poles and zeros of \( G(z) \) is limited. Figure 20-31 shows the root-locus plot in the \( z \) plane associated with the function \( G(z) = K/s(1 + \tau s) \). The function \( G(z) \) has a zero at the origin and has poles at \( z = 1 \) and \( z = 1 + \exp(-T/\tau) \). Instability occurs when the root locus crosses the unit circle.

**20.4. REMARKS ON SAMPLED-DATA SERVO SYNTHESIS**

**20.4.1. General.** The problem of synthesizing sampled-data systems may be approached by an extension of conventional methods. This problem is difficult when the system is to be designed for a given continuous output, and it is then necessary to resort to approximations.
The most common problem in designing sampled-data servomechanisms is that of selecting a basic sampling rate and the filtering and compensating networks. It has been seen in Sec. 20.3.3 that the sampling rate can be determined by the bandwidth required. After the fixed parts of the system have been selected, the filtering and compensating networks are chosen by trial-and-error procedures in order to approximate the desired response characteristics. In this procedure the conditions for physical realizability must be taken into account.

When the sampling rate is given, and if one is only interested in the response at the sampling instants, the design of the system is simpler and can be considered in terms of z transforms. Direct synthesis procedures can then be used, as will briefly be illustrated in the following examples, adapted from the work of J. Truxal.

20.4.2. First Example. Suppose the transfer function associated with an error-sampling component in a servo loop is given as

$$KG(s) = \frac{K}{s(1 + \tau s)}$$

The problem consists in synthesizing a compensating network so that the response to a unit-step input is given at two sampling instants: $r(T) = 1$ and $r(2T) = 1.2$, with the condition that $r(\infty)$ should be equal to 1.

Since there is no steady-state error, it is natural to try an open-loop transfer function with integration: say, $a/s(1 + bs)$, in which the two coefficients $a$ and $b$ are to be determined by the two conditions for the time response. The $z$ transform of this response is

$$R(z) = \frac{a(1 - \mu)z}{z^2 + [a(1 - \mu) - (1 + \mu)]z + \mu z - 1}$$

where, for brevity, $\mu$ has been written for $\text{exp}(-T/b)$.

The steady-state value, obtained from the final-value theorem (Sec. 20.2.7), is

$$r(\infty) = \lim [(z - 1)R(z)] = \frac{a(1 - \mu)}{1 - a(1 - \mu) - (1 + \mu) - 1} = 1$$

The values of $r(t)$ at the first sampling instants are given by the expansion of $R(z)$ in inverse powers of $z$:

$$R(z) = \frac{1}{z} a(1 - \mu) + \frac{1}{z^2} [-a(1 - \mu)[a(1 - \mu) - (1 + \mu) - 1]] + \cdots$$

Identification with $1/z + 1.2/z^2 + \cdots$ yields

$$a(1 - \mu) = 1 \quad 1 + \mu = 1.2$$

whence $a = 1.25$ and $b = 0.62T$. Thus the network to be inserted in the forward path must have a transfer function $F(s)$ such that

$$\frac{K}{s(1 + \tau s)} F(s) = \frac{a}{s(1 + bs)}$$

$$F(s) = \frac{a}{K} \frac{1 + \tau s}{1 + bs} = \frac{1.25}{K} \frac{1 + \tau s}{1 + 0.62Ts}$$
20.4.3. Second Example. Prediction can be effected on the basis of a certain number of the previous samples. Consider now a sampled-data servo. The $z$ transform of the response to a given input $E(z)$ can be expanded in a power series in $1/z$:

$$R(z) = \frac{KG(z)}{1 + KG(z)} E(z) = \left( a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots \right) E(z)$$

where, in general, $a_0$ is zero (as a consequence of the presence of inertia).

In terms of the values of the time functions at the sampling instants, this can be written as

$$r(nT) = a_1e[(n - 1)T] + a_2e[(n - 2)T] + \cdots$$

Therefore, the servo operates as a predictor on the basis of previous sampling values. Consider, for example, prediction of the linear extrapolation type (Fig. 20-32),

$$r(nT) = e[(n - 1)T] + [e(n - 1)T - e((n - 2)T)] = 2e[(n - 1)T] - e[(n - 2)T]$$

This type of prediction can be effected by the servo considered, provided

$$\frac{KG(z)}{1 + KG(z)} = \frac{2}{z} - \frac{1}{z^2}$$

that is

$$KG(z) = \frac{2z - 1}{(z - 1)^2}$$

If a holding device is to be included, then

$$KG(s) = \frac{1 - \exp(-Ts)}{s} \quad K'G'(s) = (1 - e^{-Ts})F(s)$$

where $K'G'(s)$ is the open-loop transfer function to be determined. In terms of $z$ transforms, the $1 - \exp(-Ts)$ factor is equivalent to multiplication by $(z - 1)/z$. Hence $F(s)$ must satisfy

$$\frac{z - 1}{z} F(z) = KG(z)$$

Identification yields

$$F(s) = \frac{1/T^2}{s^2} + \frac{3T/2}{s}$$
CHAPTER 21
MULTIPLE FEEDBACK CONTROL SYSTEMS

Summary
1. Definition. Examples.
3. Introduction to a more general theory.

21.1. DEFINITION, EXAMPLES

Earlier, feedback control systems with one variable were considered and were defined as feedback systems designed to ensure the control, with an amplification of power, of an output $r$ by an input $e$. The block diagram of such a system is shown in Fig. 21-1. The blocks $H$, $A$, and $J$ represent, respectively, the transfer functions of the compensating network, the

![Block diagram of a one-variable servo system.](image)

power and amplification stage, and the feedback loop. The feedback control system attempts to satisfy at all times the relationship

$$ e = r' \quad \text{that is,} \quad e = 0 $$

This is a differential equation of the general form

$$ F(e, r) = 0 $$

21.1.2. Definition of a Feedback Control System with Several Variables. A feedback control system with several variables, or a multiple feedback control system, is defined as a system which forms $n$ output variables ($r_1, \ldots , r_n$) from $m$ input variables ($e_1, \ldots , e_m$) to attempt to satisfy $n$ equations of the form:

$$ F_1(e_1, \ldots , e_m, r_1, \ldots , r_n) = 0 \quad \ldots \quad F_n(e_1, \ldots , e_m, r_1, \ldots , r_n) = 0 $$

This system of input-output equations is written:

$$ F_i(e, r) = 0 $$

In this case the feedback control system detects and amplifies the $n$ errors $e_i$, which are functions of the $m$ inputs and $n$ outputs. These errors act upon the $n$ outputs so as to cancel themselves and also the
functions $F_{i}(e,r)$. This definition is represented in the block diagram of Fig. 21-2. In this diagram $F$ is the sensing device, $H$ generalizes the compensating network, and the $A_{i}$ are the power servocomponents.

It is necessary to suppose that the equations $F_{i}(e,r)$ define the outputs $r$ as single-valued functions of the inputs. On the other hand, in what follows the linear case especially will be considered. The components $H$, $J$, and $A$ are then mathematically defined by transfer matrices (Sec. 10.1) such as

$$
\| \epsilon' \| = \| H \| \cdot \| \epsilon \|
$$

\textbf{21.1.3. The Components.} Very often the sensing device $F$ can be depicted as in Fig. 21-3, where the error-sensing devices $e$, are explicitly shown. For this special case the equations $F(e,r)$ have the form:

$$
F(e,r) = U(e) - V(r) = 0
$$

where the $U$ and $V$ are functions of $s$.

The component $H$ has a purpose similar to that of the compensating network in a feedback control system with one variable, i.e., to stabilize and improve the performance of the system. For the linear case, using the matrix notation,

$$
\begin{pmatrix}
\epsilon'_{1} \\
\vdots \\
\epsilon'_{n}
\end{pmatrix}
= 
\begin{pmatrix}
H
\end{pmatrix}
\begin{pmatrix}
\epsilon_{1} \\
\vdots \\
\epsilon_{n}
\end{pmatrix}
$$
Examples of $H$ Components. If $H$ is the unit matrix, one has $\varepsilon_i = \varepsilon'_i$, which is the same as eliminating the component $H$. If $H$ is a permutation matrix of order $n$, the component $H$ performs the operation of crossing connections. This crossing of connections corresponds to the determination of which errors determine which outputs. It will be shown in Sec. 21.3.3, par. 3, that there is an optimum connection.

21.1.4. Examples of Multiple Feedback Control Systems. 1. Frequency/Power Control in a Power Distribution Network. The power output of the generators must be regulated according to some program, and the electric frequency must be maintained at some constant value. These controls must be performed for reasons of economy of generator operation and to ensure the proper functioning of synchronous machines. The inputs of such a control system are of two types. The first consists of random disturbances, which are usually load variations; the second consists of the prescribed inputs, which are the operating frequency and the program of power to be furnished by the generators.

2. Transformation of Coordinates. The control system is to calculate automatically, for instance, the spherical coordinates $\rho, \theta, \psi$, from the cartesian coordinates $x, y, z$. The relations to be satisfied are

$$x - \rho \cos \theta \cos \psi = 0 \quad y - \rho \cos \theta \sin \psi = 0 \quad z - \rho \sin \theta = 0$$

A classical electromechanical device, using two components called "resolvers," works out, from the voltages $x, y, z$ and from the rotations $\theta$ and $\psi$, the deviation terms

$$\varepsilon_1 = x \sin \theta - y \cos \theta$$
$$\varepsilon_2 = (x \cos \psi + y \sin \psi) \sin \theta - z \cos \theta,$$

which are to drive motors whose shaft angles are $\psi$ and $\theta$. The cancellation of $\varepsilon_1$ and $\varepsilon_2$ entails the verification of the proposed relations. In fact, the third relation is not independent, since $x^2 + y^2 + z^2 = \rho^2$. The quantity $\rho$ is directly given as a voltage by one of the resolvers.

3. Execution, in Space, of an Imposed Trajectory by an Airplane with Human or Automatic Pilot. In the case of the human pilot flying an airplane, the pilot has the role of the $F$ and $H$ and partially of the $A_i$ blocks of Fig. 21-2. He senses the flight deviations and makes the necessary corrections. The other part of the $A_i$ blocks would correspond to the controls and dynamics of the airplane. The coupling in the present case is well known (yaw-roll).

4. Analog Computers. Example 2 was a special case. Because of the considerable importance of their applications, the following section will be exclusively devoted to electronic analog computers.

21.2. ANALOG COMPUTERS AND SIMULATORS

21.2.1. Introduction. The multiple feedback control system of Fig. 21-2, which produces outputs $r_1, \ldots, r_n$ from inputs $e_1, \ldots, e_n$, can be considered as a calculating machine used to solve the system of equations:

$$F_1(e, r) = 0 \cdots F_n(e, r) = 0$$

1 The purpose of this section is merely to show the close connecting links that exist between analog computers and multiple servo systems. For a more complete theory of analog computers see, for example, G. A. Korn and T. M. Korn, "Electronic Analog Computers," 2d ed., McGraw-Hill, New York, 1956.
If the equations $F_j$ are those of a physical system, the multiple feedback control system is an analog, or simulator, of the physical system in question.

21.2.2. Notions on Analog Computers. Analog computers are widely used nowadays to study physical systems whose equations are long and complicated, or of which experimental studies are expensive. There are two reasons for including them in a study of feedback control systems:

1. They are servomechanisms with several variables.
2. They represent an indispensable tool for the analysis and synthesis of complex feedback control systems.

21.2.3. Structure of an Analog Computer. 1. General. Actually, the type of scheme of Fig. 21-2 would hardly be applicable to universal use. Therefore, components which perform elementary operations such as addition and integration were designed. These components are placed after each other in a multiple loop which simulates differential equations. There are also analog computers for problems with nonlinear and partial differential equations such as heat transfer, flow, and flutter. The principles of the operation of the elementary component of an analog computer for linear differential equations with constant coefficients is described in the next paragraph.

2. Principle of the Operational Amplifier. The operational amplifier is the main building block of analog computers. Such an element is shown in Fig. 21-4, where $A$ is an electronic amplifier, $Z_1(s)$ and $Z_2(s)$ are complex impedances, $r(t)$ and $-r(t)$ are respectively input and output signals. By applying Kirchhoff's law at $M$, it can be shown that

\[
 Z_2(s)
\]

which in admittance form is written as

This relation is exact if (a) the amplifier gain is infinite and (b) the grid current is zero. In general the amplifiers used have a gain superior to $10^4$ and are able to reach $10^7$. 
Thus it can be seen that, from an amplifier $A$ with high gain, it is possible to build a network of which the transfer function can have a desired expression $H(s)$ by connecting an admittance $Y_1(s)$ in series and an admittance $Y_2(s)$ in parallel so that

$$\frac{Y_1(s)}{Y_2(s)} = H(s)$$

**Remark.** The minus sign at the output which depends on the number of amplification stages present will be omitted in what follows.

![Fig. 21-5.](image)

![Fig. 21-6.](image)

![Fig. 21-7.](image)

3. **Examples.**
   a. Construct the transfer function $H(s) = K$. Take

   $$Y_1 = K \quad Y_2 = 1$$

   This results in a circuit as in Fig. 21-5 (where $k = 1/K$).
   b. Construct the transfer function $H(s) = a + bs$. Take

   $$Y_1 = a + bs \quad Y_2 = 1$$

   Thus Fig. 21-6 is obtained.
   c. More generally, the scheme of Fig. 21-7 yields the equation

   $$R(s) = (a + bs)E_1(s) + dE_2(s)$$

4. **Application:** Electronic Simulation of an Airplane. As an example, the above technique will be used to build an electronic computer representing the equations of
longitudinal movement, established in Sec. 7.5.2, of an airplane flying with constant speed.

If $i$ = angle of attack
$q = d\theta/dt$ pitching rate
$\delta =$ control-surface deflection

the equations are, in terms of the Laplace transforms $I$, $Q$, $\Delta$,

$$\frac{K_t}{mV} \Delta - \left( s + \frac{K_t}{mV} \right) I = JQ$$
$$lK_\delta \Delta - \left( s + \frac{M}{J} \right) Q = -hK_\delta I$$

They are simulated respectively by the two schemes represented in Fig. 21-8. If the two terminals $q$ and the terminals $i$ are joined (taking the signs into consideration), an electronic simulator of the airplane has been built, permitting one, for instance, to study its response to a control-surface deflection which varies with time by observing $i$ or $q$ on an oscilloscope or recorder.

![Electric analog of airplane in longitudinal motion](image)

**Fig. 21-8.** Electric analog of airplane in longitudinal motion (Sec. 7.5.2).

5. **Problems Arising from the Use of Analog Computers.** The preceding outline of the very simplified principle gives little idea of the real difficulties arising in the design of an analog computer. Largely the problems are of two natures:

1. The problems of technology for different components, and especially for the amplifiers (drift, noise, etc.)
2. Theoretical problems concerning the stability and accuracy of the computer

These problems are not yet solved generally. The most powerful tool that has been used to solve them is the theory of servomechanisms as applied to multiple feedback control systems. The first work in this direction was done by F. H. Raymond in 1949. A few words will be
said about it later, in Sec. 21.3.3. The mathematical application of the theory of multiple servomechanisms does not, however, eliminate methodical experimental studies.

6. Interest in, and Use of, Analog Computers and Simulators. At the present time analog computers are indispensable auxiliaries in the study and synthesis of servomechanisms as soon as complex problems arise. In recent years their development has been most spectacular in aeronautics because of the support given by different countries to research on military problems in this field.

To reproduce as clearly as possible the conditions of the actual flight, a partially electronic and partially mechanical simulation is often required. A mechanical part, which is able to rotate (with one or several degrees of freedom) and is called a simulator, is introduced into the loop and slaved to the output variables calculated by the electronic stage. The simulator section contains actual components which are to be mounted on the airplane.

Flight simulators involve difficult problems. They must be stiff enough to avoid vibrations; on the other hand, they are expected to follow the outputs of the electronic stages with extremely low time lags, which frequently results in critical compromises. Furthermore, if actual components are used in the simulator section, the electronic stages must necessarily work in real time, which eliminates one of the advantages of analog-computer studies, that is, changing the time scale.

21.3. SOME REMARKS ON GENERAL THEORY

21.3.1. Assets of a General Theory on Multiple Feedback Control Systems. The study of certain multiple servomechanisms can be simplified to the study of several feedback control systems with one variable. If one can eliminate \( n - 1 \) outputs and obtain a unique equation defining one output \( r \), from \( m \) inputs, the study of the \( n \) equations obtained in this manner, one for each output, allows the performance to be analyzed. But this procedure sometimes cannot be applied, because the eliminations cannot be made, or because the resulting equations are complicated and hide the physical roles of various elements, especially of the control parameters.

Thus the advantage of a general theory of multiple feedback control systems is to give (a) the general structural conditions for servomechanisms and (b) the general methods for the study of a particular system in order to obtain stability and good performance from the system.

Unfortunately, at this time very few completely general results have been found by mathematical methods. The next section indicates a few of the more interesting results. They all pertain to linear multiple servomechanisms.

21.3.2. General Equations. Consider the multiple servo system of Fig. 21-2, in which (a) the components \( H, A_i \), and \( J \) are linear and (b) the \( F \) component can be so separated into the linear components \( U \) and \( V \)

\footnote{This section and the following suppose some knowledge of matrices.}
(Fig. 21-3) that the matrix relation

\[ V(R) - U(E) = 0 \]

is satisfied. For the case of zero initial conditions, the following fundamental relations are obtainable

\[ R = (I_n + AHJV)^{-1} AHU \cdot E \]
\[ E = (I_n + AHJV)^{-1} U \cdot E \]

where \( I_n \) is the unit matrix.

These relations generalize the fundamental equations of Chap. 13, which are valid for the feedback control system with one variable of Fig. 21-1. For the single-variable case, \( U \) becomes 1 and \( V \) becomes 1, and the general equations simplify to

\[ R = \frac{AH}{1 + AHJ} E \quad \text{and} \quad E = \frac{1}{1 + AHJ} E \]

from which the influence that the gain of the amplifiers has on accuracy is immediately read.

21.3.3. Study of Stability. 1. General. The characteristic equation of the servo system with one variable, \( 1 + AHJ = 0 \), becomes in the case of the multiple feedback control system

\[
\text{Determinant } (I_n + AHJV) = 0
\]

The necessary and sufficient condition for stability is that the real part of the roots of this characteristic equation must be negative. Stabilization can be performed, depending upon the case, by either altering \( H, J \), or both or by adding secondary loops. Unfortunately, it is difficult to express, in the general case, conditions for stability in conveniently appropriate rules. A few special cases follow, where the obtained conditions are easy to interpret and to apply.

2. First Case. Suppose that

\[ V = I_n \quad H = I_n \]

This is a trivial case, corresponding to \( n \) independent feedback control systems. The characteristic equation becomes \( n \) classical equations

\[ 1 + a_i(s) = 0 \]

3. Second Case.\(^1\) The following assumptions are made:

\( a: \quad V = V_0 \quad \text{numerical matrix, i.e., independent of } s \)
\( b: \quad H = H_0 \quad \text{numerical matrix} \)
\( c: \quad A(s) = a(s)I_n \quad \text{all amplifiers are identical} \)

\(^1\) This case was studied by F. H. Raymond, "Introduction à l'étude des asservissements multiples simultanés," *Bulletin de la Société Française des Mécaniciens*, no. 7, pp. 18-25, 1953.
The characteristic equation becomes:

\[ \text{Determinant} \left[ I_n + a(s)H_0V_0 \right] = 0 \]

Therefore, if \( \mu \), designates the characteristic values of the matrix \( H_0V_0 \), its roots are solutions of the \( n \) algebraic equations

\[ \mu_1a(s) + 1 = 0 \quad \cdots \quad \mu_na(s) + 1 = 0 \]

We are then led to study separately the stability of the roots of these \( n \) equations. This can be done by solving the equations or by applying numerical or graphical methods—e.g., by use of Routh's or Nyquist's criterion—the critical point lying at \(-1/\mu_i\). This again entails the study of \( n \) feedback control systems with one variable, but in this case the \( \mu_i(s) \) must be calculated. In simple cases, such as \( n = 2 \) or \( 3 \), the effect of \( H_0 \) on the \( \mu_i \) can be seen.

In particular, if \( a(s) \) has a positive real part when the real part of \( s \) is \( \geq 0 \) (which is unfortunately a very restricted hypothesis), the multiple servomechanism is stable with the sufficient condition that \( H_0V_0 \) be positive definite. The existence of this condition has led to the search for sufficient conditions for which a given matrix \( M \) would be positive definite. One of these conditions, as stated by M. Parodi, is that

\[ m_{ii} > \sum_j |m_{ij}| \quad j = 1, 2, \ldots, (i - 1), (i + 1), \ldots, n \]

which expresses the fact that the matrix \( M \) is close to \( I_n \) in that its diagonal terms are large compared to the other terms situated in their respective rows and columns.

If it is realized that the nondiagonal terms of the matrix correspond to the couplings between various systems with one degree of freedom, which latter are represented by the diagonal terms, one sees that Parodi's condition amounts to the fact that, since systems with one degree of freedom are stable by themselves, a more complex system is stable when the couplings between its component systems—i.e., a possible source of destabilization—are sufficiently weak.

4. Third Case. Suppose that (a) \( U \) and \( V \) are unit matrices,

\[ U = V = I_n \]

and (b) \( H \) is a numerical matrix \( H_0 \) (i.e., its coefficients are independent of \( s \)) and its diagonal terms are unity. In this case we are dealing with a multiple servo system consisting of \( n \) one-variable servos with unity ideal transfer functions. These \( n \) servos are coupled and it is possible to take advantage of the coupling terms to gain in simplicity and precision.\(^1\)

5. Performance Criteria. Position constants, velocity constants, etc., may be defined for multiple feedback control systems in a similar manner as for the single-variable feedback control systems. The expressions for

\(^1\) This case has been studied by M. Golomb and E. Uusin, "A Theory of Multidimensional Servo Systems," \textit{J. Franklin Inst.}, \textbf{258}(1): 28–57 (1952).
the position error, the velocity error, etc., may be found in the above reference by Golomb and Usdin.

The response time can also be discussed. It is found to be characterized by the root with the largest real part. The integral-squared error criterion can be applied to multiple feedback control systems.

6. Role of the Theory in Its Present Stage. At the present time, at least to our knowledge, no complete theory of multiple feedback control systems exists. Two facts should be emphasized:

a. It is necessary in practice, when applying mathematical methods numerically, to resort to computing machines, because of the length and complexity of the calculations.

b. The help given by the experimental methods is important. They presume the existence of a dummy, or at least of subassemblies. They are patterned after the ones used for the feedback control systems with one variable, using simple types of input and decoupling the loops each time it is possible. Simplified realizations of the system under study can also be constructed, neglecting some of the less important coupling terms. Thus the general characteristics of the system may be studied.
PART THREE

NONLINEAR SERVO SYSTEMS

CHAPTER 22

GENERAL REMARKS ON NONLINEAR SYSTEMS

Summary

1. General remarks on linearity and nonlinearity.
2. Definition and classification of nonlinear systems.
3. Cases of nonlinearity of servo systems.

22.1. GENERAL REMARKS ON LINEARITY AND NONLINEARITY

22.1.1. The Role of Linear Systems. The first two parts of this work have dealt with the general properties of linear systems, applying the results obtained to cases of servo systems and singling out general methods (transfer-function approach) and the techniques of their application. But, as has been seen, the notion of a linear system is very restrictive; many systems cannot be represented by linear differential equations with constant coefficients, and they are thus not linear. As for so-called linear systems, it was shown in Chap. 11 that they are really linear only within a given frequency range and to certain limits of approximation. In such consideration, this question naturally arises: to what extent does the theory of linear systems represent real conditions sufficiently faithfully to serve as a useful tool? In particular, are there specific cases in which phenomena cannot be explained with reference to the hypothesis of linearity?

22.1.2. Cases in which Linear Theory Is Inapplicable. 1. An example has already been encountered (Sec. 16.2.3), namely, hunting, which can be explained only by saturation, a nonlinear phenomenon. A rigorously oscillating linear system (a couple of pure imaginary poles) oscillates, in fact, with an amplitude determined by the initial conditions; whereas experience shows that a system that hunts appears to possess its own "natural amplitude." In the case of a slightly unstable system (complex poles with very small positive real parts), linear theory takes account of breakage resulting from increasing amplification, but not of the limiting of amplitude to a fixed value which is observed in hunting and which is due precisely to the nonlinear phenomenon of saturation.

2. Another example is provided by the accuracy of servo systems. Reference was made in Chap. 15 to zero deviations under certain conditions. In reality, thresholds are always present (the quantity of information can never be infinite), and it is just this fact that limits accuracy.

3. A third example is provided by on-off servo systems, e.g., when the error-sensing device gives a signal of the same sign as the error but of
constant amplitude (Fig. 22-1). In these conditions, it may be imagined that, in the case of a regulator working in the neighborhood of zero, the system oscillates constantly about its zero-deviation position. More specifically, it can be shown that, under certain very general conditions, systems of this kind oscillate about their zero position with an amplitude and frequency which depend not on the initial conditions, but on the system itself. This is what is called a limit cycle. Here we have a new concept, essentially nonlinear, for it can not be taken into account by any linear approximation to the system.

22.1.3. Some New Properties of Nonlinear Systems. 1. A Recapitulation of the Essential Properties of Linear Systems. Linear systems have the following two properties: proportionality of cause and effect and superposition.

The first of these has been used to define, by harmonic responses, an amplitude ratio depending on frequency alone. If we wish to generalize for nonlinear systems, we come up against a difficulty, since the amplitude of the output, which is periodic, but not sinusoidal, is no longer proportional, for a given frequency, to the input amplitude. With the principle of superposition, the notion of transfer function—the pivot of harmonic methods—fails, since it expresses superposition fundamentally through the convolution theorem.

2. Some Properties of Nonlinear Systems. Nonlinear systems do not possess the properties of proportionality and superposition. In the linear field, these properties involve the existence of a transfer function and of characteristic frequencies proper to the system (frequencies at which the system tends to oscillate with an amplitude depending on initial conditions); in the case of a nonlinear system, amplitude—like frequency—can depend at one and the same time on both the initial conditions and the system itself. Sometimes, as in the case of the limit cycle, both frequency and amplitude are characteristic of the system and independent of initial conditions. In other cases—e.g., when a billiard ball oscillates between parallel cushions—the system has its own specific amplitude, and the frequency depends on the initial energy conditions. One is thus faced with types of behavior which no linear approximation can explain, and which at first encounter may surprise engineers accustomed primarily to linear mechanics.

It is usual to lead into the subject by presenting a specific case of classic nonlinear oscillations, in order to accustom the reader with the notions involved. This is the procedure we shall follow here, taking as our subject certain elementary considerations concerning relaxation oscillations studied by Van der Pol. We shall then go on to a general defini-

1 It can be shown that these properties are characteristic: that is to say, if in investigating a system, hypotheses are adopted leading to the assumption of cause-and-effect proportionality and superposition, linear differential equations, possibly with variable coefficients, are certain to be obtained.
tion of nonlinearity and its more common aspects in relation to servo systems.

3. Introduction to Van der Pol's Equation. According to the linear hypothesis, the response of an unstable system increases exponentially with time. What this means in practice is the following: either the system reaches its breaking point, as can happen in the case of troops marching in step over a bridge, or the system reaches amplitudes at which the linear hypothesis ceases to be applicable. In the case of a system of the second order, three factors must be taken into consideration: (1) the storing of kinetic energy by inertias, self-inductances, or capacitances (Sec. 2.3.2); (2) the storing of potential energy by springs, capacitances, or self-inductances; and (3) the dissipation of energy in electrical or mechanical resistances, or a source of energy if the system is active. The equation of the system, taking into account its own undamped natural frequency $\omega$ and the negative damping ratio $\epsilon$, is

$$\frac{d^2x}{dt^2} - 2\epsilon \omega \frac{dx}{dt} + \omega^2 x = 0$$

To account for the existence of an oscillation of limited amplitude, one may consider a term $\epsilon$, decreasing with amplitude, thus:

$$\epsilon = \epsilon_0 (1 - \alpha x - \beta x^2 - \gamma x^3 - \cdots)$$

One then has a relaxation oscillation system. In particular, if $\epsilon$ is of the form $\epsilon = \epsilon_0 (1 - \beta x^2)$, in other words the sum of a negative resistance $\epsilon_0$ and a positive resistance $\epsilon_0 \beta x^2$ that becomes important at large amplitudes, one has the case first formulated by Van der Pol for the design and operation of radio oscillators. The equation

$$\frac{d^2x}{dt^2} - 2\epsilon_0 \omega (1 - \beta x^2) \frac{dx}{dt} + \omega^2 x = 0$$

is known as Van der Pol's equation.

4. Properties of Van der Pol's Equation. a. Limitation of Amplitude. A sustained oscillation may be considered as a periodic process during which, in the course of a single period, the energy dissipated by the positive resistance is exactly compensated by the energy supplied by the negative resistance—any inequality on one side or the other corresponding to a damped or amplified oscillation. In the case of oscillations represented by Van der Pol's equation, the total energy dissipated during a cycle of period $T$ is

$$\int_0^T 2\epsilon_0 \omega (1 - \beta x^2) \left(\frac{dx}{dt}\right)^2 dt$$

If it is supposed that the oscillations are sinusoidal—i.e., that $x = x_0 \sin \omega t$—the condition for sustaining the oscillation (i.e., for the integral above to be equal to zero) is $x_0^2 = 4/3\beta$, which is independent of $\epsilon_0$. In reality, the oscillations are not sinusoidal; a different value is obtained from that which would result from developing a Fourier series and equating the integral to zero.

1 The present paragraph constitutes an elementary qualitative introduction to some important nonlinear phenomena. It should be ignored by readers who have already had some training in nonlinear oscillations.

2 A number of other nonlinear types can also be imagined. For example, a nonlinear restoring force by Duffing's equation,

$$\frac{d^2x}{dx^2} + \alpha x (1 + \beta x^2) = 0$$
It can be shown that the time taken by the system to pass from the state of rest to the state of sustained oscillation is inversely proportional to \( \varepsilon \); that is, the time increases in proportion to the weakening of the initial negative resistance.

b. Harmonics. If, in the equation

\[
\omega^2 x + \frac{d^2 x}{dt^2} = 2\varepsilon \omega(1 - \beta x^2) \frac{dx}{dt}
\]

the first sinusoidal approximation \( x_0 = 2/\sqrt{3}\beta^{1/2} \) is substituted, a residue results; it is the solution of

\[
\omega^2 x + \frac{d^2 x}{dt^2} = \frac{4\omega^2}{\beta^{1/2}} \cos 3\omega t
\]

Continuing, the solution can be obtained in the form of a Fourier series. It is seen that the amplitude of the third harmonic is proportional to \( \varepsilon \); and the amplitude of the fifth harmonic would be found to be proportional to \( \varepsilon^2 \). Thus, the smaller the \( \varepsilon \), the purer the oscillation. The form of oscillations for increasing values of \( \varepsilon \) is shown in Fig. 22-2, where \( x \) is represented as a function of time. It is seen that, as \( \varepsilon \) increases, the oscillation tends increasingly to the squared form.\(^1\)

c. Synchronization. A Van der Pol oscillator, excited by a signal with a frequency in the neighborhood of \( \omega_1 \) and of sufficient amplitude, oscillates with a frequency \( \omega_1 \). There is no question here of amplification, for the oscillation obtained has an amplitude independent of that of the exciting signal—all that it has in common with the latter is the frequency. This fact can be expressed by the equation

\[
\omega^2 x - 2\varepsilon \omega(1 - \beta x^2) \frac{dx}{dt} + \frac{d^2 x}{dt^2} = B\omega_1^2 \sin \omega_1 t
\]

Taking \( \beta^{1/2} = 1/x_0 \) and \( 2\Delta \omega = (\omega^2 - \omega_1^2)/\omega_1 \cong 2(\omega - \omega_1) \) and seeking a solution of the form

\[x = b_1 \sin \omega_1 t + b_2 \cos \omega_1 t\]

where \( b_1 \) and \( b_2 \) are slowly varying functions,

\[
\frac{db_1}{dt} \ll \omega_1 b_1 \quad \text{and} \quad \frac{db_2}{dt} \ll \omega_1 b_2
\]

yields for the term in \( \omega_1 \), the equations

\(^1\) In certain cases, in the most general form

\[\varepsilon = \varepsilon_0(1 - \alpha x - \beta x^2 - \cdots)\]

saw-toothed oscillations are obtained.
$$b'_1 + \Delta \omega b_2 - \varepsilon \omega \left( 1 - \frac{b^3}{4x^3} \right) = 0$$
$$b'_2 - \Delta \omega b_1 - \varepsilon \omega \left( 1 - \frac{b^3}{4x^3} \right) = B \omega_1$$

where

$$b^3 = b_1^3 + b_2^3$$

These equations admit of a solution in which $b_1$ and $b_2$ are constants, which corresponds to a synchronized oscillation in $\omega_1$. It can be shown that this oscillation is stable only if $B/2x_0 > 2 \Delta \omega/\omega_1$, that is to say, if the synchronization signal is above a critical value. This critical value is greater in proportion as the synchronization frequency differs from the natural frequency of the oscillator. In the limiting case, the oscillator oscillates at its natural frequency in the absence of a synchronizing signal.

A relaxation oscillator can also be synchronized at a frequency approximating a multiple or submultiple of $\omega$ (frequency multiplication or demultiplication).

4. FREQUENCY ENTRAINMENT. If the amplitude of the synchronization signal is below the critical value, the system oscillates at a frequency between $\omega_1$ and $\omega$, the value of which is given by

$$\omega_1 - \left[ \left( \omega - \omega_1 \right)^2 - \frac{\varepsilon^2 \omega^3}{16} \right]^{1/2}$$

This phenomenon is known as frequency entrainment.\(^1\)

5. Applications.\(^2\)

a. SCANNING AND SYNCHRONIZATION OF A CATHODE-RAY OSCILLOSCOPE. The observation of a periodic phenomenon on an oscilloscope screen necessitates, for horizontal scanning, voltages of saw-toothed form with extremely straight waveforms. The image remains stationary if the scanning frequency is strictly equal to that of the phenomenon, or to submultiples of the latter. Horizontal scanning is obtained by means of oscillators derived from neon-lamp oscillators. In order to immobilize the image, recourse is had to frequency adjustment of the time base and adjustment of the synchronization.

Now the addition of an appreciable synchronization signal introduces harmonics and adversely affects the linearity of scanning; so that the ideal solution would be to adjust the frequency of the time base to the exact value of the superposed frequency. This adjustment, however, would not be stable. To effect proper synchronization, it is necessary first of all to cut out the synchronizing frequency, bring the image approximately to rest, and then introduce just sufficient synchronizing signal to immobilize it. Unfortunately, on most oscilloscopes the natural frequency of the oscillator is a function of the adjusting device, and a synchronization signal must permanently operate—which deforms the saw tooth of the horizontal scanning.

b. PHYSICAL APPLICATIONS. There are numerous common examples of natural phenomena which may be accurately represented by relaxation oscillations. The volume of a drop of water forming at the end of a dripping tap varies in accordance with this law. The factors involved are surface tension (elasticity), water pressure (motive force), mass of the drop, and loss of head in the tap (resistance). The volume of the drop increases slowly up to the moment when its weight is balanced by its surface

\(^1\) This entrainment occurs even in the case of nonsinusoidal oscillations, and it may introduce perturbations unless special precautions are taken. In the case of beat generators, the two beat frequencies should not be allowed to entrain or even to synchronize.

\(^2\) The examples quoted here are drawn for the most part from an address delivered by P. Colombani at the Centre d'Études de Mécanique du Vol, Service Technique Aéronautique, Paris.
tension; at that moment, the drop falls, and another drop begins to form in the same way.

The squeaking of a door on its hinge is also an example of relaxation oscillations. Considerable friction exists between the two components of a squeaking hinge, owing to lack of lubrication. Considerable effort is necessary to separate the two parts in contact. When the door is opened, there occurs a compression between the two facing parts of the hinge, up to the moment when the force tending to separate them is greater than the frictional force. The stress then immediately disappears, the two parts come into contact a short distance farther on, and the whole process begins again. This is again confirmed when a saw-toothed voltage is applied to a loud-speaker. The sound produced is exactly like that of a squeaking door, and if the frequency is only slightly varied, the effect is quite remarkable.

The unpleasant physiological effect of a squeaking door is explained by the profusion of very high harmonics. Similarly, when learners on the violin produce disagreeable sounds from their instruments, it is because, in "scraping" the string, the bow acts as a negative resistance and gives rise to notes rich in harmonics painful to the ear.

C. EXPERIMENTAL APPLICATIONS IN PHYSIOLOGY AND BIOLOGY.\(^1\) Attempts have been made to apply Van der Pol's equation experimentally in connection with a number of periodic phenomena encountered in the biological field—phenomena involving frequency entrainment and synchronization. Heart beats have been studied, and an electrical model has been built to simulate the auricles and ventricles and their coupling (the His fasciculus). All known heart disorders, and some unknown ones too, are reported to have been reproduced.

Flowers which close at night and open again in the morning follow a spontaneous oscillatory pattern with a 24-hr period. The period remains the same if the plant is kept in darkness, but may be synchronized by artificial light to periods of 6 or 12 hr.

Studies have also been made of population variations of two animal species, one of which preys on the other, for example, sharks and soles. A proliferation on the part of sharks leads to a decimation of the sole population. This results in more or less periodic oscillations. When expressed mathematically (Volterra), these oscillations are similar to relaxation oscillations.\(^2\) It is interesting to note, in passing, that, if this system is modified to take account of fishing by humans, one arrives at the seemingly paradoxical conclusion that fishing increases the number of fish in the sea. This conclusion has been subject, it is reported, to accurate and quantitative checking.

Finally, in the field of economics, price trends in relation to time have been investigated. On the assumption that speculation is proportional to the rate of variation in prices, a Van der Pol equation is obtained. This theory was put forward in 1931 at a meeting of the Econometric Society.

22.2. DEFINITION AND CLASSIFICATION OF NONLINEAR SYSTEMS

22.2.1. Definitions. A linear system, it will be recalled, designates any physical system which may be represented by a linear differential equation (or system of equations) with constant coefficients. A physical system which may be represented by differential equations of which at least one of the coefficients is a function of time is known as a linear system with

\(^{1}\) Reported by J. Dutilh.

\(^{2}\) A treatment of this question will be found in L. Rauch, "Theory of Oscillation of Nonlinear Systems," par. 2.5, pp. 54–60, University of Michigan, 1955.
variable coefficients. Any physical system which cannot be represented by linear differential equations is known as a nonlinear system.

22.2.2. Notes on the Definition of Nonlinear Systems. The definition of nonlinear systems given above necessitates certain remarks.

1. It is a negative definition, and therefore extremely extensive and vague. Whereas linear systems form a relatively homogeneous group, the term nonlinear system applies to very varied types of systems which have practically nothing in common with one another. A consequence of this is that, whereas methods applicable to linear systems have a remarkable unity, based on the notion of an analytic function, there is no general method applicable to nonlinear systems, unless reference is made to the fundamental laws of mechanics and electricity. From the technical viewpoint, there are methods of attacking special classes of nonlinear problems, but there is no over-all theory of nonlinear systems.

2. It is a relative definition, in at least two senses. In the first place, the definition is relative from the point of view of the degree of accuracy in terms of which one is thinking. It has been seen, for example, that a so-called linear system becomes nonlinear at the point where its threshold, which is always present, is no longer considered negligible. Conversely, it will be seen that many nonlinear systems can, under certain conditions, be classed as linear systems and treated as such to a first approximation, which is sometimes sufficient in practice. If, indeed, accuracy of definition were carried to the limit, all systems would be classed as nonlinear.

Secondly, the definition is relative to the manner in which one applies physical laws to the system. If one writes \((y^2)^{1/2}\) instead of \(y\) in an equation, and forgets to simplify, one is faced with a nonlinear problem. This is by no means a fictitious example. In the system

\[
y'' + y = \cos t \quad z' + z = y'^2 + y^{1/4}
\]

giving one whether one leaves the equations written as they are or eliminates \(y\), one has respectively a nonlinear or a linear system.

22.2.3. The Classification of Nonlinear Systems. On account of the extensive nature of their definition, nonlinear systems can very easily be classified in several different ways. Most of the possible classifications are of a mathematical kind, based on the form of the equations involved. In view of the number and variety of possible classifications, the best way of finding our bearings in the vast and complex domain of nonlinear systems is to begin by listing the types of nonlinearity most frequently encountered in those systems with which we are concerned, namely, servo systems. Once we have established this basis, it will be possible to specify one or more methods of classification which we are certain can be adapted to our problems. This is the aim followed in the next section.

---

1 Such systems are called nonlinear by some authors, who claim that they are not covered by the definition of linear systems. This is a case of erroneous terminology, for these systems possess the fundamental properties of linear systems, namely, proportionality of cause and effect and superposition. The only difference is that transfer-loci techniques, etc., cannot be applied to them without adaptation.
22.3. NONLINEARITIES MOST FREQUENTLY ENCOUNTERED IN SERVO SYSTEMS

22.3.1. General Remark. By examining a servo system from the error-sensing device to the power stage, an initial acquaintance can be made with nonlinearities which, though sometimes overlooked, are always present.

22.3.2. Sensing Stage. 1. The pick-off has generally a nonrectilinear characteristic curve, a threshold, and a saturation. Taking as a specific example an accelerometer, linear or angular, the pick-off is a mass drawn back by a spring, and the transducer is usually electrical (a potentiometer, inductance, or capacitance; see Fig. 22-3).

For the pick-off, the position of the mass is the result of the equilibrium between the inertial force, given by \( F_1 = ma \), and the withdrawal of the spring, which is not strictly proportional but is a function \( F_2(x) \). The function is generally of the form shown in Fig. 22-4; the linear approximation \( F_2 = K_2 x \) represents only the first term of the series developed from it. The relation between \( a \) and \( x \) is thus of the form shown in Fig. 22-4. This is what is known as the curvature effect.

Moreover, lubrication is never ideal; there are frictional forces which can be very approximately represented by a minimal force \( F_m \) which must be overcome if the mass is to move. Instead of being proportional to \( a \), \( x \) is given by the implicit equation

\[
F_1 = F_2 \pm F_m \quad \text{that is} \quad ma = F_2(x) \pm F_m
\]

or, again, if the curvature effect is neglected, by

\[
x = \frac{ma}{K_2} \pm \frac{F_m}{K_2}
\]

This is the threshold effect (Fig. 22-5).

Finally, the displacement of the mass is limited by stops to a maximum \( x_M \) of a few centimeters; this is saturation (Fig. 22-6). The three effects, curvature, threshold, and saturation, together give a characteristic curve
of the form shown in Fig. 22-7. To this must be added the nonlinearities of the transducer, which may be very varied. The special, nonlinear form of the potentiometer characteristic will be noted (Fig. 22-8). For every value of $x$, there is an indetermination of $v$, which is constant before saturation. On account of the impossibility of knowing what the

![Fig. 22-8. Threshold, noise, and saturation.](image)

input $x$ will be, this indetermination is of random nature; hence the term noise which is sometimes used to designate it. If, instead of being potentiometric, the sensing is inductive, there is still a saturation resulting from the end positions. A new phenomenon is then encountered: hysteresis, which results in the output being no longer a single-valued function of the input. This is represented diagrammatically in Fig. 22-9a; or, more schematically, in Fig. 22-9b.

Finally, mention should be made of the existence in certain servo systems of what are known as on-off transducers, in which the mass is
in contact with one or other of the two conducting points which are at opposite potentials. We then have, for example,

\[
\varepsilon = \varepsilon_M \quad \text{for } x > 0 \\
\varepsilon = -\varepsilon_M \quad \text{for } x < 0
\]

which is written as

\[
\varepsilon = \varepsilon_M \text{ sign } x
\]

The characteristic curve is shown in Fig. 22-10. More accurately, taking into account the very small space in which the mass is not in contact with one of the points, the curve is as shown in Fig. 22-11. This is expressed mathematically as

\[
\varepsilon = \varepsilon_M \text{ sign } (x \pm x_M)
\]

22.3.3. Other Stages. When the other stages of a servo system are considered, there are found causes of nonlinearity which in general may be included in the concepts dealt with above.

1. The amplification stage is usually electronic, or it involves relays. If it is electronic, there is in fact a curvature effect resulting from the tube characteristics, and a saturation which is that of the amplifier itself.

If the amplification stage uses relays, the latter can be represented by an on-off system if the relay does not have a median position or by an on-off system with inactive zone if it does have a median position. In the latter case, a lag must be added to allow for the change of the relay from one position to the other. Though only a few thousandths of a second, this lag is not a negligible factor in a rapid servo system.

2. The motor itself gives rise to nonlinearity by reason of its characteristics (Fig. 22-12), coulomb friction on the shaft (inactive zone), and speed and torque saturation.

3. Finally, the system to be controlled is itself not always linear, either. Its nonlinearities may vary considerably according to the nature of the system.

\[\text{Fig. 22-12.}\]

\[\text{Fig. 22-11. On-off with dead zone.}\]
To take a specific case, consider an aircraft flying by automatic pilot. We have the following factors operative:

Mechanical play in the controls (\(\frac{\pi}{4}^\circ\) when the controls are very carefully adjusted; more usually, \(\frac{\pi}{2}^\circ\)).

Aerodynamic interference with the control surfaces, equivalent to noise or inactive zone, and the limitations of deflection in the control surfaces (saturation).

Inconstant holding of course within the limits of variations, i.e., curvature effect (which is very marked for aircraft of unconventional design) plus saturation and hysteresis resulting from stall.

If the aircraft or guided missile is fitted with controls of the spoiler type, the dominant phenomenon is of the on-off type, the output of the spoiler to the left or to the right creating a torque of variable sign but of constant magnitude.

4. Many other examples could be quoted. Those involving a human operator are naturally among the most complicated for the purpose of analytical representation. In the majority of ordinary cases, nonlinearities resembling those described above are encountered.

22.3.4. Classification of Nonlinear Factors in Servo Systems, on the Basis of Four Fundamental Concepts. As we have just seen, the majority of nonlinearities in servo systems may be represented by a combination of four fundamental concepts based on the fact that a quantity (input \(e\) and output \(r\)) can deviate from the ideal linearity represented in Fig. 22-13 by a characteristic curve which is a straight line (linear relation between input and output amplitudes under unvarying conditions). These fundamental concepts are as follows:

1. The dead zone \(e_m\) (Fig. 22-14), defined by \(r(e) = r(0)\) so long as \(|e| < e_m\) (case of a symmetrical dead zone). The dead zone is the field of variation of the input, on either side of zero, in which the output shows no variation. The dead zone is often called inactive zone in the case of relays and threshold in the case of sensing devices.

2. Saturation \(r_M\) (Fig. 22-15), defined by \(|r(e)| \leq r_M\) (case of symmetrical saturation), the absolute maximum value the output \(r_M\) can assume. When, within saturation, linearity is preserved (Fig. 22-15), it is possible to define the value of the input \(\pm e_M\) corresponding to saturation.

3. The curvature effect (Fig. 22-16) corresponding to a nonrectilinear characteristic \([r(e) \neq ke]\).

4. Hysteresis (Fig. 22-17), where \(r\) depends on \(e\). This can be represented by closed curves, which usually depend on the manner in which \(e\) varies.

These basic notions of nonlinearity can, in combination with one

---

1 This is the term used to designate a comb-shaped surface normally fitted to the wing in such a way that it "spoils" the shape of the wing edge. The control of a spoiler does not involve a torque; it can be effected by means of an electromagnet, thus economising both the space and the time lag of a servomechanism. Many guided missiles are controlled by spoilers (Fig. 28-4).

2 To avoid confusion, it should be specified that, in Fig. 22-13 and the following diagrams representing the four fundamental concepts of nonlinearity, the variable on the axis of abscissas is the steady-state input amplitude of the nonlinear component.
Figs. 22-13 to 22-20. Types of nonlinearities most frequently encountered in servo systems.

another, result in almost all the types of nonlinearity encountered in practice. For example, one can have:

a. On-off with dead zone (Fig. 22-18), defined by

\[ r = 0 \quad \text{for} \quad |e| < e_m \]
\[ r = r_M \quad \text{for} \quad |e| > e_m \]

b. On-off with dead zone and hysteresis, shown in Fig. 22-19
c. **Dead zone and saturation**, defined by

\[
\begin{align*}
    r &= 0 & \text{for } |e| < e_m \\
    r &= k(e - e_m) & \text{for } e_m < |e| < e_M \\
    r &= r_M & \text{for } |e| > e_M
\end{align*}
\]

and

This is represented in Fig. 22-20. The linearity domain (Chap. 11) is characterized by the ratio \(e_M/e_m\).

Note that the basic types of nonlinearity can also be included one within the other. For example, saturation can be considered as the limiting case of the curvature effect when \(dr/de = 0\) for \(|r| > r_M\). Also, on-off is a particular case of saturation, and so on. These concepts should be regarded as only schematizations which frequently overlap. For example, the saturation system of Fig. 22-15, when it is included in an over-all system which hunts with an amplitude \(r_p > r_M\), can often be schematized approximately by an on-off with dead zone \(r_M\). Along the same lines, we may consider the representation of a symmetrical characteristic, as shown in Fig. 22-21, with three parameters \(e_m, r_M, \phi\), as capable of representing:

a. A purely linear system: \(e_m = 0, r_M = \infty\)

b. A pure dead-zone system: \(r_M = \infty\)

c. A pure saturation system: \(e_m = 0\)

d. An on-off system: \(\phi = \pi/2\)

e. Approximately, curves of the following forms:

\[
\begin{align*}
    r &= \pm e^{1/4} & e_m &= 0 \\
    r &= \pm e^2 & r_M &= \infty
\end{align*}
\]

for example, a function with horizontal tangent at the origin or with horizontal parabolic branch; in other words, the majority of the usual nonlinearities, with the exception of hysteresis.

**22.3.5. Note on the Threshold Due to Noise.** The threshold considered above is a dead zone of input values for which the output is zero. The case arises, for example, in a relay with median position. For the relay blade in the median position to engage in one of the two other positions, the exciting voltage of the relay must exceed a certain value, which is its threshold. The term *threshold* is also used to designate the inactive zone at the output end, in particular in the neighborhood of zero, where the output is blurred by noise; this is the *noise threshold.* Though this is not a nonlinear phenomenon, the fact is mentioned because of its importance in defining the performances of a servo system and its components, especially the sensing device.

Suppose that a stationary noise, with the frequency spectrum \(\Phi_n(\omega)\) and rms value

\[
\text{rms value } = \left[ \int_{-\infty}^{+\infty} \Phi_n(\omega) \, d\omega \right]^{1/2}
\]
is superimposed on the input of a sensing device with a linear characteristic \( r_1 = k_1e \).

At the output of the sensing device the noise has an rms value of \( n_1 = k_1n \), and the signal-to-noise ratio has, for every permanent value of \( r_1 \), the value

\[
\frac{r_1}{n_1} = \left[ \int_{-\infty}^{+\infty} \Phi_n(\omega) \, d\omega \right]^{\frac{1}{2}}
\]

If at the output of the sensing device we have a chain of linear systems with a transfer function \( k_2H(s) \) [with \( H(0) = 1 \)] whose output \( r_2 \) is the effective output (Fig. 22-22), then the rms value of the noise at this output is

\[
n_2 = k_1k_2 \left[ \int_{-\infty}^{+\infty} \Phi_n(\omega) |H(j\omega)|^2 \, d\omega \right]^{\frac{1}{2}}
\]

and the signal-to-noise ratio is

\[
\frac{r_2}{n_2} = 2 \left[ \int_{-\infty}^{+\infty} \Phi_n(\omega) |H(j\omega)|^2 \, d\omega \right]^{\frac{1}{2}}
\]

It is the ratio \( r_2/n_2 \) which is of interest to the user, who will not distinguish the signal \( r_2 \) from the noise \( n_2 \) when \( r_2 \) is below a certain value \( r_2 = \alpha_2n_2 \), that is, when

The ratio is below the value of \( \alpha_2 \) determined in view of the user's requirements. Hence an equivalent input threshold \( e_0 \) is determined, given by

\[
e_0 = \alpha_2 \left[ \int_{-\infty}^{+\infty} \Phi_n(\omega) |H(j\omega)|^2 \, d\omega \right]^{\frac{1}{2}}
\]

For \( |e| < e_0 \), the user can no longer distinguish the output \( r_2 \) from the noise \( n_2 \).

It will be seen from the above relationship that the threshold depends on the noise spectrum and on the transfer function of the system. In particular, the signal-to-noise ratio at the output of the sensing device, \( r_1/n_1 = \alpha_1 \), for an input value equal to \( e_0 \), is not equivalent to the ratio \( r_2/n_2 = \alpha_2 \) at the effective output

\[
\alpha_1 = \alpha_2 \left( \frac{n_2}{n_1} \right)^{\frac{1}{2}} \text{ with } \left( \frac{n_2}{n_1} \right)^{\frac{1}{2}} = \frac{\int_{-\infty}^{+\infty} \Phi_n(\omega) |H(j\omega)|^2 \, d\omega}{\int_{-\infty}^{+\infty} \Phi_n(\omega) \, d\omega}
\]

**Example.** Suppose that \( \Omega \) is a white noise with a break frequency \( \omega_1 = 200 \text{ rad/sec} \) and that \( k_2H(s) \) is the transfer function of an ideal unity-feedback servo system with a cutoff frequency \( \omega_2 = 20 \text{ rad/sec} \). Assuming that the signal-to-noise ratio \( r_2/n_2 \) necessary for the user to distinguish the output from the noise is \( r_2/n_2 = \alpha_2 = 2 \), the equivalent threshold in this case, at the input, is

\[
e_0 = 2 \left[ \int_{0}^{\Omega n_2} \omega_1 \times 1 \, d\omega \right]^{\frac{1}{2}} = 2 \left( \frac{\omega_1}{\omega_1} \right)^{\frac{1}{2}} n = 0.63n
\]

and the corresponding signal-to-noise ratio at the sensing-device output is
\[ \alpha_3 \left[ \int_0^{\omega_3} \frac{n}{\omega_1} \times 1 \, d\omega \right] \]
\[ \times \frac{n}{d\omega} = \alpha_3 \left( \frac{\omega_2}{\omega_1} \right)^{1/2} = 0.63 \]

It will be noted that, by reason of the linear\(^1\) filtering out of noise after the sensing stage, a signal-to-noise ratio at the sensing-device output below that required at the effective output suffices.

**22.3.6. Distinction between Continuous and Discontinuous Nonlinearities.** It may be useful, notably in applying mathematical theory, to distinguish between continuous and discontinuous nonlinearities. The on-off type is discontinuous ("sign" function). The curvature effect is, in general, continuous for \(r(e)\) and \(dr/de\). Dead zone and saturation are, in general, continuous for the \(r\) function, but not for its derivative \(dr/de\).

According to the method used, it may be useful to schematize a dead zone or a saturation by a continuous curve; or, alternatively, to express the nonlinearity constituted by a continuous curve by the incorporation of a dead zone or a saturation alone (Fig. 22-23).

![Fig. 22-23. Nonlinearities of the continuous and discontinuous types.](image)

**22.3.7. Another Distinction: Incidental and Purposeful Nonlinearities.** Again, we may distinguish between *incidental*, or *unwanted*, nonlinearities and those purposely introduced by the designer of the system. The latter may be termed *essential* nonlinearities.

1. Incidental nonlinearities result from limitations and imperfections in design and equipment, when the designer aims at linear solutions. They usually have an unfavorable effect on the performance of the system. Examples are curvature effect in amplifiers, thresholds in sensing devices and relays, end positions in detectors, saturation in amplifiers, etc.

   These nonlinearities arise in various ways. Curvature effects and thresholds arise more or less continuously, especially thresholds, in the functioning of regulators working most of the time in the neighborhood of zero. Saturation effects occur if the system reaches its stops (or their equivalents), which normally should not be the case; these nonlinearities are thus doubly incidental.

2. The second, or purposely introduced nonlinearities, result from the deliberate intention of the designer to achieve simplicity or to adapt the

---

\(^1\) See Sec. 12.3.4, par. 2, footnote.
system to some special purpose. The most typical example is provided by on-off sensing devices or controls whose conception is so different from that of continuous servo systems that a distinction is commonly made by referring to on-off servo systems and (theoretically) linear servo systems.

22.3.8. Conclusion. The above indications show the immense variety of nonlinearities which can be present in a servo system. Their study demands first a systematic examination of the characteristics of the various stages; in some cases this may necessitate long experimental investigations, because of the great number of parameters involved. The preceding explanations, especially of the four fundamental modes of nonlinearity, can provide a convenient language for expressing these phenomena and a guide to the study of their effects.

22.4. METHOD OF INVESTIGATION

22.4.1. Problems. Some of the problems concerning nonlinearity which may arise are given below.

1. Influence of Unwanted Nonlinearities. A servo system has been designed as theoretically linear. It is required to examine the effect of unwanted nonlinearities, that is:
   a. Determine the performance of the system if thresholds, etc., are taken into account.
   b. Determine what will happen if the system no longer remains linear, because of reaching its stops.

2. Optimal Nonlinear Servo Systems. It is required to determine whether the introduction of nonlinearities in servo systems is advantageous or disadvantageous from the point of view of performance; for example, whether for a given problem there exists a nonlinear system with performance superior to that of the optimum linear system.

3. General Theory of Nonlinear Servo Systems. Does the presence of nonlinearities in servo systems lead to the introduction of new concepts, foreign to linear systems?


The methods of approach to these various problems are very different, according to the degree of accuracy aimed at and the nature of the nonlinearities involved. These methods are outlined in the following sections.

22.4.2. Methods of Approach. 1. The first approach consists in attempting to outmaneuver nonlinearities by seeking to extend harmonic methods, which are so powerful in dealing with linear systems, to the greatest possible number of nonlinear systems. In some cases, this can be done with absolute exactitude. In others, a valid approximation is frequently possible; in yet others, such an extension is quite inapplicable.

2. The second approach consists in taking the bull by the horns and dealing with nonlinear systems as such. In this connection, the diversity
of nonlinear problems and the considerable amount of research work undertaken\(^1\) have resulted in a substantial mass of work accomplished, in Germany, during and since the last years of World War II, and elsewhere, especially since 1947. Nonlinear problems have been approached in the following ways especially:

\(a\). The complete solution of special cases with as many parameters as possible, taking into account the methods of numerical calculation available, and the expression of the results in condensed form by means of curves and charts

\(b\). The application of extension of mathematical methods of dealing with nonlinear differential equations, notably topological and numerical methods as developed by Poincaré

\(c\). The use of various new concepts derived from engineering, topology, etc.

\(d\). The systematic use of electronic calculating machines

\(22.4.3.\) Consequences. As a consequence of the above, it can be said that there is no general nonlinear theory, but there are nonlinear problems. Thus the authors had to face the following dilemma in writing the third part of this book: Part 3 could present theories relating to general types of nonlinear equations, which would have been far removed from the subject of applications to the servo field (as is the case in the majority of books on nonlinear mechanics), or it could present a great number of special nonlinear servo problems, which would have resulted in a collection of unrelated notes.

The following solution was chosen: (a) Most of the third part of the book concentrates on a small number of methods, chosen from among the more general and the more interesting didactically. (b) A systematic account is given of how these methods apply to servo problems and how the effect of nonlinearities in servomechanisms can be evaluated.

The type of servo system which we have particularly in mind in the third part of this book is a servo system which incorporates one nonlinear component, as will be explained in detail in the next section.

\(22.4.4.\) Servo System with One Nonlinear Component. A great number of the types of servo systems incorporating one nonlinear component can be imagined, according to the configuration of the system and the location of the nonlinear element. However, it can be shown that most cases can be represented by the general block diagram shown in Fig. 22-24, in which \(N\) represents the nonlinear element and \(x\) and \(y\) its input and output, respectively. The quantities \(e_1\) and \(r_1\) are related to the input \(e\) and output \(r\) of the system by linear differential equations. The transfer function \(L(e)\) is a simple combination of the transfer functions of the linear elements.

\(^1\)Research connected for the most part with automatic-control problems in the military field (notably guided missiles), the consequence being, unfortunately, that many interesting results on the practical level are withheld for military security reasons.
Case 1. The nonlinear element $N$ is not enclosed within an internal loop. In this case it is immediately seen that Fig. 22-25a is equivalent to Fig. 22-25b, which has the form of Fig. 22-24 with $1$

$$e_1 = Fe \quad r_1 = \frac{1}{G} r$$

Similarly, in the case of a nonunity feedback, Fig. 22-26a is equivalent to Fig. 22-26b. These considerations obviously apply to the case in which the blocks $F$ or $G$ involve internal loops, provided these internal feedback systems are linear.

![Diagram](image)

**Fig. 22-24.**

![Diagram](image)

**Fig. 22-25.** Two block diagrams are equivalent.

![Diagram](image)

**Fig. 22-26.** Two block diagrams are equivalent.

Case 2. The nonlinear element $N$ is included in an internal loop (Fig. 22-27). The block diagram of the system can then be redrawn by rearranging the elements outside the nonlinear element while the input and output of the latter are kept in place by taking into account the idea of Sec. 13.1.7, Note 1. This gives, successively, the block diagrams shown in Fig. 22-28a and b. Finally, the latter is equivalent to the block diagram shown in Fig. 22-29, which is of the general type of Fig. 22-24 with

$$L = G(FJH + K) \quad e_1 = Fe \quad r_1 = \left(FH + \frac{K}{J}\right) r$$

$1$ In this section $e_1 = Fe$ is written instead of $e_1 = \mathcal{L}^{-1}(FZe)$ for the sake of brevity.
In conclusion, practically all types of servo systems involving one nonlinearity can be represented by the block diagram of Fig. 22-24. This block diagram is redrawn in Fig. 22-30, using the notations that will be used throughout Part 3. It must be remembered, therefore, that the functions $e(t)$, $r(t)$, $e(t)$, and $L(s)$ represent, in Fig. 22-30, functions related by linear equations to the actual input, output, error, and block transfer functions. Two cases are of particular interest.

a. The case in which the element $N$ is of the on-off type. This is a typical case of essential nonlinearity as defined in Sec. 22.3.7. The
corresponding servomechanism is then called an on-off servo system; its properties are radically different from those of a linear servomechanism. The on-off element may belong to any of four types, according to whether or not a dead zone $\Delta$ and/or hysteresis $h$ are present. If the nonlinear element involves a delay $T$, the latter can usually be incorporated into the linear components by multiplying $L(s)$ by a factor of the type $e^{-sT}$.

b. The case in which the element $N$ is characterized by a dead zone, a saturation, and a curved characteristic. Studying the corresponding servo amounts to studying the effect of the corresponding incidental nonlinearities (as defined in Sec. 22.3.7) on the performance of a feedback control system.

22.4.5. Outline of the Following Chapters. In view of the above considerations, Chap. 23 will be devoted to methods of studying the time response of nonlinear servo systems of the two types just defined. Then Chaps. 24 and 25 will respectively outline the method of the first-harmonic approximation and Poincaré's approach, emphasis being laid on their applications to the above types of nonlinear servomechanisms. Chapters 26 and 28 will then treat more special problems related to the oscillations of nonlinear servo systems. Finally, Chap. 27 will present a few general mathematical methods which may have applications in the servo field.
CHAPTER 23

TRANSIENTS IN NONLINEAR SERVO SYSTEMS

Summary
1. Time response of on-off servo systems.
2. Time response of servo systems with one nonlinear element.
3. Conclusion.

Obtaining the time response of nonlinear systems in general is equivalent to nothing more than the general problem of solving any nonlinear equation. This chapter will be restricted to the case of servo systems incorporating one nonlinear element, as defined in Sec. 22.4.4. It has been shown that such systems can be represented by the general block diagram of Fig. 23-1, in which \( N \) represents the nonlinear element and \( L(s) \) is a transfer function that incorporates the properties of all the linear components. The cases (1) in which the element \( N \) is of the on-off type and (2) of any non-frequency-dependent type will be studied.

23.1. TIME RESPONSE OF ON-OFF SERVO SYSTEMS

23.1.1. Case in which no Hysteresis is Present. Consider an on-off control system subjected to an input (command) \( e(t) \). The output of the on-off element is a square wave \( w(t) \), the positive intervals of which correspond to those instants at which \( e(t) \) is positive—that is, \( e(t) \) is greater than \( r(t) \)—whereas the negative intervals correspond to \( e(t) \) smaller than \( r(t) \). The output \( r(t) \) is the response of the linear part to the square wave \( w(t) \).

The response of the system to the command \( e(t) \) can be visualized as follows: Suppose the system is initially at rest and \( e(t) \) is then applied at the instant \( t = 0 \) and starts by being positive. In the first instants the error \( e(t) \) is positive, hence \( w = +M \). The response of the linear part, thus actuated by the step function \( Mu(t) \), is \( M \) times its unit-step

---

\(1\) The graphical method presented here is used by different people and in different countries. In France it was first developed by C. Lepage, "Méthode de calcul de la réponse des systèmes asservis par tout-ou-rien à une excitation quelconque," C.E.M.V. no. 30, Service Technique Aéronautique, Paris, 1950.
response \( q(t) \) shown in Fig. 23-2. As time goes on, \( q(t) \) increases and will eventually catch up with \( e(t) \) (Fig. 23-3) at a time \( t = t_1 \). Then the on-off element will commutate in the reverse direction \( (w = -M) \), thus causing the linear part to be actuated by \(-2Mu(t - t_1)\). As a result, the response of the system will be

\[
r(t) = M[q(t)u(t) - 2q(t - t_1)u(t - t_1)]
\]

\((t_1 < t < t_2)\)

until the next commutation takes place at \( t = t_2 \), when again \( r(t) = e(t) \).

Thus it is seen that the output of the system is the response of the linear part \( L(s) \) to the square wave

\[
w(t) = M[u(t) - 2u(t - t_1) + 2u(t - t_2) - \cdots]
\]

\[
= M \left[ u(t) + 2 \sum_{i} (-1)^i u(t - t_k) \right]
\]

where \( t_1, t_2, \ldots \) are the instants of commutation. The expression for \( r(t) \) is

\[
r(t) = M[q(t)u(t) - 2q(t - t_1)u(t - t_1) + 2q(t - t_2)u(t - t_2) \cdots]
\]

\[
= M[q(t)u(t) + 2 \sum_{i} (-1)^i q(t - t_k)u(t - t_k)]
\]

where \( q(t) \) is the unit-step response of the linear part \( L^{-1}L(s)/s \).

These considerations make it possible to plot the function \( r(t) \) when the input \( e(t) \) and the step response of the linear part \( q(t) \) are known.

23.1.3. Remarks. 1. The same method applies when the system does not start from rest. But an additional term \( r_i(t) \) must then be added to \( r(t) \) in order to account for the initial conditions.

2. If the linear part is characterized by its transfer function \( L(s) \), the latter must have a denominator of a higher degree than the numerator in order that \( q(t) \) will not involve discontinuities. If this were not the case, it would mean that inertial terms, or their analogs, were neglected. It would then be safer to write the complete equations when applying the method.

3. For the case in which the input \( e(t) \) is an increasing function of time, the graphical procedure just outlined soon gets cumbersome. It is, then, more convenient to write \( q(t) \) as the sum of a function \( q_i(t) \) that approaches infinity as \( t \) becomes infinite, and a function \( q_f(t) \) that remains finite. Figure 23-4, where, for simplicity, \( M \) has been taken as unity, shows the situation for the case in which \( q_i(t) \) is a linear function of time, that is, in which \( L(s) \) incorporates one integration. The expression of \( r(t) \), before the first commutation occurs, is

\[
r(t) = M[q_i(t) + q_f(t)]
\]
After the first commutation has taken place—that is, for $t > t_1$, $r(t)$ becomes

$$r(t) = M[q_i(t) - 2q_i(t - t_1)]$$

The function $q_i(t) - 2q_i(t - t_1)$ is a linearly decreasing function of time with a slope equal in magnitude but opposite in sign to that of $q_i(t)$. The same remark

\[\text{Fig. 23-4.}\]

\[\text{Fig. 23-3. Response of ideal on-off servo.}\]
applies to the interval \((t_1, t_2)\) in which the function \(q_1\) is a straight line parallel to \(q_1(t)\), and so forth. This makes the graphical procedure very easy.

23.1.3. Case in which Hysteresis is Present. When hysteresis \(h\) is present, the same reasoning can be applied, the only difference being that commutations from \(-\) to \(+\) occur when \(\varepsilon - r = h/2\), and from \(+\) to \(-\) when \(\varepsilon - r = -h/2\). Figure 23-5 is self-explanatory.

23.1.4. On-Off Systems with Dead Zone and without Hysteresis. The behavior of an on-off control system with a dead zone \(\Delta\) can be explained in a similar manner. The commutations of the on-off element occur when the error function is \(+\Delta/2\) or \(-\Delta/2\). They may be of four types:

\[
\begin{align*}
0 \text{ to } + & \quad \text{when } \varepsilon = \frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} > 0 \\
+ \text{ to } 0 & \quad \text{when } \varepsilon = \frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} < 0 \\
0 \text{ to } - & \quad \text{when } \varepsilon = -\frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} < 0 \\
- \text{ to } 0 & \quad \text{when } \varepsilon = -\frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} > 0
\end{align*}
\]

Fig. 23-5. Response of on-off servo with hysteresis.
Considering, to begin with, an input $e(t)$ applied to the system starting from rest and supposing $w$ is equal to $+M$ at the instant $t = 0^+$, it is seen (Fig. 23-6) that the linear system is actuated by

$$w(t) = M[u(t) - u(t - t_1)] - u(t)$$

thus yielding the response

$$r(t) = M[q(t) - q(t - t_1) - q(t - t_2) + q(t - t_3)]$$

where, for simplicity, $q(t - t_1)$ is written for $q(t - t_1)u(t - t_1)$.

This response can be obtained graphically from the data of $e(t)$ and from a plot of $q(t)$. As in the previous case, nonzero initial conditions can be accounted for by an additional term $r_i(t)$.

![Diagram](image)

Fig. 23-6. Step response of on-off servo with dead zone.

It is to be noted that the sequence of commutations shown in Fig. 23-6 is not the only one possible; Fig. 23-7 shows a different possibility.

23.1.5. Case in which Hysteresis is Present. The above graphical method can easily be extended to the case in which the on-off element has hysteresis in addition to the dead zone $\Delta$. Figure 23-8 is self-explanatory.

23.2. TIME RESPONSE OF SERVO SYSTEMS WITH ONE NONLINEAR ELEMENT

23.2.1. Assumptions. General Outlines. Consider now the case in which the nonlinear element $N$ is not necessarily of the on-off type.
Fig. 23-7. Particular case of response when dead zone is present. Note the two consecutive positive pulses for $w(t)$.

Fig. 23-8. Response of on-off servo with dead zone and hysteresis.
Suppose this element is defined by an amplitude characteristic curve of \( w \) vs. \( \varepsilon \) (Fig. 23-9) and is not frequency-dependent. The \( w \)-vs.-\( \varepsilon \) relation will be assumed to be single-valued; that is, the case in which hysteresis is present will not be considered.

If \( w = N(\varepsilon) \) is the characteristic of the nonlinear element, then \( r(t) \) is given as the convolution integral (Sec. 4.4.5).

\[
R(t) = \mathcal{L}^{-1}L(s)W(s) = \int_0^t h(t - \lambda)w(\lambda) \, d\lambda
\]

(23-1)

in which \( h(t) \) (Fig. 23-10) is the inverse Laplace transform of \( L(s) \)—that is, the unit-impulse response of the linear part (Sec. 7.3.1)—and

\[
w(t) = N[\varepsilon(t)]
\]

If \( r(t) \) is replaced by \( \varepsilon(t) - e(t) \), Eq. (23-1) becomes

\[
\varepsilon(t) = e(t) - \int_0^t h(t - \lambda)N[\varepsilon(\lambda)] \, d\lambda
\]

(23-2)

In this equation the input \( e(t) \) and the functions \( h(t) \) and \( N(\varepsilon) \) are known. Therefore, the equation can be solved with respect to \( \varepsilon(t) \). For this purpose, the procedure developed by B. Naumov\(^1\) provides an approximate method which involves a reasonable amount of labor and whose accuracy is sufficient for practically all cases encountered in the field.

To obtain \( \varepsilon(t) \) from Eq. (23-2), the integral can be computed numerically for a given value of \( t \) by dividing the interval \((0, t)\) into \( n \) intervals equal to \( T = t/n \) and applying the classical formula of the trapezoidal rule (Fig. 23-11):

\[
\int_0^{nT} f(\lambda) \, d\lambda = T \left\{ \frac{1}{2} f(0) + f(T) + \ldots + f((n - 1)T) + \frac{1}{2} f(nT) \right\}
\]

to the function \( f(\lambda) = h(nT - \lambda)N[\varepsilon(\lambda)] \). This yields

\[
\varepsilon(nT) = e(nT) - T \left\{ \frac{1}{2} h(nT)N(\varepsilon_0) + h[(n - 1)T]N(\varepsilon_1) + \ldots + h(T)N(\varepsilon_{n-1}) + \frac{1}{2} h(0)N(\varepsilon_n) \right\}
\]

(23-3)

where, for simplicity, \( \varepsilon_k \) denotes \( \varepsilon(kT) \). Equation (23-3) can be solved with respect to \( \varepsilon(nT) \); that is, it enables one to compute \( \varepsilon(t) \).

23.2.2. Obtaining $\varepsilon(t)$. First consider the case in which $h(0)$ is zero, that is, the impulse response of the linear part involves no discontinuity. This is the case in which the denominator of $L(s)$ is at least two degrees higher than the numerator. The procedure is straightforward. Writing Eq. (23-3) for $n = 0$ gives $\varepsilon_0 = e(0)$. Then, for $n = 1$

$$\varepsilon(T) = e(T) - \frac{1}{2} Th(T)N(\varepsilon_0)$$

which gives $\varepsilon_1$, since $\varepsilon_0$ is known. Substituting $n = 2$ then yields

$$\varepsilon(2T) = e(2T) - \frac{1}{2} Th(2T)N(\varepsilon_0) - Th(T)N(\varepsilon_1)$$

which gives $\varepsilon_2$. Similarly

$$\varepsilon(3T) = e(3T) - \frac{1}{2} Th(3T)N(\varepsilon_0) - Th(2T)N(\varepsilon_1) - Th(T)N(\varepsilon_2)$$

and so forth as far as $\varepsilon_n = \varepsilon(t)$.

23.2.3. Case in Which $h(0) \neq 0$. The procedure is slightly more involved when $h(0) \neq 0$, that is, when the impulse response of the linear part involves a discontinuity.

![Trapezoidal rule for $h(t)$](image)

This occurs when the denominator of $L(s)$ is only one degree higher than the numerator. In this case the step-function response $q(t) = \int_0^t h(\lambda) \, d\lambda$ starts with a nonzero slope.

The difficulty arises from the fact that any equation, say, the $(k + 1)$th, involves both $\varepsilon_k$ and $N(\varepsilon_k)$:

$$\varepsilon_k + \frac{1}{2} Th(0)N(\varepsilon_k) = e(kT) - T[\frac{1}{2} h(kT)N(\varepsilon_0) + h((k - 1)T]N(\varepsilon_1) + \cdots + h(T)N(\varepsilon_{k-1})]$$

(23-4)

The equation cannot, therefore, be directly solved except for the case in which $N(\varepsilon_k)$ is a simple algebraic function of $\varepsilon_k$. The following graphical step-by-step procedure, due to B. Naumov, can be applied in all cases. Let

$$\Gamma_k = Th(kT) \quad S = \frac{1}{2} Th(0)$$

Equation (23-4) becomes

$$\varepsilon_k + SN(\varepsilon_k) = e(kT) - [\frac{1}{2} \Gamma_k N(\varepsilon_0) + \Gamma_{k-1} N(\varepsilon_1) + \Gamma_{k-2} N(\varepsilon_2) + \cdots + \Gamma_1 N(\varepsilon_{k-1})]$$

(23-5)

or, if the right-hand side of the latter equation is called $\Delta_k$,

$$SN(\varepsilon_k) = \Delta_k - \varepsilon_k$$

The values of $\varepsilon_k$ and $N(\varepsilon_k)$ can be obtained if the functions of $\varepsilon_k$

$$f(\varepsilon_k) = SN(\varepsilon_k) \quad g(\varepsilon_k) = \Delta_k - \varepsilon_k$$
are plotted and their intersection is taken. For successive values of $k$

\[
\begin{align*}
  k &= 0 & SN(e_0) &= \Delta_0 - e_0 & \text{with} & \quad \Delta_0 = e_0 \\
  k &= 1 & SN(e_1) &= \Delta_1 - e_1 & \text{with} & \quad \Delta_1 = e(T) - \frac{1}{2}T_1N(e_0) \\
  k &= n & SN(e_n) &= \Delta_n - e_n
\end{align*}
\]

the quantity $\Delta_k$ in the last equation being the right-hand side of Eq. (23-5). The first of these equations ($k = 0$) yields $e_0$ and $N(e_0)$, the second ($k = 1$) can then be solved to yield $e_1$ and $N(e_1)$, and so forth until $k = n$. For all these equations the function $f(x)$ is the amplitude characteristic $N(e_x)$ of the nonlinear component multiplied by the constant factor $S = T_0/2$. For the $(k + 1)$th equation the function

\[g(e_k)\]

is a straight line with a $-45^\circ$ slope, which intersects the vertical axis at $e_k = \Delta_k$ (Fig. 23-12).

23.3. CONCLUSION

The methods just outlined make it possible to obtain the time response of a servo system incorporating one nonlinear element to any input $e(t)$, in particular to any of the typical inputs listed in Sec. 3.3.1. Once the function $h(t)$ or $q(t)$ that characterizes the linear part has been determined, these methods involve a perfectly reasonable amount of work, which does not increase with the complexity of the system under consideration. Recourse may naturally be had to electronic computers in applying these methods.

However, as a result of their step-by-step procedure the methods have two limitations: (1) Studying the influence of the system parameters on the time response often involves considerable labor, which makes the methods rather unsuitable for design problems. (2) The methods enable one to determine the transient behavior of the system, but they do not lend themselves to the study of the steady state.
CHAPTER 24

THE FIRST-HARMONIC APPROXIMATION

Summary

1. The concept of the describing function.
2. Describing functions of customary nonlinear elements.
4. Their performance and compensation.
5. Additional comments on describing functions.

In view of the fact that the transfer-function approach is such a powerful tool in the realm of linear servo systems, attempts have naturally been made to extend it, so far as possible, to nonlinear systems. Different workers in various countries simultaneously had the idea of considering a sinusoidal equivalent for periodic but not sinusoidal functions, in order to describe the frequency response of nonlinear systems in terms of transfer functions. In the United States this method is generally considered as due to R. Kochenburger, whereas the French credit it to J. Dutilh, and the Russians to L. Goldfarb. The techniques suggested by these authors are substantially equivalent. They will be outlined in the present chapter.

24.1. THE CONCEPT OF THE DESCRIBING FUNCTION

24.1.1. General. The method can be considered as an attempt at a generalization of the notion of the transfer function to cover nonlinear systems. For a linear system, with input $e$ and response $r$, the transfer function has been defined, on the basis of sinusoidal responses $e = e_0 \sin \omega t$ by the amplitude ratio $A = r_0/e_0$ and the phase $\Phi$ of the response

$$r = r_0 \sin (\omega t + \Phi)$$

By reason of the linear properties of the system, $A$ and $\Phi$ are functions of the input frequency, but they do not depend on the input amplitude.


403
It is this fact which makes it possible to define a transfer function as a function of $\omega$ alone:

$$r_0/e_0 = A(\omega) \quad \Phi = \Phi(\omega)$$

When it is attempted to extend this procedure to a nonlinear system, one has to consider the response for a harmonic input $x = x_1 \sin \omega t$. This response $r(t)$ is periodic but not sinusoidal. If, by a convention, a sinusoidal function

$$w(t) = w_1 \sin (\omega t + \Psi)$$

is (arbitrarily) designated as the "equivalent" of the function $r(t)$, then one may define for the system an "equivalent transfer function"

![Fig. 24-1. Basic principle of the method.](image)

Modulus $= \frac{w_1}{x_1}$ \quad Phase $= \Psi$.

The first harmonic of $r(t)$ is usually taken as the equivalent sinusoidal function (Fig. 24-1).

The difference with respect to the linear case is twofold:

1. There is an arbitrary factor in the choice of the equivalent sinusoidal response.

2. The resulting transfer function depends not only on $\omega$ but also on the input amplitude $x_1$, since there is no longer the proportionality of cause and effect as in the linear case:

$$\frac{w_1}{x_1} = B(x_1, \omega) \quad \Psi = \Psi(x_1, \omega)$$

Graphically, in the case of a nonlinear system, we are thus no longer faced with a transfer locus graduated in frequency, but with a family of transfer loci, one for each input amplitude (Fig. 24-2).

**24.1.2. Justification of the Method.**

If one considers that the response signal of a relay has more or less the form of a square wave and, therefore, is far removed from a sine curve, one may logically be surprised that its first-harmonic approximation can lead to valid results. The essential reason is to be found in the very marked filtering out of high frequencies effected by the power stage of any servomechanism.

This point can be made clear by the following type of reasoning, which is not rigorous. Consider the signal $w(t)$ shown in Fig. 24-3. The amplitude of its first harmonic is $4/\pi$, that of the second harmonic is 0, and that of the third harmonic is $4/3\pi$, that is, one-third of the first. If the first harmonic, when it constitutes the input of a second-order system, is in the bandwidth, and the third harmonic is outside the bandwidth, the latter will be damped up to $(\frac{1}{3})^3$ times more than the former. In
other words, the third harmonic may be $3^3$, or 27, times smaller than the first at the output. Since 3 per cent is approximately the order of accuracy usually demanded in the application of transfer-function methods, it is understandable that the approximation to the first harmonic is allowable in a great many cases.

It is difficult to state this justification with greater precision, and this vagueness is, indeed, one of the weaknesses of the method.\(^1\) Suffice to say that experience, and comparison with rigorous calculations, very frequently bear out the validity of the approximation to the first harmonic.

**FIG. 24-3.**

24.1.3. The Describing Function. Suppose a harmonic function of time $x = x_1 \sin \omega t$ with a period $T = 2\pi/\omega$ is fed into a nonlinear element $N$. After the transient had died away, the output $w(t)$ will, in general, be of the form

$$w(t) = w_1 \sin (\omega t + \Psi) + w_2 \sin (2\omega t + \Psi_2) + w_3 \sin (3\omega t + \Psi_3) + \cdots$$

If $w(t)$ is assimilated to its first harmonic, the function

$$N(x_1, \omega) = (w_1/x_1)e^{i\Psi}$$  

(24-1)

can be considered as a generalized transfer function for the nonlinear element $N$. It has a magnitude $B(x_1, \omega) = w_1/x_1$ and a phase angle $\Psi(x_1, \omega)$, and can be graphically represented by a family of generalized transfer loci $N(x_1, \omega)$ graduated in $\omega$, one locus for each value of $x_1$. These loci can, of course, be represented as Nyquist or Nichols loci, or as inverse loci.

A particular case of considerable importance is that in which the com-

\(^1\) The only communication published in this connection, so far as we are aware, is by E. C. Johnson, "Sinusoidal Analysis of Feedback-control Systems Containing Nonlinear Elements," *Trans. AIEE*, 71: 169–181 (1952).
plex quantity $N$ is independent of the frequency $\omega$ and is purely dependent on the input amplitude $x_1$. In this case, the generalized transfer function is termed the describing function of the element $N$, and the generalized transfer loci can be replaced by one locus, graduated in input magnitude. This locus

$$N(x_1) = B(x_1)e^{i\Psi(x_1)} \quad (24-2a)$$

is termed the describing locus of the system. The locus

$$C(x_1) = -\frac{1}{N(x_1)} = -\frac{1}{B(x_1)}e^{-i\Psi(x_1)} \quad (24-2b)$$

can also be considered. For reasons that will become clear later, it is called critical locus when the nonlinear element $N$ is a component of a closed-loop system.

24.1.4. Application to Servo Systems Incorporating One Nonlinear Element. Consider now the case of a servo system incorporating one nonlinear element as defined in Sec. 22.4.4. Assume, furthermore, that the nonlinear element is not frequency-dependent, that is, that the element is characterized by a describing function $N(\varepsilon_1)$, or by a critical locus, graduated in magnitude. Such a system consists of two sorts of components: (1) the linear components, which are completely characterized by the transfer locus $L(j\omega)$; and (2) the nonlinear component, which is completely characterized (within the limits of the first-harmonic approximation) by its critical locus $C(\varepsilon_1)$. The generalized $R/\varepsilon$ transfer function for the system as a whole thus appears as the product

$$R/\varepsilon(\varepsilon_1,\omega) = N(\varepsilon_1) \times L(j\omega)$$

the first factor being a function of $\varepsilon_1$ alone and characterizing the nonlinear component, the second factor being a function of $\omega$ alone and characterizing the linear components. Thus the data of the two loci $L(j\omega)$, graduated in frequency, and $C(\varepsilon_1)$, graduated in magnitude, should give a complete picture of the servo system under consideration. The corresponding techniques will be outlined in the following sections.

24.2. DESCRIBING FUNCTIONS OF COMMON NONLINEAR ELEMENTS

24.2.1. On-Off Element without Hysteresis. When an on-off element with a dead zone $\Delta$ is subjected to an input $x_1 \sin \omega t$ with a period $T = 2\pi/\omega$, the output $w(t)$ is zero whenever $x_1 < \Delta/2$. If $x_1 > \Delta/2$, the output is a sequence of alternately positive and negative pulses (Fig. 24-3). The first harmonic of the Fourier expansion of $w(t)$ is $a_1 \sin \omega t + a_1' \cos \omega t$ with

$$a_1 = \frac{\omega}{\pi} \int_0^T w(t) \sin \omega t \, dt \quad a_1' = \frac{\omega}{\pi} \int_0^T w(t) \cos \omega t \, dt$$
Both integrals are easy to compute, since \( w(t) \) is zero except between the values \( t = \alpha \) and \( \beta \) and \( t = \gamma \) and \( \delta \), where

\[
\sin \omega \alpha = \sin \omega \beta = \frac{\Delta}{2x_1} \quad \sin \omega \gamma = \sin \omega \delta = \frac{-\Delta}{2x_1}
\]

Thus one finds

\[
a_1 = \frac{2M}{\pi} (\cos \omega \alpha - \cos \omega \beta) \quad a'_1 = \frac{2M}{\pi} (\sin \omega \beta - \sin \omega \alpha) \quad (24-3)
\]

In the present case (\( \lambda = 0 \)), one has \( \omega(\alpha + \beta) = \pi \), whence

\[
a_1 = \frac{4M}{\pi} \cos \omega \alpha = \frac{4M}{\pi} \left[ 1 - \left( \frac{\Delta}{2x_1} \right)^2 \right]^{1/4} \quad a'_1 = 0
\]
Fig. 24-5. Describing function $N(x_1) = B(x_1)$ for on-off element with dead zone $\Delta$. Abscissas are $x_1/\Delta$, ordinates are $B\Delta/M$.

In other words, the first harmonic is in phase with the input and its magnitude is equal to $a_1$. Hence

$$B = \frac{4M}{\pi x_1} \left[ 1 - \left( \frac{\Delta}{2x_1} \right)^2 \right]^{1/4} \quad \Psi = 0 \quad (24-4)$$

The describing function $B$ is plotted vs. $x_1$ in Fig. 24-4 for different values of $\Delta$. Figure 24-5 shows the same describing function with nondimensional variables.

These describing functions are easy to interpret. The quantity $B$ is zero so long as $x_1 < \Delta/2$, that is, so long as the forcing signal remains inside the inactive zone of the on-off element. As $x_1$ is increased above $\Delta/2$, the magnitude $B$ of the describing function quickly increases, the output being at its maximum value $M$ for all the instants at which $x(t)$ exceeds $\Delta/2$. For large values of $x_1$, on the contrary, the presence of a maximum instantaneous value $M$ for the output, however large the input amplitude may be, results in a rapid fall-off of the describing function. A maximum of $B$ occurs at $x_1 = 0.707\Delta$. Its value is $B_{\text{max}} = 4M/\pi\Delta$.

The corresponding critical loci are portions of the negative real axis, graduated in $x_1$ in accordance with the equation

$$|C(x_1)| = \frac{1}{B(x_1)} = \frac{\pi x_1}{4M} \left[ 1 - \left( \frac{\Delta}{2x_1} \right)^2 \right]^{-1/4} \quad (24-5)$$

The critical point $C(x_1)$ lies at minus infinity when $2x_1 < \Delta$. It then jumps to a minimum distance for $x_1 = 0.707\Delta$ and goes back to infinity (Fig. 24-6a). In Nichols coordinates the critical locus is a portion of the $-180^\circ$ phase vertical line (Fig. 24-6b).

1 These results are identical with the first example of Sec. 4.5.1 if the beginning of the first positive pulse is taken as the time-origin, that is, if one sets $t - \alpha = t_1$. 
24.2.2. On-Off Element with Hysteresis. When $2x_1 > \Delta + h$, the output \( w(t) \) still consists of alternately positive and negative pulses (Fig. 24-7) and the relations (24-3) hold, with

$$\omega\alpha = \arcsin \Delta + h \quad \omega\beta = \arcsin \Delta - h$$

This time \((h \neq 0)\) the first harmonic is out of phase with respect to \(x_1 \sin \omega t\) by the angle

$$\arctan \frac{a_1}{1} = \arctan \frac{\sin \omega\beta - \sin \omega\alpha}{\cos \omega\alpha - \cos \omega\beta}$$

and its amplitude is

$$Bx_1 = (a_1^2 + a_1^2)^{1/2} \quad 2M (2 - 2 \cos \omega\alpha \cos \omega\beta - 2 \sin \omega\alpha \sin \omega\beta)$$

After elementary trigonometric transformations these equations become

$$B \quad \beta - \alpha = \frac{\pi}{2}$$

The quantities \(B\Delta/M\) and \(\Psi\) are plotted in Fig. 24-8 for various values of \(h/\Delta\). The corresponding critical loci are plotted in Fig. 24-9 as Nichols loci \(-20 \log B + j(180^\circ - \Psi)\). The presence of hysteresis results, essentially, in a phase shift \(\Psi\), and the critical locus is no longer a portion of the \(-180^\circ\) phase axis. The case in which \(h = 0\) is obviously that of Figs. 24-5 and 24-6.
24.2.3. Saturation, Dead Zone. The describing function of a frequency-independent linear component $w = k x$ is obviously $B = k$, $\Psi = 0$. The describing function of a linear component with saturation $x_M$ (Fig. 24-10) is equal to $k$ for $x_1 < x_M$. For $x_1 > x_M$ the output $w(t)$ is

$$w(t) = k x_1 \sin \omega t \quad 0 < t < t_1$$

$$w(t) = k x_1 \sin \omega t_1 \quad t_1 < t < \frac{\pi}{\omega} - t_1 \quad \text{etc.}$$

where $\omega t_1 = \arcsin \left( \frac{x_M}{x_1} \right)$. 

Fig. 24-8. Describing function $N(x_1) = B(x_1) \exp \{ j\Psi(x_1) \}$ for on-off element with hysteresis $h$ and dead zone $\Delta$. 
Hence the magnitude of its first harmonic is

\[ w_1 = \frac{4\omega}{\pi} \int_0^{\pi/2} w(t) \sin \omega t \, dt = \frac{2k}{\pi} x_1 \left( \omega t_1 + \frac{\sin 2\omega t_1}{2} \right) \]

and the describing function is

\[ B(x_1) = k \quad \text{for } x_1 < x_M \]

\[ B(x_1) = \frac{2k}{\pi} \left( \omega t_1 + \frac{\sin 2\omega t_1}{2} \right) \quad \text{for } x_1 > x_M \quad (24-7) \]

with \( \Psi(x_1) = 0 \). The plot of \( B \) vs. \( x_1 \) starts at a constant value \( k \) and then falls off, soon becoming asymptotic to the hyperbola \( B = 4kx_M/\pi x_1 \) (Fig. 24-11). When both saturation \( x_M \) and dead zone \( x_m \) are present
(Fig. 22-20), the describing function $B(x_1)$ depends on three parameters, $x_m, x_M$, and $k$, the maximum value of the output being $w_M = k(x_M - x_m)$. The function $B(x_1)/k$ as a function of $x_1/x_m$ is plotted for different values of $x_M/x_m$ in Fig. 24-12. The case $x_M/x_m$ infinite with $x_m \neq 0$ is that of a linear system with dead zone. Finally, the limiting case $x_M = x_m$, that is, $k$ infinite, is that of an on-off system with a dead zone $\Delta = 2x_m$.

24.2.4. Arbitrary Frequency-independent Nonlinear Characteristic. The result just outlined enables one to obtain the describing function for any frequency-inde-
pendent nonlinear component from its characteristic curve. The latter is approximated by a small number of straight lines and the describing function obtained step by step in each interval. Consider the curved characteristic shown in Fig. 24-13, approximated by four straight lines whose slopes are \( k_1, k_2, k_3, \) and zero. The straight line \( k_1 \) is the tangent to the characteristic curve at the origin, the other three straight lines are chosen by inspection to approximate the characteristic as closely as possible.

**Fig. 24-13.** Straight-line approximation of a curved characteristic.

Let us call \( S_1, S_2, \) and \( S_3 \) the abscissas of the corresponding break points. The output \( w(t) \), for \( x_1 > S_3 \), will be:

\[
\begin{align*}
&k_1x_1 \sin \omega t \\
&k_1x_1 \sin \omega t_1 + k_2x_1(\sin \omega t - \sin \omega t_1) \\
&k_1x_1 \sin \omega t_1 + k_3x_1(\sin \omega t_2 - \sin \omega t_1) + k_2x_1(\sin \omega t - \sin \omega t_2) \\
&k_1x_1 \sin \omega t_1 + k_2x_1(\sin \omega t_3 - \sin \omega t_1) + k_3x_1(\sin \omega t_2 - \sin \omega t_3) + k_2x_1(\sin \omega t_3 - \sin \omega t_2)
\end{align*}
\]

where

\[
\begin{align*}
\sin \omega t_1 = \frac{S_1}{x_1} & \quad \sin \omega t_2 = \frac{S_2}{x_1} & \quad \sin \omega t_3 = \frac{S_3}{x_1}
\end{align*}
\]

The first harmonic of the output can easily be computed. The describing function is

\[
\begin{align*}
B(x_1) &= k_1 \quad x_1 < S_1 \\
B(x_1) &= (k_1 - k_2)f_1 + k_2 \quad S_1 < x_1 < S_2 \\
B(x_1) &= (k_1 - k_2)f_1 + (k_2 - k_3)f_2 + k_3 \quad S_2 < x_1 < S_3 \\
B(x_1) &= (k_1 - k_2)f_1 + (k_3 - k_2)f_2 + k_3f_3 \quad S_3 < x_1
\end{align*}
\]

where \( f_i \) is the describing function of a linear element with saturation (Sec. 24.2.3) divided by \( k_i \):

\[
f_i = \frac{2}{\pi} \left( \omega t_i + \frac{\sin 2\omega t_i}{2} \right) \quad i = 1, 2, 3
\]

The functions \( f_i \) are directly plotted vs. \( x_1/S_i \) for different values of \( k_i \) in Fig. 24-14. This enables a rapid computation of the describing function. The following steps are thus involved in the practical procedure:

1. Perform a broken-line approximation for the characteristic curve of the nonlinear component, the first straight line being the tangent at the origin. Note the slopes \( k_1, k_2, \ldots \) of the lines and the abscissas \( S_1, S_2, \ldots \) of their intersections.

2. In the first interval \((0,S_1)\) the describing function is \( k_1 \). In the second interval \((S_1,S_2)\) the describing function is obtained as \((k_1 - k_2)f_1 + k_2\), the quantity \( f_1 \) being read from the curve of Fig. 24-14 which corresponds to the proper value of \( k_1 - k_2 \) (note that in Fig. 24-14 the abscissa is \( x_1/S_1 \)). The same procedure applies to the following intervals \((S_2,S_3), \) etc.

This method yields with ease results of very considerable practical value. Its accuracy is of the same order of magnitude as that obtained when using the concept

\(^1\) After an unpublished paper by J. Stern, of the École Supérieure de l’Aéronautique Paris.
Fig. 24-14. Chart for obtaining the describing function of any frequency-independent nonlinear element.

of the describing function. Three to five broken lines are sufficient in most practical cases.

24.3. STABILITY OF SERVO SYSTEMS WITH ONE NONLINEAR ELEMENT

24.3.1. General Theory. It has been seen in Chap. 16 (see Sec. 16.3, the left-hand criterion) that a linear servo system with an open-loop
transfer function $K G(s) = \frac{R(s)}{E(s)}$ is stable if the critical point $-1/K$ lies at the left of the $G(j\omega)$ locus when the latter is traced out in the direction of increasing frequencies. In this case, the servo system will revert to its steady state after a disturbance. Conversely, if the $-1/K$ point lies at the right of the $G(j\omega)$ locus, the servo system is unstable and gives rise to phenomena which grow exponentially with respect to time.

These considerations can be extended to the case of a servo system incorporating one nonlinear element $N$ characterized by a describing function $N(\varepsilon_1)$. A generalized open-loop transfer function

\[
\frac{R}{E} = N(\varepsilon_1)L(j\omega)
\]

can be considered. If now $\varepsilon_1$ is assumed to be constant, $N_1(\varepsilon_1)$ is a fixed quantity, and one may consider the system as a linear servo system with the open-loop transfer-function $N(\varepsilon_1)L(s)$. The system will thus be stable for the amplitude $\varepsilon = \varepsilon_1$ if the critical point $-1/N(\varepsilon_1)$ is on the left of the $L(j\omega)$ locus. It will be unstable for $\varepsilon = \varepsilon_1$ if the $-1/N(\varepsilon_1)$ point lies on the right of the $L(j\omega)$ locus. If the critical point lay on the locus, at a point $\omega = \omega_0$, the servo system would be oscillatory for $\varepsilon = \varepsilon_1$; that is, once initiated, an oscillation of amplitude $\varepsilon_1$ and frequency $\omega_0$ would sustain itself.

These considerations form the basis of the discussions that follow. The fundamental difference should be emphasized with respect to linear systems, which are intrinsically stable or unstable: to refer to the stability of a nonlinear servo system, it is necessary to specify the conditions prevailing (in this case, the amplitude $\varepsilon_1$).

**Important Note.** When using Nichols loci, care must be taken to interchange the words “right” and “left” in enunciating the left-hand criterion. The reason for this is that the transformation from the Nyquist plane to the Nichols plane is (at least in the case of usual charts) a conformal transformation followed not only by a magnification of scales (leading to nonorthogonality of contours) but also by a change in the positive sense of the angles (Sec. 16.3.1, Remark).

**24.3.2. First Application: Hunting.** Consider a regular linear servo-mechanism with an open-loop transfer function $K G(s)$. Suppose that the system is subjected to saturation: that is, that it can be represented by the general diagram of Fig. 24-15, the describing function of the nonlinear element decreasing from one to zero as $\varepsilon_1$ increases above satura-
tion. Thus, the critical locus consists of that part of the negative real axis which extends from \(-1\) to infinity (see Fig. 24-16 in which the arrows indicate the direction of increasing \(\omega\) and \(\varepsilon_1\)).

If the gain \(K\) is sufficiently increased, the \(KG(j\omega)\) locus will go beyond the \(-1\) point and, therefore, intersect the critical locus at a point \(C\) which corresponds to the amplitude \(\varepsilon_c\) on the critical locus and to the frequency \(\omega_c\) on the transfer locus. Application of the left-hand criterion shows that the system is stable for amplitudes greater than \(\varepsilon_c\) and unstable for amplitudes smaller than \(\varepsilon_c\). This means that small values of \(\varepsilon\) tend to increase (which expresses the fact that the system, if saturation is neglected, is an unstable linear system) and that large values of \(\varepsilon\) tend to decrease (as a result of saturation). The quantity \(\varepsilon_c\) is a convergent point of equilibrium; that is, whatever disturbance the system may be subjected to, it will finally oscillate at frequency \(\omega_c\) and amplitude \(\varepsilon_c\). This is the phenomenon of hunting (Sec. 16.2.3).

Mathematically, it is usually said that the intersection of the \(KG(j\omega)\) locus with the critical locus gives rise to a stable limit cycle which is approached by the system, whatever the initial conditions. The frequency of the limit cycle is that for which the argument of \(G(j\omega)\) is \(-180^\circ\). Its amplitude \(\varepsilon_c\), so far as the variable \(\varepsilon\) is concerned, is given by the value of \(\varepsilon_1\) on the critical locus at its intersection with the \(KG\) locus; the amplitude anywhere in the loop can easily be deduced from it: it is \(A(\omega_c)\varepsilon_c\), if \(A(\omega)\) is the modulus of transfer function that relates the variable considered to \(\varepsilon\).

**Note.** It is now clear that hunting is an essentially nonlinear phenomenon with an amplitude that is not a function of the initial conditions. This is an essential difference as compared with the mathematical case of an oscillatory linear servo system.

24.3.3. **Second Application: On-Off Servos.**
Consider an on-off servo with an \(L(s)\) transfer function of the regular type and an on-off element that may have a dead zone \(\Delta\) and hysteresis \(h\). The \(L(j\omega)\) transfer locus and the \(N(\varepsilon_1)\) critical locus are shown in Fig. 24-17. This case is very similar to the case just dealt with (hunting): there is a limit cycle. It is one of the major disadvantages of on-off servo systems; sustained oscillations exist and occur at full torque, which impairs the accuracy of the system and may be a potential danger to the equipment.\(^1\)

The case of Fig. 24-18 derives from that of Fig. 24-17 by increasing the dead zone \(\Delta\), which results in magnifying the critical locus. The system

\(^1\) In certain cases, however, such limit cycles are deliberately aimed at; see Chap. 28.
is always stable; the limit cycle has been eliminated by the increase in dead zone.

Figure 24-19 introduces a new feature. There are two intersections of the $L(j\omega)$ locus with the critical locus: $C'$ and $C''$, with respective amplitudes $\varepsilon'_c$, $\varepsilon''_c$ and frequencies $\omega'_c$ and $\omega''_c$. It is seen that $C''$ corresponds to a stable limit cycle. But this is not the case for $C'$. In fact, the system is stable for $\varepsilon_1 < \varepsilon'_c$ and unstable for $\varepsilon_1 > \varepsilon'_c$. Thus the existence of the intersection $C'$ shows that the system will react differently to small and to large disturbances. The effect of small disturbances will die out in time because of the stability for $\varepsilon_1 < \varepsilon'_c$, whereas larger disturbances will lead to the limit cycle $C''$.

Note that in the absence of hysteresis the limit cycles always occur at a frequency for which the phase of $L(j\omega)$ is $-180^\circ$.†

24.3.4. Other Cases. Other cases may arise. Some typical ones corresponding to an $L(s)$ with double integration are shown in Fig. 24-20. Case $a$ is always unstable. Case $b$ is stable for small disturbances and unstable for large disturbances, but there is no limit cycle. Case $c$ shows, in addition, a limit cycle for small disturbances. Cases $b$ and $c$ correspond to a stabilization of Case $a$ by addition of phase lead; Case $c$ corresponds to Case $b$ except that the dead zone is smaller.

24.3.5. Discussion of the Effect of Threshold, Lag, and Hysteresis in the Case of Regular On-Off Servo Systems. Consider an on-off servo whose open-loop transfer function $L(s)$ is regular, for example

$$L(s) = \frac{K}{s(1 + Ts)(1 + T's)}$$

† This important result had already been pointed out by L. McColl, "Fundamental Theory of Servomechanisms," p. 85, footnote, Van Nostrand, Princeton, N.J., 1945.
where \( T' \) may express the delay of the on-off element (Sec. 22.4.4, Case 2, par. a).

1. **On-Off Element with Negligible Hysteresis \((h = 0)\) and Lag \((T' = 0)\).** If the relay has an inactive zone, there is no intersection of the transfer locus and the critical locus; therefore, there is no limit cycle (Fig. 24-21a). If the relay has no inactive zone, it may be said that there is a point of intersection at the origin, that is, a limit cycle of infinite frequency and zero amplitude (Fig. 24-21b). This case does not, however, correspond to any physical reality.

2. **The Effect of Lag.** In this case, the transfer locus always cuts the negative real axis. If the relay has no inactive zone, there is still a limit cycle (Fig. 24-22) which occurs at the frequency at which the sum of the

\[
\begin{align*}
&\text{Fig. 24-21.} & \text{Fig. 24-22. Limit cycle for regular servo with ideal on-off element.}
\end{align*}
\]

\[
\begin{align*}
&\text{Fig. 24-23. Elimination of limit cycle by increasing dead zone.} & \text{Fig. 24-24. Lag increases amplitude and decreases frequency of limit cycle.}
\end{align*}
\]

phase shifts of the linear components is equal to \(\pi\). Its amplitude is given by the graduation of the critical locus.

These two quantities are basic factors in considering the effect of the various parameters. If the relay has an inactive zone, there is no limit cycle when the inactive zone is above a certain critical value (Fig. 24-23a and b). When the inactive zone of the relay is below the critical value, it is seen that the greater the lag, the lower the frequency and the greater the amplitude of the limit cycle (Fig. 24-24).

3. **The Effect of Hysteresis.** The critical locus in this case is no longer on the real negative axis. Its intersection with the transfer locus (taking account of lag, of course) gives the limit cycle. It can be suppressed if the inactive zone of the relay is high enough (above the critical value) (Fig. 24-25a). Hysteresis has the effect of reducing the limit-cycle frequency, but of generally increasing its amplitude (Fig. 24-25b).

**24.3.6. Remark Concerning the Stability of a System on the Verge of Instability.** The stability of the zero-error position of the system depends upon the position of
the starting point \( \varepsilon_1 = 0 \) of the critical locus with respect to the transfer locus of the linear part \( L(j\omega) \). Consider the case in which this point lies on the stable side, but extremely close to the \( L(j\omega) \) locus, so that a slight change in the parameters of the system, e.g., a small additional lag \( \Delta T \), will suffice to push it over into the unstable region.

It can be seen that two cases are possible.\(^1\) When the critical locus starts toward the stable region (i.e., toward the left of the transfer-locus traced out in the direction of increasing frequencies, see Fig. 24-26a) the presence of the additional lag results in a limit cycle whose amplitude \( \Delta \varepsilon \) is small when \( \Delta T \) is small. If the additional lag disappears, this limit cycle tends toward zero, and the system comes back to its zero-error equilibrium position; that is, there is reversibility. However, if the critical locus starts toward the unstable region (Fig. 24-26b), the presence of the additional lag will result in a limit cycle which already existed before \( \Delta T \) was introduced and whose amplitude is not infinitely small with \( \Delta T \). If the additional lag disappears, the system will not regain its previous equilibrium position but will stay on this limit cycle; that is, there is no reversibility.

These concepts have been rigorously examined by N. Bautin (Sec. 27.1.4). The stability boundary is said to be of the nondangerous type in the first case and of the dangerous type in the second case.

24.2.7. Alternate Method for Determining the Stability of a Limit Cycle. Consider a limit cycle obtained by taking the intersection of the \( L(j\omega) \) transfer locus and the \(-1/N(x_1)\) critical locus. The angular frequency \( \omega = \omega_0 \) and the amplitude \( x_1 = x_0 \) of the limit cycle are the roots of the equation

\[
L(j\omega)N(x_1) + 1 = 0
\]

\(^1\) In his paper "Näherungskriterien für Stabilität und Gefährlichkeit in nichtlinearen Regelkreisen," in "Regelungstechnik: moderne Theorien und ihre Verwendbarkeit," pp. 149–151, Oldenbourg, München, 1957, K. Magnus gives a different approach to this question, and applies it to more general cases.
If the real and the imaginary parts of the left-hand side are $X$ and $Y$, the latter equation becomes

$$X(\omega, x_1) + jY(\omega, x_1) = 0$$

It can be considered that the effect of a disturbance consists of replacing the limit cycle $x_1 = x_0 \exp (j\omega t)$ by a function of time differing from it by a small difference in amplitude $\Delta x_1$ and in frequency $\Delta \omega$, and which is slightly damped (positively or negatively) with a coefficient $\delta$: $x = (x_0 + \Delta x_1) \exp [j(\omega_0 + \Delta \omega + j\delta)t]$. The incremental quantities $x_1$, $\Delta \omega$, and $\delta$ satisfy

$$X(\omega_0 + \Delta \omega + j\delta, x_1 + \Delta x_1) + jY(\omega_0 + \Delta \omega + j\delta, x_1 + \Delta x_1) = 0$$

If this equation is limited to the first term of its Taylor expansion and if the real and the imaginary parts are separately equated to zero, one obtains

$$\frac{\partial X}{\partial x_1} \Delta x_1 + \frac{\partial X}{\partial \omega} \Delta \omega - \frac{\partial Y}{\partial \omega} \delta = 0$$

$$\frac{\partial Y}{\partial x_1} \Delta x_1 + \frac{\partial Y}{\partial \omega} \Delta \omega + \frac{\partial X}{\partial \omega} \delta = 0$$

whence by eliminating $\Delta \omega$

$$\left[ \left( \frac{\partial X}{\partial \omega} \right)^2 + \left( \frac{\partial Y}{\partial \omega} \right)^2 \right] \delta = \left( \frac{\partial X}{\partial x_1} \frac{\partial Y}{\partial \omega} - \frac{\partial Y}{\partial x_1} \frac{\partial X}{\partial \omega} \right) \Delta x_1$$

The condition for the stability of the limit-cycle is that the disturbed function should be positively damped if its amplitude is greater than $x_0$, and negatively if its amplitude is smaller than $x_0$; in other words, the condition for stability is that $\delta$ be of the same sign as $\Delta x_1$. Hence$^1$

$$\frac{\partial X}{\partial x_1} \frac{\partial Y}{\partial \omega} - \frac{\partial Y}{\partial x_1} \frac{\partial X}{\partial \omega} > 0$$

$^1$ This condition was first obtained by J. Loeb, "Phénomènes héréditaires dans les servomécanismes. Un critère général de stabilité," Annales des Télécommunications, 6(12): 346–356 (1951). It was also found by E. Popov, "Priiblshennyj metod
Now, if $L(j\omega)$ and $-1/N(x_1)$ are written as complex numbers

$$L(j\omega) = U(\omega) + jV(\omega) \quad C(x_1) = -\frac{1}{N(x_1)} = P(x_1) + jQ(x_1)$$

then

$$X(\omega,x_1) = U(\omega) - P(x_1) \quad Y(\omega,x_1) = V(\omega) - Q(x_1)$$

and the condition for stability is

$$\frac{\partial U}{\partial \omega} \frac{\partial Q}{\partial x_1} - \frac{\partial V}{\partial \omega} \frac{\partial P}{\partial x_1} > 0$$

This means that the vectorial product $dL/d\omega \times dC/dx_1$ should be positive. In other words, for a stable limit cycle the $L(j\omega)$ locus, traced out in the direction of increasing frequencies, leaves the increasing-$x_1$ portion of the critical locus on the left-hand side.\(^1\) This can be easily checked on the examples given in Secs. 24.3.2 to 24.3.6.

### 24.4. PERFORMANCE AND COMPENSATION

**24.4.1. Damping and Performance.** For linear servomechanisms, stability is not the only consideration; of great importance is the question of how the system approaches its steady-state position, that is, whether the transient performance is acceptable. Information concerning the transient is obtained from the $KG$ locus by noting the resonance frequency $\omega_R$, which characterizes the rapidity of the transient, and the corresponding resonance ratio $Q$, which characterizes its damping.

A somewhat similar situation occurs in a nonlinear servo system. If it is stable, how does it approach its equilibrium position? If it has a limit cycle, how does it tend toward the limit cycle? These questions can be answered by extending the methods applied to linear systems. For each amplitude $\varepsilon_1$ a $Q$ and an $\omega_R$ can be defined (for the “equivalent” linear system at that amplitude) and read from the Nichols chart by noting the $\lambda$ of the Nichols contour tangent to the locus $N(\varepsilon_1)L(j\omega)$ and the frequency at the point of contact. These quantities $Q$ and $\omega_R$ provide information concerning the damping and the rapidity of the transient at the amplitude $\varepsilon_1$. As the transient proceeds, the amplitude $\varepsilon_1$ changes. The new $N(\varepsilon_1)L(j\omega)$ Nichols locus at each amplitude $\varepsilon_1$ is obtained by translation of the $L(j\omega)$ locus by the vector $N(\varepsilon_1)$. When no hysteresis is present, $N(\varepsilon_1)$ is a real quantity $B(\varepsilon_1)$ and the translation is vertical, which makes the procedure extremely simple if the $L(j\omega)$ locus has been drawn on millimetric tracing paper (compare with Sec. 13.3.3).

For example, in the case of Figs. 24-27 and 24-28,\(^2\) it is seen that $Q$ is great for small amplitudes $\varepsilon_1$ and diminishes for large amplitudes ($Q < 1.3$ for $\varepsilon_1 > 0.9$). On the other hand, $\omega_R$ is a maximum near $\varepsilon_1 = 0.3$. This means that major disturbances are adequately damped, but the corresponding transients are slow. Conversely

\(^1\) On the right-hand side if Nichols coordinates are used (Sec. 16.3.1, Remark).
\(^2\) Adapted from Kochenburger, op. cit.
minor disturbances (and also the terminal phase of the transient state following major disturbances) give rise to a transient which is rapid but poorly damped.

24.4.2. Compensation. 1. General. As in the case of linear systems, the transfer locus \( L(j\omega) \) can be reshaped in order to improve the system performance with reference to some criteria. In particular, the introduction of compensating networks can help solve the following dilemma: a limit cycle can be done away with by increasing the dead zone of the nonlinear element, but at the expense of the accuracy.

Generally speaking, the \( Q \)- and \( \omega_B \)-\( \varepsilon_1 \) curves are the engineer's guide as he studies how to compensate a servo system with one nonlinear element. By properly reshaping the \( L(j\omega) \) locus, or sometimes the \( N(z_1) \) function, the designer can attempt to obtain desired transient performance for given conditions. The following examples are intended to show how integral and lead compensation can be applied to nonlinear servo systems. Only linear controllers will be considered here. The effect of such a controller with a transfer function \( F(s) \) is essentially to replace \( L(s) \) by \( L(s)F(s) \) (Fig. 22-25).

2. Integral Compensation. Figure 24-29 shows the noncompensated open-loop transfer locus and the new locus after insertion of an undercompensated integral controller:

\[ F(s) = \frac{1 + Ts}{1 + bTs} \quad T = 7, \ b = 10 \]

In the absence of hysteresis it is seen that the critical value of the threshold, i.e., the minimum value necessary for eliminating the limit cycle, is lowered by the presence of the controller in the ratio 6.7 (17 db). On the other hand, it is seen in Fig. 24-29 that the resonance is extremely sharp in the low-frequency region (Q is 7 db when the critical point is the point C, which corresponds to \( \epsilon_k = 5 \)). This could be made up for by introducing a limiter of the amplitude.

3. Phase-lead Compensation. The following example is a good illustration. The discontinuous curve of Fig. 24-30 is the Nichols locus of an actual air-to-ground missile piloted by spoilers (Sec. 22.3.3; the locus takes the mechanical and aerodynamical lags of the spoilers into account). If no compensation is applied, three limit cycles may appear. Only the first (at 3.4 rad/sec) and the third (at slightly more than 8 rad/sec) are stable; the middle intersection represents a separation between disturbances leading to each of these two cycles.

The introduction of a phase-lead correction replaces the uncompensated locus by the locus drawn in continuous line. It is seen that the low-frequency limit cycle has disappeared. The frequency of the high-frequency one is slightly increased (9 rad/sec), and its amplitude is decreased.

**Fig. 24-29. Example of integral compensation.**

### 24.5. ADDITIONAL COMMENTS ON DESCRIBING FUNCTIONS

**24.5.1. Generalization for the Case of Random Inputs.**

1. When a nonlinear component is subjected to a sinusoidal input \( x_1 \sin \omega t \) (Fig. 24-1), the describing function is defined by considering the first harmonic \( w_1 \sin (\omega t + \psi) \) of the output \( w(t) \). When the component is frequency-independent and involves no hysteresis, the describing function is equal

---

to the real quantity $B = \frac{w_1}{x_1}$, which can be considered as the "gain" of the system for the sinusoidal input considered. The function $w_1 \sin (\omega t + \Psi)$ is the first term in the Fourier expansion of $w(t)$. It can be shown as an application of the theory of Fourier series that, in the case just described, the value $w_1$ of the amplitude is that which minimizes the mean value of $[w(t) - w_1(t)]^2$. Therefore, the gain $B(x_1)$ minimizes the mean-square value of $[w(t) - B(x_1)x(t)]^2$ for a sinusoidal $x(t)$.

2. Consider now a random input $x(t)$. It is no longer possible to describe the output $w(t)$, which is random and not periodic, by its first harmonic: a "describing function" cannot be derived from an equation of the type of (24-1). But the condition of minimization can be applied to random functions and a real describing function, or "equivalent gain" $B$, associated with a given random input can be defined by the condition that $B$ should minimize the mean value of $[w(t) - Bx(t)]^2$. The random function $x(t)$ must be described by its probability distribution $P(x)$, which will be assumed independent of time (case of a stationary function). Since time averages are then equal to ensemble averages, the gain $B$ must minimize the ensemble average

$$M = [w(x) - Bx]^2$$

Knowing the probability distribution $P(x)$ and the amplitude characteristic of the nonlinear element $w(x)$, $B$ can be obtained from this condition. Knowing that the mean value of any function $f(x)$ is

$$\int_{-\infty}^{+\infty} f(x)P(x) \, dx$$

it is seen that

$$M = \int_{-\infty}^{+\infty} [w(x) - Bx]^2 P(x) \, dx$$

† The probability distribution is such that $P(x_i) \, dx_i$ is the probability that $x$ lies within the interval $(x_i, x_i + dx_i)$. 

Fig. 24-30. Example of phase-lead compensation.
The value of $B$ which minimizes $M$ is given by
\[
\frac{\partial M}{\partial B} = 0 \quad \text{that is} \quad -2 \int_{-\infty}^{\infty} 2x[w(x) - Bx]P(x)\,dx = 0 \quad (24-8)
\]
whence
\[
B = \frac{\int_{-\infty}^{+\infty} w(x)xP(x)\,dx}{\int_{-\infty}^{+\infty} x^2P(x)\,dx} = \frac{\int_{-\infty}^{+\infty} w(x)xP(x)\,dx}{x^2} \quad (24-9)
\]

**Example.** Assume for the input amplitude $x$ a gaussian distribution with a standard deviation (mean value of $x^2$) $\sigma^2$:
\[
P(x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp \frac{-x^2}{2\sigma^2}
\]
Suppose that $w(x)$ has the form shown in Fig. 24-31, the gain $B$ is found to be
\[
B = \frac{2}{\sigma(2\pi)^{1/2}} \int_{0}^{x_M} \exp \frac{-x^2}{2\sigma^2} \, dx = \frac{2}{(\pi)^{1/2}} \int_{0}^{1/\sigma} \exp (-Z^2) \, dZ
\]
where
\[
\sigma_n = \frac{\sigma \sqrt{2}}{x_M}, \quad Z = \frac{x}{\sigma \sqrt{2}}, \quad \frac{x}{x_M \sigma_n}
\]

The equivalent gain $B$ is shown as a function of $\sigma_n$ in Fig. 24-32, which should be compared with Fig. 24-11. The latter concerns the case of a sinusoidal input. Note that for $\sigma_n \ll 1$ the system behaves like a linear system.

3. Now let the nonlinear element be inserted in the loop of a servo system, where the input $e(t)$ is a stationary random function (Fig. 22-30). Assume that, according to the above theory, the nonlinear element $N$ is replaced by its equivalent gain $B$; the signal $x$ at the input of the nonlinearity is, in terms of Laplace transforms,
\[
\frac{1}{1 + BL(s)} \text{ } E(s)
\]

The equivalent gain $B$ depends on the amplitude probability distribution of $e(t)$ and can be computed if this probability distribution is known. This makes it possible to consider the system as a linear system. In particular, if the spectral density of the random input, $\Phi_e(\omega)$, is known, the spectral density of the output is
\[
\Phi_r(\omega) = \left| \frac{BL(j\omega)}{1 + BL(j\omega)} \right|^2 \Phi_e(\omega) \quad (24-10)
\]
Example of Application. Nikiforuk and West consider the system shown in Fig. 24-33 where:

a. The nonlinear element is the one described in Fig. 24-31.
b. The probability distribution at the input of the nonlinear element is gaussian, with the standard deviation \( \sigma \). This probability distribution is not the same as the distribution at the input \( e \), since

\[
X(s) = \frac{G_1(1 + \frac{T_d}{s})T_s^2}{T_s^2 + BG_1(1 + \frac{T_d}{s})} E(s)
\]

However, it can be reasonably assumed that, if the distribution of \( e(t) \) is gaussian, the distribution of \( x(t) \) can also be approximated by a gaussian function. Therefore

![Figure 24-33](image)

the equivalent gain \( B \) can be derived in terms of \( \sigma_n = \sigma \sqrt{2/x_M} \), and is represented in Fig. 24-32.

c. The spectral density of the input is assumed to be

\[
\Phi_e(\omega) = \frac{a^2}{(1 + T_f^2\omega^2)^3} \quad \text{whence} \quad \overline{\epsilon^2} = \frac{\pi a^2}{2T_f}
\]

The mean-square value of the input to the nonlinear element is then

\[
\sigma^2 = a^2 \int_{-\infty}^{+\infty} \frac{G_1(1 + \frac{T_d\jmath\omega}{s})T_s^2(\jmath\omega)^2}{(T_s^2(\jmath\omega)^2 + G_1B(1 + \frac{T_d\jmath\omega}{s})[T_s^2(\jmath\omega)^2 + 2T_f\jmath\omega + 1])^2} d\omega
\]

which can be computed in terms of \( B \) and of the system parameters. In particular, for a given system, the nondimensional value \( \sigma_n = \sigma/x_M \) is obtained as a function of \( B \) and of \( e_n = (\overline{\epsilon^2}/x_M)^{1/2} \). Figure 24-34 shows \( B \) vs. \( \sigma_n \) for different values of \( e_n \).

Intersection of the curves shown in Figs. 24-32 and 24-34 yields the value of \( B \) and \( \sigma \) for each \( e_n \). \( B \) being thus determined, the power spectral density at any point, in particular at the output, can be computed by means of Eq. (24-10). Typical output spectra are described in Fig. 24-35 for different values of \( e_n \). It is seen that the presence of the nonlinearity has the effect of decreasing the bandwidth and increasing the resonance ratio.

24.5.2. Application of the Root-locus Method. The root-locus method can be extended to servo systems with one nonlinear element characterized by a describing function. When the nonlinear element has no hysteresis, the generaliz-
Fig. 24-35. Typical output spectra. $T$, $T_d$, $T_1$, and $G_1$ have the same values as in Fig. 25-34. Curve $a$ corresponds to the linear case; curves $b$ and $c$ correspond to the nonlinear case with $\varepsilon_1 = 1.60$ and 4.65, respectively. [After P. Nikiforuk and J. West, "The Describing-function Analysis of a Non-linear Servo Mechanism Subjected to Stochastic Signals and Noise," Proc. Inst. Elec. Engrs. (London), Monograph 207 M, 1956.]

tion is straightforward, the effect of the nonlinearity being to multiply the gain by a real factor that is a function $B$ of the error amplitude $\varepsilon_1$. Thus, for given conditions for the system, the corresponding value of the describing function defines a point on the root locus. If this point lies in the left half plane, the system is stable under the conditions considered, and the point of operation tends to move on the root locus in the direction of decreasing $\varepsilon_1$. Conversely, if the point lies in the right-hand half plane, the point will tend to move in the direction of increasing $\varepsilon_1$. These qualitative considerations are valuable only in the region near the imaginary axis. In particular, intersections of the root locus with the imaginary axis correspond to limit cycles with the amplitude determined by the value of $B(\varepsilon_1)$. These limit cycles are stable if smaller values of $\varepsilon_1$ correspond to points in the right-hand half plane and larger values to points in the left-hand half plane.

When hysteresis is present, the discussion is not so straightforward.\(^1\)

\(^1\)This case has been studied by A. P. Paris, "The Analysis and Synthesis of Contactor Servomechanisms," University of British Columbia thesis, 1954, and by C. H. Wilts in his course on servomechanisms given at the California Institute of Tech-
because the describing function of the nonlinear component is complex. A possible procedure consists in plotting the constant-magnitude and constant-argument loci (Secs. 9.3.1 and 9.3.7) for the function $L(s)$ (see Fig. 24-36; recall that the $\Phi = \pi$ locus is the root locus). For each value of $\Phi$, the values of $B(\varepsilon_1)$ that correspond to $\Psi = \pi - \Phi$ are read from the describing locus (Fig. 24-37) and placed on the $\Phi$ locus. It can be seen that the locus of the points thus obtained crosses the imaginary axis at points which correspond to limit cycles (for the case shown in the figure, $P_1$ is a stable and $P_2$ an unstable limit cycle).

24.5.3. Forced Oscillations of Servo Systems Incorporating One Nonlinear Element. 1. General. The oscillations studied in Sec. 24.3 (limit cycles) tend to develop with an amplitude and a frequency which are characteristic of the system and not of the input to which the system is subjected. They may even develop in the absence of any input. Hence they are termed the self-oscillations of the system, or the oscillations of the autonomous system. It may also be of interest to analyze the behavior of the system when it is subjected to a periodic, e.g., harmonic, input, termed the forcing input. It may always be assumed (Sec. 13.1.7, Note 1) that the forcing input is applied at the control point: $e_f(t) = \varepsilon_1 \sin \omega_f t$. The problem is to investigate under what conditions a permanent regime of the periodic type can develop at the forcing frequency, that is, to investigate the possibility of forced oscillations in the system. Such oscillations are usually characterized by the expression for the input to the nonlinear element, i.e., $\varepsilon(t)$ for the system shown in Fig. 22-30. If upper harmonics are neglected, the function $\varepsilon(t)$ in the forced regime has the form $\varepsilon_f(t) = \varepsilon_1 \sin (\omega_f t + \phi)$. The quantities $\varepsilon_1$ and $\phi$ are functions of $\varepsilon_1$ and $\omega_f$; they define the frequency response of the nonlinear system.

For the system shown in Fig. 22-30, $e_f$ and $\varepsilon_f$ are related by

$$\frac{\varepsilon_f}{E_f}(j\omega) = \frac{1}{1 + N(\varepsilon_1)L(j\omega)}$$

whence for $\omega = \omega_f$

$$\frac{\varepsilon_1}{\varepsilon_f} = \frac{1}{1 + N(\varepsilon_1)L(j\omega_f)}$$

$$\phi = \tan^{-1} \frac{1}{1 + N(\varepsilon_1)L(j\omega_f)}$$

The present paragraph follows Dr. Wilts' course notes, pp. 186–189, with the permission of the author.
The first of these equations, in which \( L(j\omega_f) \) is a given complex quantity, can be solved with respect to \( \varepsilon_1 \), thus giving the amplitude of the possible forced oscillations measured at the input of the nonlinear element. Since \( N \) is a function of \( \varepsilon_1 \), the equation is most conveniently solved graphically by plotting the function \( \varepsilon_1 = \varepsilon_1 |1 + N(\varepsilon_1)L(j\omega_f)| \) in the \((\varepsilon_1, \varepsilon_2)\) plane and finding its intersections with parallels to the \( \varepsilon_1 \) axis.

The solutions so obtained can be very complicated. It is often found that many forced oscillations with different amplitudes are possible for given \( \varepsilon_1 \) and \( \omega_f \), but some of them are unstable and do not appear physically. A general discussion is beyond the scope of the present chapter. The following paragraphs are merely intended to explain two important nonlinear phenomena: the cutoff amplitude and the jump, combinations of which account for the behavior of most nonlinear servo systems subjected to periodic forcing inputs.

2. Cutoff Amplitude. The phenomenon of cutoff amplitude arises in systems where forced oscillations with a frequency \( \omega_f/2\pi \) are possible only if the amplitude \( \varepsilon_1 \) of the forcing signal is greater than a critical value \( \varepsilon_c \) termed cutoff amplitude, or synchronization threshold. This happens when the relation that exists between \( \varepsilon_1 \) and \( \varepsilon_2 \) has the shape shown in Fig. 24-38: no forced oscillation can occur when \( \varepsilon_1 < \varepsilon_c \), whereas two are possible for \( \varepsilon_c < \varepsilon_1 < \varepsilon_c' \) and one for \( \varepsilon_1 > \varepsilon_c' \).

This is the case for most on-off servo systems. Consider the case of an ideal on-off control with an amplitude \( M \):

\[
N(\varepsilon_1) = \frac{4}{\pi} \frac{M}{\varepsilon_1} = \frac{M'}{\varepsilon_1}
\]

where \( M' = 4M/\pi \). One has

\[
\varepsilon_f = \frac{1}{1 + (M'/\varepsilon_1)L(j\omega_f)} \quad \text{whence} \quad \varepsilon_1 = |\varepsilon_1 + M'L(j\omega_f)|
\]

The latter equation can easily be interpreted graphically. If \( P_f \) is the \( \omega = \omega_f \) (fixed) point of the \( M'L(j\omega) \) locus, the quantity on the right-hand side is the distance to that point from the \(-\varepsilon_1 \) point on the real axis. As \( \varepsilon_1 \) varies, this distance goes through a minimum which is the vertical distance of \( P_f \) from the real axis. Hence \( \varepsilon_c = M'[V(\omega_f)] \), where \( jV(\omega) \) is the imaginary part of \( L(j\omega) \). If \( \varepsilon_1 < \varepsilon_c \), there can be no forced oscillation with the frequency \( \omega_f/2\pi \). If \( \varepsilon_1 > \varepsilon_c \), the circle centered at \( P_f \) with a radius \( \varepsilon_1 \) has two intersections with the real axis, giving rise to two possible forced oscillations.\(^1\)

If different values of \( \omega_f \) are considered and if the \( L(j\omega) \) locus is assumed to have the shape shown in Fig. 24-39, a cutoff amplitude \( \varepsilon_c \) exists at each forcing frequency. When \( \omega_f \) is close to the frequency \( \omega_0 \) at which self-oscillations occur, \( \varepsilon_c \) becomes very small. In the limit the synchronization threshold is zero when

\(^1\) If one of the intersections lies on the positive real axis, it does not give rise to an oscillation, since \(-\varepsilon_1 \) is an inherently negative quantity. When the two intersections lie on the negative real axis, it can be shown that the oscillation with smaller amplitude is unstable and therefore does not appear physically (Sec. 26.4.3, par. 2).
The self-oscillation; in fact the latter does not require the presence of a synchronizing signal \( e_f \) to develop.

3. The Jump Phenomenon. When a nonlinear system is subjected to sinusoidal inputs with constant frequency \( \omega_f/2\pi \) and variable amplitude \( e_1 \), it is often found that the amplitude \( \varepsilon_1 \) of the forced oscillation varies discontinuously. This happens when the \( \varepsilon_1 \) vs. \( e_1 \) plot for a given frequency \( \omega_f/2\pi \) has the shape shown in Fig. 24-40 with vertical tangents at \( e_1 = e'_1 \) and \( e''_1 \). When \( e'_1 < e_1 < e''_1 \), there are three possible values for \( \varepsilon_1 \). If in the course of the test the input amplitude is increased from zero (Fig. 24-41), the smallest of the three values is first obtained; but if \( e_1 \) is increased beyond \( e''_1 \), the operating point jumps upward onto the upper branch and suddenly goes over to the greatest of its three possible values. Conversely, if \( e_1 \) is decreased from large values, \( \varepsilon_1 \) will first assume the largest of its three possible values and as \( e_1 \) is decreased beyond \( e'_1 \) a jump downward will occur. To summarize, the \( \varepsilon_1 \)-vs.-\( e_1 \) variation follows a hysteresis loop. The points on the middle branch between \( e'_1 \) and \( e''_1 \) correspond to unstable oscillations that do not appear physically.

The jump phenomenon is frequently encountered in nonlinear mechanics. Examples can be found in the servo field (Prob. 57).

---


24.5.4. Validity of the First-harmonic Approximation. The validity of the method described above is unfortunately conditioned by the validity of the approximation, which consists in systematically neglecting the effect of harmonics of second and higher orders.

In the case of systems with one degree of freedom, this approximation is generally allowable. In all cases which the authors personally have encountered, the error exceeded 10 per cent in relation to the exact calculations in only exceptional cases; for example, for limit-cycle frequencies, results are usually exact to within 5 per cent. For evaluating the performance of the system, the qualitative results of the method are appreciably confirmed by experiment. Such is the case for the discussion of the stability of limit cycles as compared to rigorous methods and to experiment.

However, cases may arise in which the first-harmonic approximation leads to incorrect results (see Sec. 26.2.5). Fortunately most of these cases are examples that have been thought up on purpose. In any event, there is no sure criterion that specifies under what conditions and within what limits the first-harmonic approximation can be applied safely. As a result, the method should be applied with great circumspection when studying systems with several degrees of freedom or when investigating the low-frequency behavior.

24.5.5. The Usefulness of the Method. Taking into account this reservation, the method is of very considerable usefulness. In the first place, the introduction of the notion of a describing function makes it possible to visualize the effect of variations in different parameters more intuitively than by calculation. Secondly, the method makes it possible not only to determine the limit cycle of an on-off servo system, but to determine the system's behavior at least qualitatively.

Methods based on first-harmonic approximation are, in fact, currently in use in design laboratories, at least for reaching first approximations. Multiple-input-amplitude frequency-response techniques have been developed for this purpose.
CHAPTER 25

THE POINCARÉ APPROACH IN THE PHASE PLANE

Summary

1. The phase plane.
2. Application to simple nonlinear servo systems.
3. Optimum on-off servo systems.

25.1. THE PHASE PLANE

25.1.1. General Remarks and Definitions. The notion of the phase plane, introduced by H. Poincaré, may be usefully applied to numerous linear and nonlinear engineering problems. Consider a physical system whose degree of freedom is determined by a differential equation which (time being the independent variable) may be written as:

\[ \frac{d^2x}{dt^2} = f(x, \frac{dx}{dt}) \]  \hspace{2cm} (25-1)

If

\[ \frac{dx}{dt} = y \]  \hspace{2cm} (25-2)

then Eq.(25-1) is equivalent to the system

\[ \frac{dx}{dt} = y \]
\[ \frac{dy}{dt} = f(x,y) \]

This is a particular case of the more general system which may be written in the form:

\[ \frac{dx}{dt} = P(x,y) \]  \hspace{2cm} (25-3)
\[ \frac{dy}{dt} = Q(x,y) \]

which is linear if \( P \) and \( Q \) are of the form

\[ P = ax + by \quad Q = cx + dy \]

\( (a, b, c, \text{ and } d \) being independent of \( x \) and \( y \), and if possible of time) and which is nonlinear otherwise.

At any given moment the physical condition of the system is defined
by the values of $x$ and $y$, that is, by a point on the coordinate plane of $x$ and $dx/dt$, called the phase plane. The evolution of the system with respect to time is represented by a curve in the phase plane, graduated in $t$; such a curve is known as the phase trajectory of the system and is an integral curve of the system represented by Eq. (25.3). The examination of the system may, therefore, be limited to that of the trajectories of its representative points in the phase plane, i.e., of the integral curves of the system of Eq. (25.3).

Experimentally, the phase plane can be visualized by applying to the deflector plates of a cathode-ray oscilloscope voltages proportional to $x$ for the horizontal deflection and to $dx/dt$ for the vertical deflection; the phase trajectories are then shown on the screen.

25.1.2. The Tracing of Trajectories in Practice. It is possible to trace to a first approximation the trajectories of a system in the phase plane in the following manner. It will be noted that the slope of the tangent to the trajectory which passes through the point $(x,y)$ is $dy/dx = Q/P$. Several tangents may be drawn, and the broken lines thus obtained can then be smoothed; this is known as the method of isoclines.

It can be shown¹ that it is even possible to quantify the slopes; at each point of intersection the slope of the tangent is taken as being whichever of the four numbers $0, 1, -1, \text{ or } \infty$ is nearest to the exact value; for example,

\[
\begin{align*}
0 & \quad \text{if} \quad -\frac{1}{2} < \frac{Q}{P} < \frac{1}{2} \\
1 & \quad \text{if} \quad \frac{1}{2} < \frac{Q}{P} < 2
\end{align*}
\]

and so on. Using this method, an open trajectory remains open, a closed trajectory remains closed, and this enables qualitative results to be obtained.

25.1.3. Special Case. Liénard's Construction.² Consider a system of the second order with a nonlinear restoring factor. Its equation

\[
\frac{d^2x}{dt^2} + 2x \frac{dx}{dt} + f(x) = 0
\]

may be written

\[
\begin{align*}
\frac{dx}{dt} & = y \\
\frac{dy}{dt} & = -f(x) - 2ay
\end{align*}
\]

If it is assumed that $u = y + 2zx$, these equations become, in the Liénard plane $(x,u)$,

¹ This property, demonstrated by S. Birkhoff, was pointed out to the authors by L. Gauthier.
\[
\frac{dx}{dt} = u - 2xx \\
\frac{du}{dt} = -f(x) \\
\frac{du}{dx} = -\frac{f(x)}{u - 2xx} \\
\left[ \left( \frac{dx}{dt} \right)^2 + \left( \frac{du}{dt} \right)^2 \right]^{1/2} = [f^2(x) + (u - 2xx)^2]^{1/2}
\]

From this can be derived a simple construction for the normal \(MN_1\) at any point \(M(x_1, u_1)\) (Fig. 25-1). In particular, it is obvious that a trajectory has a vertical tangent at every point where it cuts the \(x\) axis.

**Remark.** This method is more generally valid for an equation of the form

\[
\frac{dx}{dt^2} + g(x) \frac{dx}{dt} + f(x) = 0
\]

in which \(\frac{dx}{dt} = z\) and in which a change of variable is effected, namely

\[
y = z + G(x) \quad \text{with} \quad G(x) = \int_0^x g(x) \, dx
\]

whence

\[
\frac{dy}{dx} = \frac{f(x)}{y - G(x)}
\]

It is thus possible to analyze the effect of nonlinear damping.\(^1\)

### 25.1.4. Interpretation of Trajectories.

By following the course of the point \((x, y)\) on its trajectory graduated in \(t\), an idea may be obtained of the behavior of the system with respect to time. The abscissa gives the position of the moving point, and the ordinate its speed. The time taken to cover an arc \(AB\) of the phase trajectory is given by

\[
t = \int_{x_A}^{x_B} \frac{dx}{dx/dt}
\]

and is the area included between the \(x\) axis, the straight lines \(x = x_A\) and \(x_B\), and the inverse axial phase trajectory curve. It may be noted also that the phase velocity at the point \((x, y)\), namely \((x^2 + y^2)^{1/2}\), is given by its distance to the origin.

### 25.1.5. Application to a Second-order Linear System.

To assist the reader in assimilating these concepts, we shall give the phase-plane representation of the evolution of a second-order linear system (Chap. 6) when the system is left to itself under

---

nonzero initial conditions (free oscillations), e.g., unit-step response when the initial velocity condition is \( y = 0 \). The equation

\[ \frac{d^2x}{dt^2} + 2\omega_n \frac{dx}{dt} + \omega_n^2 x = 0 \]

may be written in the form

\[ \frac{dx}{dt} = y \]
\[ \frac{dy}{dx} = -2\omega_n u \]

from which we see that

\[ \frac{dy}{dx} = -\omega_n^2 x - 2\omega_n \]

The isocline of slope \( q \) is a straight line passing through the origin; its equation is

\[ x = q + 2\omega_n \]

For simplification, take \( \omega_n = 1 \) (change of time scale). It is easily seen from Fig. 25-2 that:

1. If \( z = 0 \), the tangent to the trajectory at any point is perpendicular to the radius vector; the trajectories are circles. This is the case for an oscillatory system (sinusoidal oscillation at constant amplitude).

Fig. 25-2. Phase portrait of linear second-order system (z is the damping ratio).

2. If \( z > 0 \), the trajectory, from the point corresponding to the initial conditions, is a spiral about the origin. This is the case for damped oscillations. But in proportion as \( z \) increases, the trajectory tends toward the origin without spiraling. This is the case for an aperiodic system (\( z > 1 \)).

25.1.6. Application to the Case of Coulomb Friction. Consider a second-order system (mass and spring, for example) in which the frictional force does not develop in proportion to the speed (viscous friction), but which involves a **coulomb friction**
which develops practically constantly in magnitude but whose sign is opposite to that of the speed:

\[ F = -M \text{ sign } \frac{dx}{dt} \]

The equation of the system may be written

\[ \frac{d^2x}{dt^2} x = -\frac{dx}{dt} \]

(25-4)

and it is therefore a nonlinear system. It is possible to examine the system from its behavior under initial conditions up to the moment when \( dx/dt \) becomes zero (the moment of "change of state") by integrating Eq. (25-4) and subsequently integrating the equation with the opposite sign under new initial conditions, and so on. But the phase-plane method gives a much more intuitive and expressive view of the situation. By writing

\[ x + M \text{ sign } \frac{dx}{dt} = u \]

that is, by bringing the point

\( A (x = h, y = 0) \)

or the point \( B (x = -h, y = 0) \) to the origin, according to the sign of \( y \), the equation may be written as

\[ \frac{du}{dt} = y \quad \frac{dy}{dt} = -u \]

and hence \( u^2 + y^2 = \text{const} \)

If, therefore, initial conditions correspond to the point \( M_0 \) (to give a specific example, \( x_0 > 0, y_0 < 0 \), the point \( M \) follows in a clockwise direction the circle with center \( A \) and radius \( AM_0 \) until it meets the \( x \) axis at a point \( M_1 \). From there, it follows, still in the same clockwise direction, the circle with center \( B \) and radius \( BM_1 \) until it meets the \( x \) axis in \( M_2 \); the trajectory then continues along the circle with center \( A \) and radius \( AM_2 \), and so on, and it is composed of semicircles with centers \( A \) and \( B \) alternately (Fig. 25-3). The movement terminates when the trajectory cuts the \( x \) axis between \( A \) and \( B \) (at \( M_4 \) in the figure).

25.2. APPLICATIONS TO SOME SIMPLE NONLINEAR SERVO SYSTEMS

25.2.1. Transient Response of a Special On-Off Servo System. Consider an on-off servo system (Fig. 25-4) in which the motor is controlled by a relay whose inactive zone may be neglected. It is assumed, moreover, that the response is purely inertial, and that the relay and motor lags may be neglected. If proportional-plus-derivative control is assumed, the torque of the motor is given by

\[ C = M \text{ sign } \left( \varepsilon + \lambda \Gamma \frac{dz}{dt} \right) \quad \text{with} \quad M > 0 \]
From this it may be deduced that

\[ J \frac{d^2 r}{dt^2} = M \text{ sign} \left( \varepsilon + \lambda \frac{d\varepsilon}{dt} \right) \]

or

\[ J \frac{d^2 \varepsilon}{dt^2} = J \frac{d^2 \varepsilon}{dt^2} + M \text{ sign} \left( \varepsilon + \lambda \frac{d\varepsilon}{dt} \right) \]

Consider now the transient state of the system when utilized as a regulator \((e = 0)\), from initial conditions \(\varepsilon_0\) and \((d\varepsilon/dt)_0\), at the instant \(t = 0\),

\[ \begin{array}{c}
\varepsilon + \lambda \frac{d\varepsilon}{dt} \\
\end{array} \]

\[ \begin{array}{c}
\varepsilon_0 \\
\end{array} \]

\[ \begin{array}{c}
\text{Compensating} \\
\text{network} \\
\end{array} \]

\[ \begin{array}{c}
\text{Relay} \\
\text{Motor} \\
\text{Inertia} \\
\end{array} \]

\[ \begin{array}{c}
r \\
\end{array} \]

\[ \begin{array}{c}
\text{Fig. 25-4.} \\
\end{array} \]

such that (to take a specific example) \((\varepsilon + \lambda \frac{d\varepsilon}{dt})_0 < 0\). We have, up to the time of the first change of state,

\[ J \frac{d^2 \varepsilon}{dt^2} = J \frac{d^2 \varepsilon}{dt^2} - M \]

from which, integrating twice, we obtain

\begin{align*}
\frac{d\varepsilon}{dt} &= + \frac{M}{J} (t - t_1) \quad \text{with} \quad t_1 = - \frac{J}{M} \left( \frac{d\varepsilon}{dt} \right)_0 \\
\varepsilon - \varepsilon_1 &= + \frac{M}{2J} (t - t_1)^2 \quad \text{with} \quad \varepsilon_1 = \varepsilon_0 - \frac{J}{2M} \left( \frac{d\varepsilon}{dt} \right)_0^2
\end{align*}

The equation of the phase trajectory in the plane \((\varepsilon, d\varepsilon/dt)\), obtained by eliminating \(t - t_1\), is

\[ \varepsilon - \varepsilon_1 = \left( \frac{J}{2M} \right) \left( \frac{d\varepsilon}{dt} \right)^2 \]

and is a parabola with a horizontal axis, vertex \(\varepsilon_1\), and focal distance \(M/J\). The representative point follows the curve from the initial point \(M_0\) in a clockwise direction up to the point \(M_1\), where it meets the line on which the commutation (or change of state) occurs, i.e., the line with equation

\[ \varepsilon + \lambda \frac{d\varepsilon}{dt} = 0 \]

After \(M_1\) the trajectory follows another parabola, having the same axis and focal distance, but with opposite concavity. It is thus seen (Fig. 25-5) that the trajectory of the transient state is formed, when \(\lambda < 0\), of divergent parabolic arcs. If the system were linear, the introduction of a derivative term with \(\lambda < 0\) would indeed correspond to a destabilizing of the system. The result is, therefore, qualitatively the same with an on-off control.
If \( \lambda > 0 \) (the case corresponding to a servo system with conventional control), the phase trajectory is not divergent. It tends to converge (Fig. 25-6), but a close examination of the graphic constructions reveals that the parabolic arcs do not converge to the origin itself. The convergence stops at a point \( P \) on the line on which the change of state occurs (which may be obtained during the first cycle, as is the case in Fig. 25-7), where the slope of the tangent to the parabola has an absolute value higher than that of the line on which the change of state occurs. This commutation line can then no longer be departed from.

An analysis of the situation shows that this paradox is linked with the hypotheses relating essentially to the infinite speed of the change of state. In fact, it will be seen later (Sec. 25.2.2, par. 2) that by adding a commutation lag, the phase trajectory can converge to the origin.

25.2.2. On-Off Servo System with Viscous Friction. 1. Ideal Regulation (without lag or threshold). Consider the servo system described in Sec. 25.2.1, but suppose that there is viscous friction at the output so that

\[
J \frac{d^2 \tau}{dt^2} + f \frac{d\tau}{dt} = M \text{sign} \left( \varepsilon + \lambda \frac{d\varepsilon}{dt} \right) \quad M > 0
\]
In the case of regulator operation \((e = 0)\) we have

\[
J \frac{d^2e}{dt^2} + f \frac{de}{dt} = -M \text{ sign} \left( e + \lambda \frac{de}{dt} \right)
\]

which may be written in the form

\[
\frac{d^2e}{dt^2} + a \frac{de}{dt} = -b \text{ sign} \left( e + \lambda \frac{de}{dt} \right)
\]

It may be shown, as in Sec. 25.2.1, that the trajectories follow characteristic curves with a horizontal asymptote, with an ordinate \(\pm b/a\), divided into two families of opposite concavities (according to the value of the sign function). Curves of the same family are separated one from another by translation parallel to the \(e\) axis. Figure 25-8 shows the two curves of opposite concavities which pass through the origin.

![Fig. 25-8. Trajectories of system shown in Fig. 25-4 when load involves viscous friction. \(H\) is the Hamel locus of the system.](image)

The passage from one characteristic curve to another occurs on the commutation line, or line of change of state, \(e + \lambda \frac{de}{dt} = 0\). Graphically (Fig. 25-9) the presence of the asymptote means that, if \(\lambda\) is negative, the phase trajectories begin by diverging but are limited for an infinite value of \(t\) by a limit cycle consisting of two arcs of these characteristic curves linking up on the commutation line. This cycle corresponds to a unique periodic solution which does not depend on initial conditions. This solution consists of the arcs of the two characteristic curves passing through the point of intersection of the commutation line and a curve \(H\) which is the locus of initial conditions corresponding to a periodic solution.\(^1\)

This curve is defined parametrically by the equations:

\[
e = \frac{b}{a} \left( \tanh \frac{aT}{2} - \frac{T}{2} \right) \quad \frac{de}{dt} = -\frac{b}{a} \tanh aT
\]

\(^1\) The curve \(H\) is the Hamel locus of the system. See Sec. 26.2.
and is completed by symmetry in relation to the origin. The quantity \( T \) is the half-period of the corresponding solution.

The curve \( H \) is shown in Figs. 25-8 and 25-9. It admits of the horizontal asymptote \( \frac{dx}{dt} = -b/a \). It is tangential at the origin to the \( y \) axis and is entirely contained in the fourth quadrant. The line of change of state, therefore, cuts it only if \( \lambda < 0 \). When \( \lambda > 0 \), as in conventional phase-lead compensated servo systems, the trajectories tend to converge; but, as in the case where there is no friction \( f \), they come to an abrupt stop at a point \( P \) on the line of change of state (Fig. 25-9b). This paradox is due, it will be seen, to the idealization of the system, specifically, to the assumption that the lag is negligible.

![Diagram](image)

**Fig. 25-9.** (a) Shows commutation line and limit cycle.

2. General Case: Presence of a Lag \( \tau \). Consider the same servo system, but assume the existence of a constant lag \( \tau \) in the direct loop. The quantity \( \tau \) accounts for the time of change of state of the relay, plus the time for the motor to start up. It can be shown\(^1\) that the points of change of state are situated on two straight lines, symmetrical in relation to the origin. One of these lines is the locus of the positive commutation points (the sign function passing from +1 to -1), and the other is the locus of the negative points of change of state (from -1 to +1). These two lines will be designated as \( \Delta_+ \) and \( \Delta_- \). The equation of \( \Delta_+ \) in the plane \((\varepsilon, d\varepsilon/dt)\) is

\[
\varepsilon + A \frac{d\varepsilon}{dt} = B
\]

with

\[
A = e^{\sigma\tau} \left[ \lambda - \frac{1}{a} (1 - e^{-\sigma}) \right] - \frac{a}{b} B = \tau + (1 - e^{-\sigma}) \left( A - \frac{1}{a} \right)
\]

\(^1\) A property established for the first time, so far as the authors are aware, by B. Hamel, "Contribution à l'étude mathématique des systèmes de réglage par toutou-rien," *C.E.M.V.* no. 17, Service Technique Aéronautique, Paris, 1949.
It will be noticed that the slope of the lines $\Delta$ is different from that of the no-lag line of change of state, the slope of the latter being $-1/\lambda$. But it is seen graphically (Fig. 25-10) that, whatever $\lambda$ may be, the line $\Delta_+$ cuts the Hamel locus $H$, giving a periodic solution. There is, therefore, a definite periodic solution characterized by the intersection of $H$ and $\Delta_+$. If $\lambda < 0$, the point of departure (initial conditions) is inside the limit cycle; if $\lambda > 0$, it is outside.

The properties of the limit cycle, which do not depend on initial conditions, may be deduced from knowledge of the point of intersection of $H$ and $\Delta$. The amplitude is obtained from its coordinates, and the period

![Diagram](image)

**Fig. 25-10.** Commutation lines and limit cycle in presence of lag.

![Diagram](image)

**Fig. 25-11.**

is read off on the $T$ scale of the curve $H$. It is thus seen that, if $b$ is changed, the amplitude of the limit cycle varies, but its period does not. The influence of $\lambda$ can also be discussed (each of the families of lines $\Delta_+$ and $\Delta_-$ passes through a fixed point when $\lambda$ varies).

3. **Another Application: Transient Response of a Linear Servo System with Saturation.** Consider the angular-position servo system in Fig. 25-11. Assume that the motor develops a torque proportional to a linear combination of the error and its derivative,

$$\Gamma = C \left( e + \lambda \frac{de}{dt} \right)$$

but which cannot exceed a saturation value

$$|\Gamma_{\text{max}}| = M \quad (M > 0)$$
Suppose for simplicity that there is no lag at the motor control and that the load is purely inertial. It will also be assumed that \( \lambda \) is positive, as in the case of conventional phase-lead compensation. Such a system is linear so long as \( |C(\varepsilon + \lambda \frac{d\varepsilon}{dt})| < M \) and will be designated as \( S_1 \). The relationship between its input \( \varepsilon \) and error \( \varepsilon \) is

\[
J \frac{d^2\varepsilon}{dt^2} = J \frac{d^2\varepsilon}{dt^2} + C\lambda \frac{d\varepsilon}{dt} + C\varepsilon
\]

But when \( |C(\varepsilon + \lambda \frac{d\varepsilon}{dt})| > M \), it behaves as a nonlinear system (which will be designated as \( S_2 \)) and is defined by

\[
J \frac{d^2r}{dt^2} = M \text{ sign}\left(\varepsilon + \lambda \frac{d\varepsilon}{dt}\right)
\]

It is then identical to the system considered in Sec. 25.2.1.

![Fig. 25-12. Saturation.](image)

The phase trajectories of the systems \( S_1 \) and \( S_2 \) when they act as regulators \([\varepsilon(t) = 0]\) have been previously described. In the case of \( S_1 \), the phase trajectories are spirals of second-order systems (Sec. 25.1.4) which spiral about the origin. They are circles if \( \lambda = 0 \), and they do not spiral if \( \lambda \) is great. In the case of \( S_2 \), the phase trajectories are parabolas with horizontal axes and focal distance \( M/J \) as in Sec. 25.2.1. The separation between the two cases occurs at the changes of state, on the two lines \( \delta_+ \) and \( \delta_- \), which are symmetrical about the origin and of slope \(-1/\lambda\). The line \( \delta_+ \) corresponds to the change of sign from positive to negative; its equation is

\[
\varepsilon + \lambda \frac{d\varepsilon}{dt} = \frac{M}{C}
\]

Between these straight lines are arcs of spirals, and outside them are
arcs of parabolas (Fig. 25-12). The regulator is seen to tend in an oscillatory manner toward its zero position.

25.2.3. Other Generalizations. It is possible to generalize for the following cases:

1. Presence of a threshold; that is, no error torque is developed when

\[ |e + \frac{de}{dt}| < m \]

2. Presence of a restoring torque \( J \frac{d^2r}{dt^2} + f \frac{dr}{dt} + kr \)

3. Presence of a threshold, a restoring torque, and a lag all together

The limit cycles are always obtained from the intersections of the line of change of state with the Hamel locus \( H \), the latter being the locus of initial conditions giving periodic solutions. For example, in the case of a threshold \( m \), the relevant portion of the curve is of the shape shown in Fig. 25-13. It will be noted that the presence of a sufficiently appreciable threshold causes the limit cycle to disappear. In general, a threshold reduces the frequency and increases the amplitude of the limit cycle, when the latter is present.

25.3. OPTIMUM ON-OFF SERVO SYSTEMS

25.3.1. Principle. Consider again the on-off servo system with which Sec. 25.2.1 was introduced (conventional derivative control, negligible lag, maximum torque \( M \), and inertial load). The fact that, when disturbed from its zero-error position, this system comes back to that position with maximum acceleration is itself more satisfactory, from the point of view of rapid correction, than is the behavior of a linear servo system which in fact never functions at 100 per cent efficiency. A possible disadvantage, on the other hand, is that, in the case of an on-off system, there is a risk of overshooting the mark and producing oscil-
lations. This disadvantage can be avoided if changes of state can be systematically forecast so that there is no overshooting; the system begins by shooting toward the zero-deviation position and decelerates strongly at exactly the right moment (response shown in Fig. 25-14).

This condition is fairly simple to achieve in practice in the simple special case considered here. It has been seen, in fact, that the phase trajectories are equal parabolas \( P \) with horizontal axes. If the change of state occurs on one of the two parabolas \( P \) which pass through the origin (that is, \( P_1 \) and \( P_2 \)), the representative point will go directly to the origin and remain there—at least, if the input maintains zero acceleration. The corresponding condition is that the changes of state correspond to the annulling of

\[
\frac{2M}{J} \varepsilon + \left( \frac{d\varepsilon}{dt} \right)^2 \text{sign} \frac{d\varepsilon}{dt} = 0
\]

In these conditions, we have an optimal servo system for a motor torque limited to \( M \) from the point of view of the speed of the return to the position of zero deviation. For this special performance, the system is superior to any linear servo system limited to the same maximum torque.

*Note.* The above result is due to the fact that all motor systems have a limited torque. A comparison has been made, in fact, of all the characteristics of motor torque in relation to the input signal, admitting the same saturation value \( M \); in these conditions, the best system from the point of view of speed of response is that whose characteristic curve is the one shown in Fig. 25-14, but it leads to infinite acceleration derivatives.

It is worth noting also that this torque saturation is contrary to the strict principle of linearity. This explains why, in Fig. 25-12, where a torque saturation is introduced, the purely linear system tends more rapidly toward the origin than the saturated system; the linear system allows of torque values above saturation.

**25.3.2. Practical Applications.** To sum up, the servo system of Sec. 25.3.1 is excellent for input disturbances of high position and speed steps. However, it presents in general two serious disadvantages:

1. Mediocre performance for very different types of input, especially for slowly varying inputs.
2. The presence of a limit cycle, resulting from ever-present lags and small disturbances. This limit cycle always involves maximum torque,
and it is not without danger to the equipment. A remedy is to make the servo operate linearly for small deviations (Fig. 25-15), adding if necessary a certain amount of derivative control. In this way it is possible to combine the advantages of a nonlinear servo system operating with optimum torque (a high speed of correction for strong disturbances) with those of a linear system which operates gradually and gently damps small disturbances.

![Torque Diagram](image)

**Fig. 25-15.** Optimum on-off servo with linear operation in the region of small disturbances.

It is possible\(^1\) to control by means of the variable

\[
\frac{2M}{J} \left( \varepsilon + \lambda \frac{d\varepsilon}{dt} \right) + \left( \frac{d\varepsilon}{dt} \right)^2 \text{sign} \frac{d\varepsilon}{dt}
\]

For small deviations the second term may be neglected. For large deviations it is possible, by limiting the term \(\lambda \frac{d\varepsilon}{dt}\), to reduce this case to that just described. In the phase plane the trajectories after commutation are coincident with the parabolas \(P_1\) and \(P_2\) within the linear term.

The technique of servomechanisms of this type is likely to take great extension in the coming years,\(^2\) together with the development of electronic function generators.

**Remark.** In many practical cases the sign of the acceleration cannot be instantaneously reversed, that is, there can be no discontinuity in the slope of the phase trajectory. In such cases the situation can be analysed by considering trajectories in a three-dimensional phase space.\(^3\)

\(^1\) This concept is due to R. M. Howe, "Automatic Control Course Notes," University of Michigan, Ann Arbor, 1953.

\(^2\) A discussion of this question will be found in J. Loeb, "Les servomécanismes à programme," *Automatisme, 1*(7):235–242 (1956).

CHAPTER 26

OSCILLATIONS OF ON-OFF CONTROL SYSTEMS

Summary

1. Introduction.
2. Self-oscillations of on-off regulators.
3. Forced oscillations of on-off servo systems.
4. Discussion of their stability.

The theoretical and practical importance of periodic phenomena in on-off servo systems has already been pointed out. The approximate method, which consists in neglecting the higher harmonics of the variables involved, provides a rapid and convenient way of studying the oscillations for an on-off servo; it may also give some information concerning the behavior of the system as it approaches a limit cycle or an equilibrium position. In most practical cases, the first-harmonic approximation leads to satisfactory approximations for the frequency and amplitude of limit cycles.

However, at least one disadvantage of this approach should be pointed out: there is practically no means of evaluating the error resulting from use of the method, there is found no criterion that might specify the extent to which the approximation is valid, and under what assumptions it is valid. In the present chapter, a method which does not have that disadvantage is presented. It is based on research work done in Europe in the postwar years independently by different people who conceived similar, if not equivalent, approaches: essentially B. Hamel1 in France, who started from the phase-plane representation, and J. Tsykin2 in Russia, who uses a representation of his own. The present chapter attempts to tie these approaches together and to discuss the close links that exist between them.


26.1. INTRODUCTION

26.1.1. Assumptions. Feedback control systems incorporating one nonlinear element of the on-off type will be considered. As has been shown in Sec. 22.4.4, such systems can be represented by the general block diagram of Fig. 26-1, where \( L(s) \) is a transfer function and \( N \) represents an on-off element. The on-off element \( N \) may have a dead zone \( \Delta \) and hysteresis \( h \); the maximum value of its output is designated by \( M \).\(^1\) The variables \( e(t) \), \( \varepsilon(t) \), \( r(t) \) are the command, the error, and the output, or they are functions of time related to these quantities by linear differential equations (see Figs. 22-26 and 22-27 to 22-30).

26.1.2. Conditions for Periodic Phenomena in the Absence of Dead Zone. 1. First consider the case in which the on-off element has no hysteresis, and suppose that periodic phenomena take place in the system. Let \( T \) be the corresponding period, and \( \omega = 2\pi/T \) the angular frequency. The situation can be analyzed by studying either the phase-plane trajectories or the time responses (see the first row in Table 26-1).

In the phase plane \((e, de/dt)\) periodic phenomena are represented by a limit cycle, i.e., by a closed curve traced out completely in \( T \) sec.\(^2\) A commutation of the on-off element occurs each time the \( e = 0 \) line is crossed. The commutation is from \(-\) to \( +\) when the \( e = 0 \) line is crossed from left to right. Let the instant of such a commutation be taken as the time origin, and \( P_0 \) be the corresponding position of the point \( P \) \((e, de/dt)\). If periodicity at \( T \) is assumed, the point \( P \) will be

\(^1\) The methods presented here can be extended to the case of nonsymmetrical or multistep on-off elements.

\(^2\) This is true whatever the order of the system, since only the periodic regime is considered. In fact, knowledge of the closed cycle that represents \( e \) and \( de/dt \) of a periodic solution also defines all the other time derivatives of \( e \) for that solution.
at $P_0$ after one period, that is at time $t = T = 2\pi/\omega$. If, furthermore, the limit cycle is assumed to be symmetric with respect to the origin (an assumption to be checked later), then $d\varepsilon/dt$ is positive at $t = 0$, and the point $P$ will be at $P_1$ after a half-period $T/2 = \pi/\omega$, $P_1$ being symmetric of $P_0$ with respect to the origin, i.e., being defined by

$$P_1 \left( t = \frac{T}{2} = \frac{\pi}{\omega} \right); \quad \varepsilon = 0; \quad \frac{d\varepsilon}{dt} < 0$$

2. Using time responses, it is seen that, if periodic oscillations are assumed, the output $w$ of the on-off element is a square wave $w_T(t)$ with a period $T$. The response $r(t)$ of the linear part $L(s)$ is given by

$$R(s) = L(s)W_T(s)$$

The function $r(t)$ is fed back and subtracted from the input, so that the on-off element is actuated by $\varepsilon(t) = e(t) - r(t)$, that is, in the case of a regulator, by $\varepsilon(t) = -r(t)$. If the oscillation is supposed to be symmetrical, $\varepsilon(t)$ is zero at each half-period. Furthermore, if the time origin
(t = 0) is taken at the instant of a commutation from − to +, then the instant t = T/2 = \pi/\omega will correspond to a commutation from + to −:

$$\varepsilon \left( \frac{\pi}{\omega} \right) = 0 \quad \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) < 0$$  \hspace{1cm} (26-1)

3. If hysteresis is present in the on-off element, but if the dead zone is still assumed to be negligible (see second row in Table 26-1), similar reasoning can be applied. If t = 0 is taken at the instant of a commutation from − to +, that is, ε(0) being equal to \(\frac{h}{2}\) with \(\frac{d\varepsilon}{dt}\) positive, the condition for the existence of a symmetric periodic oscillation with a period T is

$$\varepsilon \left( \frac{\pi}{\omega} \right) = -\frac{h}{2} \quad \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) < 0$$  \hspace{1cm} (26-2)

26.1.3. Case in Which a Dead Zone Is Present.  1. First consider the case in which the on-off element has no hysteresis (see third row in Table 26-1) and take the time origin at the instant of a commutation from 0 to +. Let \(P_0\) be the corresponding point in the phase plane, with \(\varepsilon = 0\) and \(\frac{d\varepsilon}{dt} > 0\). In the case of a symmetric oscillation with a period T, a commutation from 0 to − occurs at half-period and is represented by the point \(P_1\) symmetric to \(P_0\) with respect to the origin:

$$P_1 \left( t = \frac{T}{2} = \frac{\pi}{\omega} \right): \quad \varepsilon \left( \frac{\pi}{\omega} \right) = -\frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) < 0$$  \hspace{1cm} (26-3a)

This commutation is similar to that which occurs at half-period in the zero–dead zone case. But before t has reached the value T/2, another commutation has occurred from + to 0, at a time

$$t = \rho \frac{T}{2} = \rho \frac{\pi}{\omega} \quad \text{with} \quad 0 < \rho < 1$$

Thus, if \(P'_1\) is the corresponding point in the phase plane, a second set of conditions for periodicity is

$$P'_1 \left( t = \rho \frac{T}{2} = \rho \frac{\pi}{\omega} \right): \quad \varepsilon \left( \rho \frac{\pi}{\omega} \right) = \frac{\Delta}{2} \quad \frac{d\varepsilon}{dt} \left( \rho \frac{\pi}{\omega} \right) < 0$$  \hspace{1cm} (26-3b)

2. Now, if hysteresis h as well as a dead zone is present (see fourth row in Table 26-1), the conditions for periodicity are:

$$P_1: \quad \varepsilon \left( \frac{\pi}{\omega} \right) = -\frac{\Delta - h}{2} \quad \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) < 0$$  \hspace{1cm} (26-4a)

for the commutation at half-period \((t = T/2)\), and

$$P'_1: \quad \varepsilon \left( \rho \frac{\pi}{\omega} \right) = \frac{\Delta - h}{2} \quad \frac{d\varepsilon}{dt} \left( \rho \frac{\pi}{\omega} \right) < 0 \quad (0 < \rho < 1)$$  \hspace{1cm} (26-4b)

for the commutation within the half-period \((t = \rho T/2)\).

26.2. SELF-Oscillations of ON-OFF Regulators

26.2.1. Introductory Example. Consider to begin with the simple case in which \(L(s)\) is of the form \(L(s) = K/(s(1 + \tau s))\) and the on-off element has no dead zone.
Suppose that the system operates as a regulator, that is, \( e(t) = 0 \). After a zero-to-plus commutation the system is described by the equation

\[
\tau \frac{d^2 \epsilon}{dt^2} + \frac{d \epsilon}{dt} = KM \quad \text{or} \quad \tau \frac{d^2 \epsilon}{dt^2} + \frac{d \epsilon}{dt} = -KM
\]

Integration of this equation starting from the initial conditions \( \epsilon_0 = h/2 \), \((d\epsilon/dt)_0 = \epsilon'_0 \) (see point \( P_0 \) in Fig. 26-2) gives

\[
\frac{d\epsilon}{dt} = \epsilon'_0 e^{-t/\tau} - KM(1 - e^{-t/\tau})
\]

\[
\epsilon = \epsilon_0 - K Mt + \tau (\epsilon'_0 + KM)(1 - e^{-t/\tau})
\]

The time \( t_1 \) at which \( \epsilon_0/dt \) has become \( -\epsilon'_0 \) is given by

\[
\epsilon'_0 e^{-t_1/\tau} - KM(1 - e^{-t_1/\tau}) = -\epsilon'_0
\]

whence

\[
\epsilon'_0 = KM \frac{1 - e^{-t_1/\tau}}{1 + e^{-t_1/\tau}} = KM \tanh \frac{t_1}{2\tau}
\]

and \( \epsilon_0 \) can be expressed as a function of \( t_1 \) by

\[
\epsilon_0 = -KM \left[ \tau \tanh \frac{t_1}{2\tau} - \frac{t_1}{2} \right]
\]

These results can also be expressed as follows: Starting from the initial conditions \((\epsilon_0, \epsilon'_0)\) the error and its time derivative have become \( -\epsilon_0 \) and \( -\epsilon'_0 \), respectively, at a time \( t_1 \) such that

\[
\epsilon_1 = -\epsilon_0 = KM \left[ \tau \tanh \frac{t_1}{2\tau} - \frac{t_1}{2} \right]
\]

\[
\left( \frac{d\epsilon}{dt} \right)_{t_1} = -\epsilon'_0 = -KM \tanh \frac{t_1}{2\tau}
\]

(26-5)

In the phase plane the equations (26-5) represent a locus \( H \), graduated in \( t_1 \). This locus has been shown in Fig. 25-8 and is the \( a = 0 \) locus of Fig. 26-6, where \( \tau = 1 \) and \( \omega = \pi/T \). Figure 26-3 immediately shows that this locus intersects the \( \epsilon = -h/2 \) line, provided intersection lies in the lower half plane.
line. It is thus seen that a periodic solution is possible, the period being twice the value of $t_i$ at the point of intersection.

The $H$ locus is the Hamel locus for the system under consideration. It is seen that, the smaller $h$, the smaller $t_i$, that is, the higher the frequency of the periodic solution and the smaller its amplitude. In the limit, the purely theoretical case $h = 0$ corresponds to a limit cycle that would have zero amplitude and infinite frequency.

If $h = 0$ but the system incorporates a controller (Fig. 26-4), then,

$$\frac{dv}{dt} + \tau \frac{dv}{dt} = -KM \text{sign} \left( v + \lambda \frac{dv}{dt} \right)$$

and the periodic solutions are given by the intersection of the Hamel locus with the commutation line $v + \lambda \frac{dv}{dt} = 0$. A limit cycle exists only if $\lambda < 0$ (Fig. 25-9).

![Diagram for Fig. 26-4](image)

26.2.2. Generalization. More generally, suppose $L(s)$ is any transfer function. First, consider the case in which $L(s)$ has no integration and incorporates no pure lag factor $e^{-\tau s}$. Suppose that all the poles of $L(s)$ are single (for this assumption, see Sec. 6.5.4, case 3) so that $L(s)$ can be written as

$$L(s) = \frac{P(s)}{Q(s)} = \frac{A_1}{s - p_1} + \frac{A_2}{s - p_2} + \cdots = \sum_{i=1}^{n} \frac{A_i}{s - p_i}$$

where

$$A_i = \frac{P(s_i)}{Q'(s_i)} \quad Q'(s) = \frac{dQ}{ds}$$

Under periodic conditions with a period $T$, the output $w(t)$ of the on-off element has the shape shown in Fig. 26-5. If the time origin $t = 0$ is chosen at an instant at which $w(t)$ jumps from $-M$ to $+M$ and the system has reached its steady state, the response $r(t)$ of the system to the input $w(t)$ can be obtained in the interval $(0, T/2)$.

In this interval,

$$w(t) = Mu(t) - M \left[ u \left( t + \frac{T}{2} \right) - u(t) \right] + \cdots + M \left[ u \left( t + \frac{2pT}{2} \right) - u \left( t + \frac{(2p - 1)T}{2} \right) \right] + \cdots$$

$$w(t) = M \left[ u(t) + \sum_{i=1}^{n} (-1)^K u \left( t + \frac{KT}{2} \right) - u \left[ t + \frac{(K - 1)T}{2} \right] \right]$$

If $q(t)$ is the unit-step response of the linear part $L(s)$,

$$\mathcal{L} q(t) = \frac{L(s)}{s} \quad q(t) = A_0 + \sum_{i=1}^{n} \frac{A_i}{p_i} e^{p_i t}$$
then the output \( r(t) \) will be, for \( 0 \leq t < T/2 \),

\[
r(t) = Mq(t) - M \left[ q \left( t + \frac{T}{2} \right) - q(t) \right] + \cdots + M \left[ q \left( t + \frac{2pT}{2} \right) - q \left( t + \frac{(2p - 1)T}{2} \right) \right] + \cdots
\]

The general term of this expression can be written as follows:

\[
q \left( t + \frac{2kT}{2} \right) - q \left[ t + \frac{(2k - 1)T}{2} \right] = - \sum_{i=1}^{n} A_i e^{p_i t}(1 - e^{p_i T/2})e^{(2k-1)p_i T/2}
\]

Thus the contribution of the pole \( s = p_i \) to \( r(t) \) will be

\[
r_i(t) = Mq_i(t) + M \frac{A_i}{p_i} e^{p_i t}(1 - e^{p_i T/2})(1 - e^{p_i T/2} + e^{2p_i T/2} \cdots)
\]

\[
= Mq_i(t) + M \frac{A_i}{p_i} \frac{1 - e^{p_i T/2}}{1 + e^{p_i T/2}} = Mq_i(t) - M \frac{A_i}{p_i} e^{p_i t} \tanh \frac{p_i T}{4}
\]

Hence, for \( t = 0 \),

\[
r(0) = \sum_{i=1}^{n} r_i(0) = Mq(0) - M \sum_{i=1}^{n} \frac{A_i}{p_i} \tanh \frac{p_i T}{4}
\]

and similarly,

\[
\left( \frac{dr}{dt} \right)_0 = Ml_0 - M \sum_{i=1}^{n} A_i \tanh \frac{p_i T}{4}
\]

where \( l_0 \) is the initial value \( l(0+) = \lim_{s \to \infty} sL(s) \) of the unit-impulse response \( \mathcal{L}^{-1}L(s) \).

At \( t = T/2 \) the values will be the opposite,

\[
r \left( \frac{T}{2} \right) = -r(0) \quad \frac{dr}{dt} \left( \frac{T}{2} \right) = - \frac{dr}{dt} (0)
\]

Hence the values of \( \varepsilon \) and \( d\varepsilon/dt \) at \( t = T/2 \) [it will be recalled that \( \varepsilon(t) = -r(t) \)] are

\[
\varepsilon \left( \frac{T}{2} \right) = Mq_0 - M \sum_{i=1}^{n} \frac{A_i}{p_i} \tanh \frac{p_i T}{4} \quad q_0 = \lim_{s \to \infty} L(s) = 0 \quad (26-6)
\]
\[
\frac{de}{dt} \left( \frac{T}{2} \right) = Ml_0 - M \sum_{i=1}^{n} A_i \tanh \frac{p_i T}{4} \quad \quad l_0 = \lim_{s \rightarrow -\infty} sL(s) \tag{26-7}
\]

where
\[
A_i = \frac{P(p_i)}{Q'(p_i)} \quad Q'(s) = \frac{dQ}{ds} \quad L(s) = \frac{P}{Q} \tag{26-8}
\]

Equations (26-6) and (26-7) define a Hamel locus graduated in \( T \) in the \( (\varepsilon, \frac{de}{dt}) \) plane. An intersection of the Hamel locus with the \( \varepsilon = -\frac{h}{2} \) line corresponds to a possible periodic oscillation for the system, provided this intersection lies in the \( \frac{de}{dt} < 0 \) half plane.

If the function \( L(s) \) has one integration, the above expressions become
\[
\varepsilon \left( \frac{T}{2} \right) = -M \frac{T}{4} \frac{P(0)}{Q_1(0)} - M \sum_{i=1}^{n-1} A_i \tanh \frac{p_i T}{4} \quad Q(s) = sQ_1(s) \tag{26-9}
\]
\[
\frac{de}{dt} \left( \frac{T}{2} \right) = Ml_0 - M \sum_{i=1}^{n-1} A_i \tanh \frac{p_i T}{4} \quad l_0 = \lim_{s \rightarrow -\infty} sL(s) \tag{26-10}
\]
\[
A_i = \frac{P(p_i)}{p_i Q'(p_i)} \quad Q'_1(s) = \frac{dQ_1}{ds} \tag{26-11}
\]

If two integrations are present, then
\[
\varepsilon \left( \frac{T}{2} \right) = -M \frac{T}{4} \lim_{s \rightarrow 0} \frac{d}{ds} \left( \frac{P}{Q_1} \right) - M \sum_{i=1}^{n-2} A_i \tanh \frac{p_i T}{4} \quad Q(s) = s^2Q_2(s) \tag{26-12}
\]
\[
\frac{de}{dt} \left( \frac{T}{2} \right) = Ml_0 - M \frac{T}{4} \frac{P(0)}{Q_1(0)} - M \sum_{i=1}^{n-2} A_i \tanh \frac{p_i T}{4} \quad l_0 = \lim_{s \rightarrow -\infty} sL(s) \tag{26-13}
\]
\[
A_i = \frac{P(p_i)}{p_i s Q'_2(p_i)} \quad Q'_2(s) = \frac{dQ_2}{ds} \tag{26-14}
\]

Equations (26-6) to (26-14) are the general expressions for the Hamel loci. Intersection of a Hamel locus with the \( \varepsilon = -\frac{h}{2} \) line corresponds to a periodic solution if this intersection lies in the lower half plane \( \frac{de}{dt} < 0 \). More generally, periodic solutions are given by the intersections of the Hamel locus with the change-of-state line \( \varepsilon + \lambda \frac{de}{dt} = -\frac{h}{2} \) in the case of a proportional-plus-derivative controller preceding the on-off element.

26.2.3. Examples. Figures 26-6 to 26-11\(^1\) show the Hamel loci for the following transfer functions:

<table>
<thead>
<tr>
<th>( L(s) )</th>
<th>Figure</th>
<th>( L(s) )</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{(1+s)(1+as)} )</td>
<td>26-6</td>
<td>( \frac{1}{s(s^2 + 2\zeta s + 1)} )</td>
<td>26-9</td>
</tr>
<tr>
<td>( \frac{1}{s^2 + 2\zeta s + 1} )</td>
<td>26-7</td>
<td>( \frac{1}{s^2(1+s)(1+as)} )</td>
<td>26-10</td>
</tr>
<tr>
<td>( \frac{1}{s(1+s)(1+as)} )</td>
<td>26-8</td>
<td>( \frac{1}{s^2} \frac{1+as}{1+s} )</td>
<td>26-11</td>
</tr>
</tbody>
</table>

\(^1\) Drawn by H. Lepage at the École Supérieure de l'Aéronautique, Paris.
Their plotting is straightforward. However, when \( L(s) \) has complex poles, it is advisable to write the \( \tanh \) in trigonometric form.\(^1\) These figures are graduated in \( \omega = 2\pi/T \), that is, in the angular frequencies of the possible oscillations. For most of them the low-\( T \)—i.e., high-frequency—portion has been represented separately on a larger scale.

These curves make possible a complete discussion of the possibility for periodic oscillations in on-off servo systems without dead zone when the \( L(s) \) function has any of the above expressions, and also if derivative control is added in the form \( 1 + \lambda s \) before the on-off element. The graphical discussions are very easy. They enable one to draw qualitative and quantitative conclusions concerning the influence of the parameters \( a \) of the linear part, \( \lambda \) of the controller, and \( h \) of the on-off element on the existence, the period and, to a certain extent,\(^2\) the magnitude of the limit cycle.

\(^1\) This is facilitated by the use of charts for complex hyperbolic tangent, for example A. Kennelly, "Chart Atlas" and "Tables of Complex Hyperbolic and Circular Functions," Harvard University Press, Cambridge, 1914.

\(^2\) "To a certain extent" because the time expression of \( r(t) \) can be obtained only by working it out completely. However, the horizontal distance from the origin of the
Fig. 26-7. Hamel loci for $L(s) = 1/(s^2 + 2ζs + 1), \, ζ < 1, \, M = 1, \, ε = [1/(1 - ζ^2)^{1/2}]A \sin (θ - φ), \, and \, de/dt = [1/(1 - ζ^2)^{1/2}]A \sin θ$, where $\tan φ = (1 - ζ^2)^{1/2}/\zeta$ and $Ae^{jθ} = \tanh (π/2ω) [iζ + j(1 - ζ^2)^{1/2}]$.

26.2.4. Alternative Representation. A somewhat different approach, developed by Russian authors, follows. If symmetric periodic phenomena occur at a frequency $2π/ω$, the output of the on-off element is a square wave that can be expressed as

$$w(t) = Mu(t) - 2Mu \left( t - \frac{π}{ω} \right) + 2Mu \left( t - \frac{2π}{ω} \right) - \cdots$$

if $t = 0$ is chosen to be the beginning of a positive pulse, or, in Fourier form,

point at which the Hamel locus intersects the change-of-state line gives an idea of the amplitude of the limit cycle and allows a comparison between different adjustments of similar systems.

\(^1\) In some very particular cases, $w(t)$ can have a different shape ("Complex oscillations" for Russian authors) by being zero inside the half-period. In such cases, the relevant Fourier expansion should be used, which leads to techniques similar to those outlined in Sec. 26.1.3.
Hamel loci for $L(e) = 1/(e(1+e)(1+e))$, $M=1$

- $e = \frac{\pi}{2\omega} \frac{1}{1-a} \tanh \frac{\pi}{2\omega} \frac{1-a}{1-a} \tanh \frac{\pi}{2\omega}$
- $\frac{de}{dt} = \frac{1}{1-a} (\tanh \frac{\pi}{2\omega} \frac{1-a}{1-a} \tanh \frac{\pi}{2\omega})$

Fig. 28-8.
Hamel loci for $L_0 = \frac{1}{a(a^2 + 2 \xi \omega^2 + 2 \xi + 1)}$, $\xi < 1$, $M = 1$

\[ \varepsilon = \frac{1}{(1-\xi^2)} A \sin(\theta - 2\varphi) + \frac{\pi}{2\omega} \]

\[ \frac{d\varepsilon}{dt} = \frac{1}{(1-\xi^2)} A \sin(\theta - \varphi) \]

\[ \varphi = \arctan \left( \frac{1-\xi^2}{\xi} \right) A e^{\theta} \cosh \left( \frac{\pi}{2\omega} (\xi + \sqrt{1-\xi^2}) \right) \]

Fig. 26-9.
Fig. 26-10. Hamel loci for $L(s) = 1/s^2(1 + s)(1 + as)$, $M = 1$, and

$$
\varepsilon = (1 + a) \frac{\pi}{2\omega} - \frac{1}{1 - a} \tanh \frac{\pi}{2\omega} + \frac{a^2}{1 - a} \tanh \frac{\pi}{2\omega}
$$

$$
\frac{ds}{dt} = -\frac{\pi}{2\omega} + \frac{1}{1 - a} \tanh \frac{\pi}{2\omega} - \frac{a^2}{1 - a} \tanh \frac{\pi}{2\omega}
$$
Fig. 26-11. Hamel loci for $L(s) = (1 + as)/s^3(1 + s)$, $M = 1$, and

$$
\varepsilon = (1 - a) \left( \frac{\pi}{2\omega} - \tanh \frac{\pi}{2\omega} \right) \quad \frac{dx}{dt} = -\frac{\pi}{2\omega} + (1 - a) \tanh \frac{\pi}{2\omega}
$$
\[ w(t) = \sum_{n} \left( \sin \omega t + \frac{1}{2} \sin 3\omega t + \frac{1}{6} \sin 5\omega t \right) \cdot 4M \sin \omega t \]

where \( \sum_{n} \) indicates the summation for all positive odd values of \( n \) (Sec. 4.5.1 and Fig. 4-17).

If the linear element is characterized by its transfer function \( L(s) \),

\[ L(j\omega) = A(\omega)e^{j\Phi(\omega)} \]

the response \( r(t) \) is

\[ r(t) = \frac{4M}{\pi} \left\{ A(\omega) \sin [\omega t + \Phi(\omega)] + \frac{1}{2} A(3\omega) \sin [3\omega t + \Phi(3\omega)] + \cdots \right\} \]

\[ r(t) = \frac{4M}{\pi} \sum_{n} \frac{1}{n} A(n\omega) \sin [n\omega t + \Phi(n\omega)] \]

and its time derivative is

\[ \frac{dr}{dt} = \frac{4M}{\pi} \omega \sum_{n} A(n\omega) \cos [n\omega t + \Phi(n\omega)] \]

Noting that \( \varepsilon(t) = -r(t) \) in the regulator case, one can write the conditions for periodicity (26-1) in the form

\[ \varepsilon \left( \frac{\pi}{\omega} \right) = \frac{4M}{\pi} \sum_{n} \frac{1}{n} A(n\omega) \sin \Phi(n\omega) = -\frac{h}{2} \]

\[ \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) = \frac{4M}{\pi} \omega \sum_{n} A(n\omega) \cos \Phi(n\omega) < 0 \]

These equations can be interpreted as follows: Consider the function of \( \omega \)

\[ \Lambda(\omega) = \frac{1}{\omega} \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) + j\varepsilon \left( \frac{\pi}{\omega} \right) \]  \hspace{1cm} (26-15)

which is a complex quantity whose imaginary and real parts are

\[ \text{Im} \Lambda(\omega) = \varepsilon \left( \frac{\pi}{\omega} \right) \quad \text{Re} \Lambda(\omega) = \frac{1}{\omega} \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega} \right) \]  \hspace{1cm} (26-16)

The conditions for a periodic solution become

\[ \text{Im} \Lambda(\omega) = -\frac{h}{2} \quad \text{Re} \Lambda(\omega) < 0 \]

Now, if we draw the vector \( \textbf{OM} \) that represents the complex quantity \( \Lambda(\omega) \), then, as \( \omega \) varies, the point \( M \) traces a curve in the complex plane. This locus characterizes the linear part \( L(s) \) of the system; we shall call
it the Tsypkin locus\textsuperscript{1} of the system. The conditions for periodic solution can now be interpreted as the intersections of the Tsypkin locus with the $-\frac{h}{2}$ horizontal line (or, in the absence of hysteresis, with the real axis), provided these intersections lie in the left half plane (Fig. 26-12).

26.2.5. Determining the Tsypkin Locus from the Transfer Locus $L(j\omega)$. The real and imaginary parts of $\Lambda(\omega)$ are:

\[
\text{Re} \Lambda(\omega) = \frac{4M}{\pi} \sum A(n\omega) \cos \Phi(n\omega)
\]

\[
\text{Im} \Lambda(\omega) = \frac{4M}{\pi} \sum \frac{1}{n} A(n\omega) \sin \Phi(n\omega)
\]

If the transfer function $L(j\omega)$ is written as the sum of its real and imaginary parts

\[
L(j\omega) = U(\omega) + jV(\omega)
\]

with

\[
U(\omega) = A(\omega) \cos \Phi(\omega), \\
V(\omega) = A(\omega) \sin \Phi(\omega)
\]

it is immediately seen that

\[
\text{Re} \Lambda(\omega) = \frac{4M}{\pi} \sum U(n\omega)
\]

\[
\text{Im} \Lambda(\omega) = \frac{4M}{\pi}
\]

or, in explicit form,

\[
\text{Re} \Lambda(\omega) = \frac{4M}{\pi} \left[ U(\omega) + U(3\omega) + U(5\omega) + \cdots \right]
\]

\[
\text{Im} \Lambda(\omega) = \frac{4M}{\pi} \left[ \frac{1}{2} V(\omega) + \frac{1}{2} V(3\omega) + \cdots \right]
\]

These relations enable one to draw the Tsypkin locus from the $U + jV$ transfer locus. This can be done either from the Nyquist plot of $L = U + jV$ (Fig. 26-13), or from the separate plots of $U$ vs. $\omega$ and $V$ vs. $\omega$. The latter procedure is facilitated if the plots of $U$ and $V$ are drawn by using a logarithmic scale for the frequency (Fig. 26-14).

Since most transfer loci $L(j\omega)$ tend toward the origin at high frequencies, it is, in general, sufficient to consider the third and fifth harmonics. Furthermore it is seen that the high-frequency portion of the Tsypkin locus is extremely close to the corresponding portion of the transfer locus, since even the third harmonic $V(3\omega)/3$ quickly becomes negligible as $\omega$ increases.

\textsuperscript{1} Russian authors (Tsypkin, Ajserman, etc.) call it the hodograph of the function $\Lambda(\omega)$.
Fig. 26-13. Obtaining the Tsyplkin locus from the transfer locus \( L(j\omega) \): \( L \) being any point \( (\omega_0) \) on the transfer locus, mark the points \( L_1 (\omega = 3\omega_0) \), \( L_5 (\omega = 5\omega_0) \), etc. Then obtain \( L'_1 \) from \( L_1 \) by dividing the ordinate by 3; \( L'_5 \) from \( L_5 \) by dividing the ordinate by 5, etc. The \( \omega_0 \) point on the Tsyplkin locus is given by \( OT = OL + OL'_1 + OL'_5 + \cdots \).

Fig. 26-14. Obtaining the Tsyplkin locus from separate plots of \( U(\omega) \) and \( V(\omega) \) vs. \( \log \omega \). \( \text{Re} \lambda(\omega) \) and \( \text{Im} \lambda(\omega) \) are obtained by graphical addition of the curves shown. (After J. Tsyplkin, "Teoriya relejnykh sistem avtomaticheskovo regulirovaniya," p. 175, Gostekhizdat, Moscow, 1956.)

Fig. 26-15. Comparison of the Tsyplkin locus \( T \) and the Nyquist locus \( L \). (After J. Tsyplkin, "Teoriya relejnykh sistem avtomaticheskovo regulirovaniya," p. 356, Gostekhizdat, Moscow, 1956.)
This is the basis of the first-harmonic approximation outlined in Chap. 24. A comparison of the Tsypkin locus $\Lambda(\omega)$ and the transfer locus $L(j\omega)$ shows that this approximation usually leads to correct results for high-frequency oscillations, but that it may lead to incorrect conclusions concerning low-frequency oscillations. An example quoted by Tsypkin\textsuperscript{1} concerns the function

$$L(s) = K/(1 - T_1s)(1 + T_2s)$$

The shape of the Tsypkin locus (Fig. 26-15) shows that one possible periodic solution exists whenever \(h/2\) is smaller than \(KM(T_1 - T_2)/(T_1 + T_2)\), whereas the describing-function technique would indicate two oscillations for small values of \(h\) and none for higher values.

Examples of Tsypkin loci are plotted in Figs. 26-16 to 26-21 (drawn by H. Lepage) for the functions $L(s)$ already represented by Figs. 26-6 to 26-11 in the form of Hamel loci. In most of these figures the high-frequency portions of the loci have been represented separately on a larger scale.\textsuperscript{2}

26.2.6. Comparison of Hamel and Tsypkin Loci. It is easily seen that the Hamel and Tsypkin representations are equivalent. In fact, letting $T = 2\pi/\omega$, we have

$$\text{Re} \, \Lambda(\omega) = \frac{1}{\omega} \frac{de}{dt} \left( \frac{T}{2} \right)$$
$$\text{Im} \, \Lambda(\omega) = \varepsilon \left( \frac{T}{2} \right)$$


\textsuperscript{2}It is to be noted that actual plotting of the Tsypkin locus is not necessary, in practice, to determine the possible self-oscillations of the system. It is sufficient to draw the $\text{Im} \, \Lambda(\omega)$-vs.-$\omega$ curve, obtain its intersections with the $-h/2$ line and check the sign of $\text{Re} \, \Lambda(\omega)$ at the frequencies of the intersection points.
In other words, the Tsypkin locus is obtained from the Hamel locus by interchanging the coordinates and dividing $d\varepsilon/dt$ by $\omega = \pi/T$. The Hamel representation has the two following advantages:

a. It makes use of the phase-plane variables $\varepsilon$ and $d\varepsilon/dt$, which is an aid to visualizing the system's behavior.

b. The introduction of a derivative control in the form of a $1 + \lambda s$ controller is especially easy to study.

On the other hand, the Tsypkin representation has the advantage that its high-frequency region is generally very close to the transfer locus $L(j\omega)$. This makes the Tsypkin locus more familiar to servo engineers trained in linear techniques. Furthermore, it is especially suitable for judging the validity of the results yielded by the approximate method outlined in Chap. 24.

26.2.7. Extension When a Dead Zone Is Present. The methods just outlined can be extended to the case in which the on-off element has a dead zone $\Delta$. This has been done independently by Hamel and Tsypkin, using somewhat different techniques.
The Tsypkin approach will be briefly outlined. It consists in noting that, when the system operates under periodic conditions, the output \( w_T(t) \) of the nonlinear element has the shape shown in Fig. 26-22, because of the presence of the dead zone. The problem of finding a periodic regime for the system consists in finding the function

\[
\varepsilon = -\frac{\tau d}{2} - \frac{KM}{\tau} t_1
\]

\[
\frac{d\varepsilon}{dt} = KM(\Delta - KM_t) \frac{1 - e^{-\nu r}}{(1 - e^{-\nu r}) + KM_t(1 + e^{-\nu r})}
\]

where \( r, K, M, \) and \( \Delta \) are the quantities defined in Sec. 26.2.1.
Fig. 26-19. Tsypkin loci for $L(s) = 1/s(s^3 + 2\xi s + 1)$, $\xi < 1$. 
Fig. 26-21. Tsypkin loci for $L(s) = \frac{1 + \alpha s}{s^4(1 + s)}$.

Fig. 26-22.
$w_T(t)$, that is, the two parameters $\omega = 2\pi/T$, which characterizes its frequency, and $\rho$, which characterizes its shape.

The Fourier expansion of $w_T(t)$ is (Sec. 4.5.1)

$$w_T(t) = \frac{4M}{\pi} \sum_{o} \frac{1}{n} \sin n\rho \frac{\pi}{2} \cos \left( n\omega t - n\rho \frac{\pi}{2} \right)$$

where the summation is extended to include the positive odd values of $n$. Under these conditions the output $r(t) = -\varepsilon(t)$ and its time derivative are

$$r(t) = \frac{4M}{\pi} \sum_{o} \frac{1}{n} A(n\omega) \sin n\rho \frac{\pi}{2} \cos \left[ n\omega t - n\rho \frac{\pi}{2} + \Phi(n\omega) \right] \quad (26-19)$$

$$\frac{1}{\omega} \frac{dr}{dt} = -\frac{4M}{\pi} \sum_{o} A(n\omega) \sin n\rho \frac{\pi}{2} \sin \left[ n\omega t - n\rho \frac{\pi}{2} + \Phi(n\omega) \right] \quad (26-20)$$

where

$$A(\omega)e^{i\Phi(\omega)} = L(j\omega)$$

The conditions for periodicity are twofold:

a. For the commutation $P_1$ (from zero to minus) these equations correspond [see Eq. (26-3a)] to those intersections of the $-(\Delta + \lambda)/2$ line with the $\Lambda(\omega)$ locus that lie in the left half plane, $\Lambda(\omega)$ being defined as

$$\Lambda(\omega) = \frac{1}{\omega} \frac{dz}{dt} \left( \frac{\pi}{\omega} \right) + jz \left( \frac{\pi}{\omega} \right) \quad (26-21)$$

The function $\Lambda(\omega)$ depends on $\rho$; that is, there is an infinity of Tsykin loci $\Lambda(\omega)$, one for each value of $\rho$. Intersection of each locus with the $-(\Delta + \lambda)/2$ line indicates a possible periodic solution and gives the corresponding $\omega$ and $\rho$ (Fig. 26-23).

b. For the commutation $P'_1$ (from + to zero), one is similarly led [Eq. (26-4b)] to consider a second family of Tsykin loci

$$\Lambda'(\omega) = \frac{1}{\omega} \frac{dz}{dt} \left( \rho \frac{\pi}{\omega} \right) + jz \left( \rho \frac{\pi}{\omega} \right) \quad (26-22)$$
and their intersection with the \((\Delta - h)/2\) line. There is one \(\Lambda'(\omega)\) locus for each value of \(\rho\). Each intersection indicates a possible periodic solution with the corresponding \(\omega\) and \(\rho\) (Fig. 26-23).

Thus, a periodic solution must comply with the conditions expressed by the intersections of the \(\Lambda(\omega)\) and \(\Lambda'(\omega)\) loci with the commutation lines. The values of \(\omega\) and \(\rho\) can thus be found by plotting \(\omega\) vs. \(\rho\) for the two sets of intersections and choosing the values that are common to them.

26.3.8. Plotting the Loci. Expressions for the Tsypkin loci when a dead zone is present are obtained by substituting \(t = \pi/\omega\) and \(t = \rho\pi/\omega\) in Eqs. (26-19) and (26-20). One thus obtains

![Diagram](image.png)

**Fig. 26-24a.** Families of Tsypkin loci for \(L(s) = 1/s(1 + s)\), \(M = 1\), when dead zone is present. Parameter \(\rho\) is defined by Fig. 26-22.
\[ \text{Im } \Lambda(\omega) = \frac{2M}{\pi} \sum \left[ \frac{\sin \frac{n\omega}{n}}{n} \left( \frac{1 - \cos n\sigma \pi}{n} V(n\omega) \right) \right] \]

\[ \text{Re } \Lambda(\omega) = \frac{2M}{\pi} \sum \left[ \frac{(1 - \cos n\sigma \pi)U(n\omega) - \sin n\sigma \pi V(n\omega)}{n} \right] \]

\[ \text{Re } \Lambda'(\omega) = \frac{2M}{\pi} \sum \left[ (1 - \cos n\sigma \pi)U(n\omega) + \sin n\sigma \pi V(n\omega) \right] \]

\[ \text{Im } \Lambda'(\omega) = \frac{2M}{\pi} \sum \left[ \frac{\sin n\sigma \pi}{n} U(n\omega) - \frac{1 - \cos n\sigma \pi}{n} V(n\omega) \right] \]

**Fig. 26-24b.** Tsypkin loci for \( L(s) = 1/(s(1 + s)(1 + 0.1s)), \ M = 1 \) with dead zone.
These expressions make it possible for each value of \( \rho \) to plot \( \Lambda(\omega) \) and \( \Lambda'(\omega) \) from the Nyquist locus \( L(j\omega) \) or from the separate plots of \( U(\omega) \) and \( V(\omega) \) by using techniques similar to those indicated in Fig. 26-13 or 26-14.

In the Hamel form, algebraic expressions generalizing Eqs. (26-9) to (26-14) can also be given. They are, when no integration is present,

\[
\varepsilon \left( \frac{T}{2} \right) = -M \sum_{i}^{n} \frac{A_i \exp \left( p_i T/2 \right) - \exp \left[ p_i (1 - \rho) T/2 \right]}{\exp (p_i T/2) + 1}
\]

\[
\frac{d\varepsilon}{dt} \left( \frac{T}{2} \right) = Ml_0 - M \sum_{i}^{n} A_i \frac{\exp (p_i T/2) - \exp \left[ p_i (1 - \rho) T/2 \right]}{\exp (p_i T/2) + 1}
\]

---

Fig. 26-24c. Twykin loci for \( L(s) = 1/(s(1 + s)(1 + 0.2s)) \), \( M = 1 \) with dead zone.
oscillations of on-off control systems

\[ \varepsilon \left( \frac{T}{2} \right) = -M \sum_{i=1}^{n} \frac{A_i \exp \left( \rho_i T/2 \right) - 1}{\rho_i \exp \left( \rho_i T/2 \right) + 1} \]

\[ \frac{d\varepsilon}{dt} \left( \frac{T}{2} \right) = -M \sum_{i=1}^{n} \frac{A_i \exp \left( \rho_i T/2 \right) - 1}{\exp \left( \rho_i T/2 \right) + 1} \]

and, when one integration is present,

\[ \varepsilon \left( \frac{T}{2} \right) = -M \rho \frac{T}{4} P(0) - M \sum_{i=1}^{n-1} \frac{A_i \exp \left( \rho_i T/2 \right) - \exp \left[ \rho_i (1 - \rho) T/2 \right]}{\rho_i \exp \left( \rho_i T/2 \right) + 1} \]

Fig. 28-24d. Tsypkin loci for \( L(s) = 1/s(1 + s)(1 + 0.5s) \), \( M = 1 \) with dead zone.
Fig. 26-24e. Tsypkin loci for \( L(s) = 1/s(1 + s)(1 + 0.9s) \), \( M = 1 \) with dead zone.

\[
\frac{de}{dt} \left( \frac{T}{2} \right) = Ml_0 - M \sum_{i=1}^{n-1} A_i \frac{\exp(p_i T/2) - \exp[p_i(1 - \rho)T/2]}{\exp(p_i T/2) + 1}
\]

\[
e \left( \frac{T}{2} \right) = -M \rho \frac{T}{4} \frac{P(0)}{Q_1(0)} + -M \sum_{i=1}^{n-1} \frac{A_i \exp(p_i T/2) - 1}{p_i \exp(p_i T/2) + 1}
\]

\[
\frac{de}{dt} \left( \frac{T}{2} \right) = -M \sum_{i=1}^{n-1} A_i \frac{\exp(p_i T/2) - 1}{\exp(p_i T/2) + 1}
\]

In these expressions, \( A_i \) and \( Q_1(0) \) are the quantities defined in Sec. 26.2.2. The sets of loci corresponding to the function \( L(s) = 1/s(1 + s)(1 + 0.9s) \) are plotted in Tsypkin form for different values of \( \alpha \) in Figs. 26-24a to e.\(^1\) It is to be noted that, for

\(^1\) Drawn by H. Lepage.
\( \rho = 1, \Lambda(\omega) \text{ and } \Lambda'(\omega) \) are the Tsypkin loci of Fig. 26-8. Loci for \( L(s) = [\exp(-s^2)]/s \) can be found in Tsypkin's work, page 229.

26.3. FORCED OSCILLATIONS IN ON-OFF SERVO SYSTEMS

26.3.1. General. A nonlinear system subjected to a periodic input is said to exhibit forced oscillations when it produces periodic phenomena at the frequency of the input (forcing frequency \( \omega_f \)), and not at the frequency of the self-oscillations of the system. This is a synchronization phenomenon. An example of such phenomena has been encountered in Sec. 22.1.3 for systems of the Van der Pol type. Its occurrence in on-off servo systems will now be studied.

The forcing input may be applied anywhere in the loop. Therefore (Sec. 13.1.7, Fig. 13-11) it is possible to assume that it is applied at the same point as the control input. Thus, the study of forced oscillations becomes the problem of studying under what conditions an on-off servo system will exhibit periodic phenomena at the frequency \( \omega_f \) when it is subjected to a command \( e_f \) periodic with the angular frequency \( \omega_f \).

26.3.2. Case in Which No Dead Zone Is Present. In the \((e, de/dt)\) plane the periodic forcing function \( e_f(t) \) is represented by a closed curve \( E_T \) which is traced out in one period \( T_f = 2\pi/\omega_f \). In the particular case

of a harmonic forcing function \( e_f(t) = e_0 \sin(\omega_f t + \phi) \) this closed curve is an ellipse (Fig. 26-25). If \( e_f(t) \) is a more complicated function of time, the \( E_T \) locus may be more complicated (Fig. 26-26). It will be assumed that the forcing function has zero mean value, the \( E_T \) locus being thus centered at the origin. The maximum value \( e_0 \) of \( e_f(t) \) will be called the magnitude of the forcing function.

Suppose the system operates under periodic conditions at the forcing
frequency and that a commutation from 0 to + is taken as the origin of time \(P_0\). At half-period \(t = T_f/2\) the position of the point \(P_1(\varepsilon, d\varepsilon/dt)\) in the phase plane will be given by

\[
\varepsilon \left( \frac{\pi}{\omega_f} \right) = e \left( \frac{\pi}{\omega_f} \right) - r \left( \frac{\pi}{\omega_f} \right), \quad \frac{d\varepsilon}{dt} \left( \frac{\pi}{\omega_f} \right) = \frac{de}{dt} \left( \frac{\pi}{\omega_f} \right) - \frac{dr}{dt} \left( \frac{\pi}{\omega_f} \right)
\]

The quantities \(e(\pi/\omega_f)\) and \((de/dt)(\pi/\omega_f)\) represent the vector \(E_T\) for \(t = \pi/\omega_f\). The quantities \(-r(\pi/\omega_f)\) and \(-(dr/dt)(\pi/\omega_f)\) represent the vector \(O_H\) defined by Eqs. (26-6) and (26-7), \(H_f\) being the point of the Hamel locus that corresponds to the forcing frequency. Thus the point \(P_1\) is located on the \(\varepsilon_T\) locus, which is identical with the \(E_T\) locus but is centered at \(H_f\) (Fig. 26-27). Intersection of this \(\varepsilon_T\) locus with the \(\varepsilon = -h/2\) line corresponds to a possible periodic solution, provided that this intersection lies in the lower half plane. If there is no such intersection, there can be no forced oscillation.

**26.3.3. Examples. Conclusion.** The following configurations are most commonly encountered:

**Case 1.** There are two intersections in the lower half plane (Fig. 26-28). This means that forced oscillations are possible, each intersection corresponding to a possible "phase" with respect to the periodic forcing function \(\varepsilon_f(t)\).

**Case 2.** The curve \(\varepsilon_T\) does not intersect the \(\varepsilon = -h/2\) line (Fig. 26-29). This means that the system cannot exhibit forced oscillations with respect to the forcing input under consideration.

**Case 3.** The curve \(\varepsilon_T\) is tangent to the \(-h/2\) line (Fig. 26-30). This is the critical case lying between Cases 1 and 2. The amplitude of the forcing input is just sufficient to give rise to forced oscillations in the system. If the amplitude were smaller, the curve \(\varepsilon_T\) would be smaller, while still centered at point \(H_f\), and would not intersect the \(-h/2\) line. If, on the other hand, the amplitude were greater, there would be two intersections. The corresponding amplitude for the forcing function is the critical amplitude \(\varepsilon_c\) at the angular frequency \(\omega_f\). It is easy to see that the critical amplitude is smaller if the point \(H_f\) lies near the intersection of the Hamel locus with
the \(-h/2\) line, i.e., if the forcing frequency \(\omega_f\) is close to the frequency \(\omega_0\) at which the system exhibits self-oscillations (in the limit, \(\varepsilon\) is zero when \(\omega_f = \omega_0\), which means that the system can oscillate at the frequency \(\omega_0\) in the absence of any external stimulus). In other words, the synchronization of the system at the frequency \(\omega_f\)

\[ \begin{align*}
  \frac{d\varepsilon}{dt} & \quad \text{at } \omega = \omega_f. \\
  \end{align*} \]

The considerations just outlined enable one to draw conclusions concerning the influence of parameters on the existence of forced oscillations. They also enable one to determine the forced oscillations quantitatively, when they exist. They can, of course, be expressed by using Tsypkin loci instead of Hamel loci. They can also be extended to the case in which a dead zone is present.

26.4. STABILITY OF PERIODIC STATES

26.4.1. General. Consider an on-off servo system, with or without a dead zone. Suppose conditions have been found (i.e., values of \(\omega\) and \(\rho\) under which periodic phenomena are possible, either self- or forced oscillations. The question that now arises is whether such a periodic regime is stable, in other words, whether the system, after a disturbance, will resume its periodic regime. This is obviously the key to whether the corresponding periodic regime will physically exist.
26.4.2. Expressing the Deviation from Periodic Conditions.¹ Consider the case in which no dead zone is present. Let \( e_r(t) \) be the error under periodic conditions with a period \( T = 2\pi/\omega_1 \). Suppose the system is operating "in the neighborhood" of this periodic condition, i.e., that the actual error is

\[
e(t) = e_r(t) + e_d(t)
\]

where \( e_d(t) \) is the deviation of the actual error from the error under periodic conditions. The function \( e_d(t) \) is assumed to have small average amplitude as compared to \( e_r(t) \). Furthermore, it will be assumed that the time derivative of \( e_d(t) \) does not take too large values. This assumption is a reasonable one when the power stage, and/or the system to be controlled, acts as a low-pass filter that filters out high-frequency components, i.e., rapid variations of \( r(t) \) or \( e(t) \).

Under such conditions the response \( w(t) \) of the on-off element for the input \( e_r + e_d \) will consist (Fig. 26-32) of a sequence of square waves.

¹ The method outlined in this section is adapted from J. Tzypkin, "Teorija relejnykh sistem avtomaticheskovo regulirovaniia," chaps. 8 and 9, Gostekhizdat, Moscow, 1955.
This sequence, however, is not periodic, since commutations occur when \( t \) is a root of the equation

\[
\varepsilon_r(t) + \varepsilon_d(t) = 0 \quad \text{that is,} \quad t = t_1, t_2, \ldots
\]

and not when

\[
\varepsilon_r(t) = 0 \quad \text{that is,} \quad t = \frac{\pi}{\omega_1}, \frac{2\pi}{\omega_1}, \ldots
\]

In other words, \( w(t) \) will differ from the response \( w_r(t) \) under periodic conditions by a sequence of alternatively positive and negative pulses with a magnitude \( 2M \). These pulses occur at the instants \( t = \pi/\omega_1, t = 2\pi/\omega_1, \) etc. Their durations are the quantities \((t_1 - \pi/\omega_1), (t_2 - 2\pi/\omega_1), \) etc.

Because of the assumption that \( \varepsilon_d(t) \) is small and does not vary too rapidly, the pulse occurring at the instant \( n\pi/\omega_1 \)

\[
2M \left[ u \left( t - n \frac{\pi}{\omega_1} \right) - u(t - t_n) \right]
\]

can be represented by the impulse

\[
(-1)^n \frac{d\varepsilon_r}{dt} \left( \frac{n\pi}{\omega_1} \right) \delta \left( t - n \frac{\pi}{\omega_1} \right)
\]

where \( \delta(t) \) is the Dirac or unit-impulse function (Sec. 4.3.2). When \( \varepsilon_d(t) \) becomes very small \((d\varepsilon/dt \ll d\varepsilon_r/dt, \) see Fig. 26-33), this is equivalent to

\[
2M \left[ \frac{d\varepsilon_r}{dt} \left( \frac{n\pi}{\omega_1} \right) \right] \delta \left( t - n \frac{\pi}{\omega_1} \right)
\]

Therefore, the deviation \( w_d(t) \) of the response of the on-off element from its expression \( w_r(t) \) under periodic conditions can be expressed as

\[
w_d(t) = 2M \sum_{n=0}^{\infty} \left[ \frac{\varepsilon_r \left( \frac{n\pi}{\omega_1} \right)}{\frac{d\varepsilon_r}{dt} \left( \frac{n\pi}{\omega_1} \right)} \delta \left( t - n \frac{\pi}{\omega_1} \right) \right]
\]

Taking into account the periodicity of \( \varepsilon_r(t) \)

\[
\left| \frac{d\varepsilon_r}{dt} \left( n \frac{\pi}{\omega_1} \right) \right| = \cdots = \left| \frac{d\varepsilon_r}{dt} \left( \frac{\pi}{\omega_1} \right) \right|
\]

\(^1\) More exactly, their beginning or their end takes place at the instants \( n\pi/\omega_1 \), according to the sign of \((-1)^n\varepsilon_d(n\pi/\omega_1)\).
(the expression will be noted $\varepsilon'$ for brevity), the expression for $w_d(t)$ can be written

$$w_d(t) = \frac{2M}{\varepsilon'} \sum_{0}^{\infty} \varepsilon_d \left( \frac{n\pi}{\omega_1} \right) \delta \left( t - \frac{n\pi}{\omega_1} \right)$$

Thus, $w_d(t)$ appears as the modulation of the function of time $(2M/\varepsilon')\varepsilon_d(t)$ by pulses occurring at regular intervals $t = \pi/\omega_1, 2\pi/\omega_1, \ldots, n\pi/\omega_1, \ldots$

26.4.3. Studying the Deviation $\varepsilon_d(t)$. 1. As a consequence, studying the stability of the periodic regime defined by $\varepsilon(t) = \varepsilon_f(t)$ is equivalent to studying the behavior of the function $\varepsilon_d(t)$ when $w_d(t) = w(t) - w_f(t)$ is fed into the linear part $L(s)$ of the system and when the loop is closed. This is a problem of stability of sampled-data servo systems.

More precisely, the stability of a limit cycle with frequency $\omega_1$ is the same as the stability of the sampled unity-feedback servo system with an open-loop transfer function $KG(s) = (2M/\varepsilon')L(s)$, the sampling frequency being $\Omega = 2\omega_1$. In what follows, the frequency $\omega_1$ will be called $\omega_f$ in the case of a forced oscillation and $\omega_1$ in the case of a self-oscillation.

As seen in Sec. 20.3.5, stability can be investigated by applying the generalized Nyquist criterion to the sampled transfer locus

$$KG^*(j\omega) = \left(\frac{2M}{\varepsilon'}\right) L^*(j\omega)$$

If the system is regular and open-loop stable (which we will assume), stability will occur if the intersection of the $KG^*(j\omega)$ locus with the real axis for $\omega = \Omega/2 = \omega_1$ takes place at the right of the critical point $-1$. Instability will occur if the $\omega_1$ point lies at the left of the $-1$ point:

$$KG^*(j\omega_1) > -1 \quad \text{stable}$$
$$KG^*(j\omega_1) < -1 \quad \text{unstable}$$

2. Stability of Forced Oscillations. Let $\omega_f$ be the forcing frequency. For $\omega = \Omega/2 = \omega_f$ the point at which the $KG^*(j\omega)$ locus intersects the real axis should, for stability, lie at the right of the $-1$ point, thus

$$\frac{2M}{\varepsilon'} L^*(j\omega_f) > -1$$

Multiplying both sides successively by $\varepsilon'/2M$ and by the sampling period $\pi/\omega_f$ yields, equivalently,

$$\frac{\pi}{\omega_f} L^*(j\omega_f) > -\frac{\pi}{\omega_f} \frac{\varepsilon'}{2M}$$

The left-hand side of this expression can be shown to be equal to

$$U^*(\omega_f) = \frac{\pi}{2M} \text{Re} \Lambda(\omega_f)$$
where \( \Lambda(\omega) \) is the expression for the Tsykin locus [Eq. (26-17)]. The right-hand side is the product of \( \pi/2M \) and the abscissa of the operative point at which the \( \varepsilon_T \) locus centered at \( \Lambda(\omega_T) \) intersects the \( -h/2 \) line (Fig. 26-34).

Hence the condition for stability: the operative point should lie at the left of \( \Lambda(\omega_T) \). In the typical case shown in Fig. 26-34, where two forced oscillations are theoretically possible with different phases with respect to \( \varepsilon_T(t) \), it is thus seen that only that oscillation corresponding to point \( P_1 \) will physically occur. That corresponding to point \( P_2 \) is unstable.

Similar reasoning can be applied when a dead zone is present.

3. Stability of Self-oscillations. In the case of self-oscillations, \( \omega_T = \omega_0 \) and the sampled open-loop transfer locus \( (2M/\varepsilon')L^*(j\omega) \) passes through the critical point at \( \omega = \omega_0 \). This means that the application of the generalized Nyquist criterion is less straightforward.

The difficulty arises from the fact that the function \( 1 + (2M/\varepsilon')L^*(j\omega) \) is zero for \( s = j\omega_0 \). Its stability depends on whether the other zeros of this function have negative real parts or not. This can be investigated by studying the stability of the zeros of the function after the \( s = j\omega_0 \) zero has been removed or, equivalently, by studying the position of the locus

\[
S(j\omega) = \frac{1 + (2M/\varepsilon')L^*(j\omega)}{s - j\omega_0}
\]

with respect to the origin in the neighborhood of \( \omega = \omega_0 \). If \( S(j\omega_0) \) is positive, it follows from the generalized Nyquist criterion that the system is stable. If \( S(j\omega_0) \) is negative, the system is unstable.

For \( \omega = \omega_0 \) both the numerator and the denominator of \( S(j\omega) \) are zero. The limit of \( S(j\omega) \) for \( \omega = \omega_0 \) is the derivative of the numerator with respect to \( s = j\omega \) for \( \omega = \omega_0 \). This quantity has the same sign as \( dL^*(j\omega)/ds \). Now

\[
\frac{dL^*}{ds} = \frac{1}{j} \frac{dL^*}{d\omega} = \frac{1}{j} \frac{dU^*}{d\omega} + \frac{dV^*}{d\omega}
\]

with

\[
\frac{dU^*}{d\omega} = \sum_{n=0}^{\infty} \frac{dU(\omega + 2n\omega_0)}{d\omega} \quad \frac{dV^*}{d\omega} = \sum_{n=0}^{\infty} \frac{dV(\omega + 2n\omega_0)}{d\omega}
\]

Noting that, for \( \omega = \omega_0 \)

\[
\frac{dU(\omega + 2n\omega_0)}{d\omega} = \frac{1}{2n + 1} \frac{dU[(2n + 1)\omega]}{d\omega} \quad \frac{dV(\omega + 2n\omega_0)}{d\omega} = \frac{1}{2n + 1} \frac{dV[(2n + 1)\omega]}{d\omega}
\]
it is seen that
\[
\frac{dU^*}{d\omega_0} = 0 \quad \frac{dV^*}{d\omega_0} = 2 \sum_{0}^{+\infty} \frac{1}{2n+1} \frac{dV[(2n+1)\omega]}{d\omega_0}
\]

The first of these equations shows that, for the particular case of \( \Omega = 2\omega_0 \), the sampled locus arrives at \( \omega = \omega_0 \) perpendicularly to the real axis. Thus the stability is determined by \( dV^*/d\omega_0 \); that is, stability occurs if the critical point is approached from the lower half plane as \( \omega \) increases through \( \omega_0 \), and instability occurs if it is approached from the upper half plane (Fig. 26-35).

![Fig. 26-35.](image)

\( (a) \frac{dV^*}{d\omega_0} > 0: \) stable

\( (b) \frac{dV^*}{d\omega_0} < 0: \) unstable

It is possible to extend the analysis further by evaluating \( dV^*/d\omega_0 \). Comparison with the expression for the Tsypkin locus (Eq. (26-18)) shows that

\[
\frac{dV^*}{d\omega_0} = \frac{\pi}{2M} \frac{d \text{Im} \Lambda(\omega_0)}{d\omega_0}
\]

Hence stability depends on the sign of \( d \text{Im} \Lambda(\omega_0)/d\omega_0 \), which is easily read from the plot of the Tsypkin locus. The sign is positive if the intersection with the \(-h/2\) line occurs from below for increasing \( \omega \); it is negative if the intersection occurs from above (Fig. 26-36). Equivalently, using the Hamel representation, intersections occurring from left to right for increasing \( \omega \) correspond to positive \( dV^*/d\omega_0 \), and to stable self-oscillations; intersections occurring from right to left correspond to unstable self-oscillations (Fig. 26-37).

![Fig. 26-36.](image)

![Fig. 26-37.](image)
The conditions just stated are necessary and sufficient for the case in which the \((2M/\varepsilon')L^*(j\omega)\) locus does not intersect the real negative axis beyond the \((-1)\) point for a value \(\omega \neq \omega_0\), as in Figs. 20-27 and 26-35. Conversely, for the case in which at least one such supplementary intersection exists (Fig. 26-38), it can be shown\(^1\) that the conditions stated are just necessary conditions. In other words, intersections from above (or, in the Hamel representation, from the right) always correspond to unstable self-oscillations. Intersections from below (or from the left) may be stable or unstable.

![Fig. 26-38. Case in which conditions for stability are just necessary: (a) is unstable; (b) may be stable or unstable.](image)

These results should be compared with those given by the method of the first-harmonic approximation. When \(\omega_0\) is high, the Tsypkin locus does not differ greatly from the open-loop transfer locus, and the results of the discussion (see Secs. 24.3.3 and 24.3.7) made on the basis of the left-hand criterion applied to the transfer locus will hold. For self-oscillations occurring at relatively low frequencies, however, the Tsypkin locus may differ appreciably from the transfer locus, and the first-harmonic approximation may lead to incorrect conclusions.

\(^1\) For more details concerning this particular question, see G. Nejmark, "O periodicheskih reshimakh i ustoichivosti releynykh sistem," *Avtomatika i Telemekhanika*, 14(5): 556-569 (1953).
CHAPTER 27

ADDITIONAL METHODS APPLICABLE TO NONLINEAR SYSTEMS

Summary
1. Poincaré’s theorems.
2. Limit cycles and stability.

It is the purpose of this chapter to outline briefly a few methods applicable to nonlinear systems. These methods are more or less well known in the general field of nonlinear mechanics. Their application to specific automatic-control problems can broaden the engineer’s outlook and enable him not only to solve specific problems but also to understand the fundamental nature of certain classes of nonlinear problems. It is from this viewpoint that such mathematical methods are briefly presented in the following sections. It is hoped that the introductory comments that follow will help the reader to gain a better understanding of the theoretical background underlying the properties of nonlinear servo systems studied in Chaps. 23 to 26 and will encourage him to make a more thorough investigation of the subject.

27.1. POINCARÉ’S THEOREMS

27.1.1. General. Singular Points. The use of the phase plane is actually based on a number of theorems by H. Poincaré that apply to systems governed by simultaneous differential equations of the form

\[
\frac{dx}{dt} = P(x,y,z, \ldots) \quad \frac{dy}{dt} = Q(x,y,z, \ldots) \quad \frac{dz}{dt} = R(x,y,z, \ldots)
\]

(27-1)

where \(x, y, z, \ldots\) are the variables that characterize the system and \(P, Q, R, \ldots\) are functions of these variables.

We have already considered the particular case in which the system is characterized by one variable \(x\), with \(y = \frac{dx}{dt}\). In this case the system is completely defined by the trajectories of the \((x,y)\) point in the phase plane (Sec. 25.1.1) and the theorems stated can be easily visualized.

Consider a system defined by

\[
\frac{dx}{dt} = P(x,y) \quad \frac{dy}{dt} = Q(x,y)
\]

(27-2)

where \(y = \frac{dx}{dt}\), that is, \(P(x,y) = y\). It follows from the fundamental
theorem of Cauchy on the uniqueness of the solution of simultaneous
differential equations with given initial conditions that there is one, and
only one, trajectory passing through a point in the phase plane. This
property holds true if the system satisfies classic conditions known as the
Cauchy-Lipschitz conditions, which in particular presuppose the continuity of the $P$ and $Q$ functions. Exceptions are the points where simultaneously

$$ P(x,y) = 0 \quad Q(x,y) = 0 $$

These points are called singular, or critical, points.

Fig. 27-1. Various kinds of singular points: (a) center, (b) saddle, (c) node, (d) star, (e) focus.

If the system under consideration is linear, there is only one such point; in the general case, there are several. At every singular point, $x$ and $dx/dt$ are zero; the system is in a state of equilibrium. To examine the point, the origin is brought to it by coordinate translation.

27.1.2. Various Kinds of Singular Points. In the linear case, the following types of singular points exist; they are defined by the behavior of the trajectories in their neighborhood.

1. Center (Fig. 27-1a). The trajectories are concentric circles or ellipses. None pass through the origin, which is a position of stable equilibrium. An example is $P = y, Q = -x$, whence $x \, dx + y \, dy = 0$; this is the case of an undamped second-order system.

2. Saddle (Fig. 27-1b). Several trajectories pass through the origin, which is a point of unstable equilibrium. An example is $P = y, Q = x$, whence $x \, dy - y \, dx = 0$. 
3. **Node** (Fig. 27-1c). Several trajectories pass through the origin, where they all have the same slope. Example: \( P = y, \ Q = x - 3y \). This corresponds to an overdamped motion (e.g., Fig. 25-2 with \( \zeta > 1 \)).

4. **Star** (Fig. 27-1d). As for the node, several trajectories pass through the origin. Example: \( P = \lambda x, \ Q = \lambda y \), with \( \lambda \) real.

5. **Focus** (Fig. 27-1e). This corresponds to the existence of damped or divergent oscillations (stable or unstable focus). An example is \( (x + y) \ dx + (y - x) \ dy = 0 \); whence in polar coordinates, \( \rho = \exp(K\theta) \). The stability is controlled by the sign of \( K \). (See Fig. 25-2 with \( \zeta < 1 \).)

The familiar gravity pendulum offers an excellent way of visualizing the nature of singular points. The equilibrium position of the pendulum is stable: in the usual case in which the pendulum is just slightly damped, it is a stable focus; it is a center in the limiting case in which the damping can be considered negligible; if heavy damping is introduced (e.g., pendulum in oil, eddy currents, etc.), the oscillations disappear and the equilibrium position becomes a stable node. The case of an unstable focus \( (-1 < \zeta < 0) \) or node \( (\zeta < -1) \) would be that of a pendulum negatively damped by some external means (or, equivalently, that of an ordinary pendulum if one mentally reverses the direction of time). A saddle point corresponds to the unstable equilibrium position of the pendulum: the center of gravity is above the pivot; that is, gravity acts as an antirestoring force \( (k < 0) \). Reaching the saddle point is possible only by starting from very particular initial conditions (launching the pendulum in the correct direction with the exact required momentum to stop at the upper equilibrium position; see on Fig. 27-1b the particular trajectory that goes through the saddle point).

In the case of nonlinear systems, the same types of singular points occur. New phenomena arise only if the \( P(x,y) = 0 \) and \( Q(x,y) = 0 \) curves happen to be tangent.

**27.1.3. Determination of the Nature of a Singular Point in the Linear Case.** Let

\[
\frac{dx}{dt} = ax + by \quad \frac{dy}{dt} = cx + dy \tag{27-3a}
\]

Consider the auxiliary variable

\[
z = ax + by \tag{27-3b}
\]

and let us try to find the values of the constants \( \alpha \) and \( \beta \) in order that \( z(t) \) be an exponential function of time, i.e., in order that

\[
\frac{dz}{dt} = Sz \tag{27-4}
\]

where \( S \) is a constant. Substitution of \( dx/dt \) and \( dy/dt \) from Eq. (27-3a) into Eq. (27-3b) yields

\[
\alpha(ax + by) + \beta(cx + dy) = S(ax + by)
\]

whence the two algebraic equations

\[
(a - S)\alpha + c\beta = 0
\]
\[
b\alpha + (d - S)\beta = 0
\]
These simultaneous equations in $\alpha$ and $\beta$ have a nonzero solution if the determinant
\[
\begin{vmatrix}
 a - S & c \\
 b & d - S
\end{vmatrix} = 0
\]  \hspace{1cm} (27-5)

This yields a second-degree equation in $S$ (S equation or characteristic equations) of the form
\[ S^2 + pS + q = 0 \]
with
\[ p = -(a + d) \quad q = ad - bc \]

Let $S_1$ and $S_2$ be its solutions, then
\[
\frac{dz_1}{dt} = S_1z_1 \quad \frac{dz_2}{dt} = S_2z_2
\]
whence
\[ z_1 = Ke^{S_1t} \quad z_2 = Ke^{S_2t} \]  \hspace{1cm} (27-6)

The nature of the singular point is governed by the nature of the roots $S_1$ and $S_2$. Complex conjugate roots correspond to a focus, real roots of the same sign give rise to a node, and real roots of opposite sign give rise to a saddle. The discussion immediately follows. $S_1$ and $S_2$ are real or complex according to the sign of $p^2 - 4q$. Furthermore,
\[ S_1 + S_2 = -p \quad S_1S_2 = q \]

Hence the results summarized in Fig. 27-2 in the form of a discussion in the $(p,q)$ plane:

a. The $q < 0$ half plane corresponds to saddle points.

b. The $q > 0$ half plane corresponds to nodes outside and focuses inside the $p^2 = 4q$ parabola.

c. Stable singular points correspond to $p > 0$, unstable ones to $p < 0$.

d. The $p = 0$, $q > 0$ half axis corresponds to singular points of the type of centers which are intermediates between stable and unstable foci.

27.1.4. Nature of a Singular Point. Nonlinear Case. In the nonlinear case, when the singular point has been brought to the origin by a change of variable, we have a system of the form
\[
\frac{dx}{dt} = ax + by + \text{terms of the second and higher degrees} \\
\frac{dy}{dt} = cx + dy + \text{terms of the second and higher degrees}
\]

whence, by the same process,

\[
\frac{dz_1}{dt} = S_1z_1 + \text{terms of the second and higher degrees} \\
\frac{dz_2}{dt} = S_2z_2 + \text{terms of the second and higher degrees}
\]

It can be shown that, for the purpose of the present problem, terms of degree greater than one may be neglected, provided the real parts of \( S_1 \) and \( S_2 \) are not zero. This property, first proved by A. Liapunov,\(^1\) of course applies to nonlinear systems with more than two variables. It allows the stability of the singular point \((x = 0, y = 0)\) of the nonlinear system defined by Eq. (27-7) to be studied from its first-degree approximation, defined by

\[
\frac{dx}{dt} = ax + by \\
\frac{dy}{dt} = cx + dy
\]

This is the so-called first-degree approximation of Liapunov. Referring to Eq. (27-6), the following properties immediately result: (1) If the roots of the \( S \) equation for the first-degree approximation all have negative real parts (that is, if the linearized system is asymptotically stable), the singular point is stable. (2) If at least one root of the \( S \) equation for the first-degree approximation has a positive real part (that is, if the linearized system is unstable), the singular point is unstable.

For the case in which one root has zero real part, or two roots are pure imaginary, the other roots having negative real parts (that is, if the linearized system is stable without being asymptotically stable), stability or instability cannot be deduced from the first-degree approximation but depends on the higher-degree terms (see below).

The practical and theoretical importance of Liapunov's theorems is considerable. The theorems constitute the background for the frequent approximation that consists in "linearizing" nonlinear systems by replacing nonproportional functions of the variables\(^2\) by the first terms of their series expansion. A typical example of this procedure is small-

\(^1\) See Bibliography, par. 5.1.

\(^2\) Note that some nonlinearities of the essential type, e.g., on-off elements or nonlinear elements with a dead zone, are not directly relevant to consideration by Liapunov's method. The latter can, however, be applied to nonlinearities of the saturation type (Fig. 22-21 with \( \varepsilon_m = 0 \)). If, then, \( \varphi \) is made to approach \( \pi/2 \), the case of the on-off element is arrived at.
signal approximation. It is felt intuitively that stability of the linear system guarantees actual stability, because it crushes “in the bud” (i.e., in regions which involve linear operation quite approximately) any tendency the system may have toward divergence. Liapunov’s theorems express this fact in a precise form. They state that, however the performance of the linearized system may differ from that of the actual system, the stability or instability of the latter in the vicinity of the singular point is not altered by the linearizing approximation except in borderline cases. It should be noted, however, that Liapunov’s theorems give no information about the region in which (i.e., how far from the equilibrium position) stability can be deduced from the first-degree approximation; rather, they concern only the “local” stability, i.e., the stability of the singular point. (This is stability in the small; see Glossary, Note 3).

Notions on the Behavior in the Critical Case. The critical case in which Liapunov’s theorems give no information was studied by Russian authors, especially by N. Bautin, ¹ who showed how the stability (or instability) in the critical case depends upon the terms of second or higher degree. For a second-order system

\[
\frac{dx}{dt} = ax + by + a_{20}x^2 + a_{11}xy + a_{02}y^2 + a_{10}x^2 + \cdots
\]

\[
\frac{dy}{dt} = cx + dy + b_{20}x^2 + b_{11}xy + b_{02}y^2 + b_{01}y^2 + \cdots
\]

the stability is controlled by the sign of a quantity \( L \) which is dependent upon the coefficients of the second-degree terms and also of two third-degree terms. The system is stable if \( L \) is negative, unstable if \( L \) is positive. The expression for \( L \) is given by

\[
-\frac{4b}{\pi} (ad - bc)^{\frac{3}{2}} = [ac(a_{11}^2 + a_{11}b_{02} + a_{02}b_{11}) + ab(b_{11}^2 + a_{20}b_{11} + a_{11}b_{20})] \\
+ c^2(a_{11}a_{02} + 2a_{02}b_{02}) - 2ac(b_{01}^2 - a_{20}a_{02}) - 2ab(a_{30}^2 - b_{20}b_{02}) \\
- b^2(2a_{30}b_{02} + b_{12}b_{20}) + (bc - 2a^2)(b_{11}b_{02} - a_{11}b_{20})] \\
- (a^2 + bc)[3(cb_{02} - b_{20}) + 2a(a_{21} + b_{12}) + (ca_{12} - bb_{21})]
\]

In the case of more complex systems the expression for \( L \) becomes extremely complicated, taking more than one page of printed text for a third-order system ² and four pages for a fourth-order one. ³

When \( L \) is negative, it can be shown that the system has reversibility. In other words, if an incidental change in the parameters of the system (say the addition of a lag \( \Delta T \)) causes instability, then there will result a limit cycle whose amplitude is infinitely small with \( \Delta T \) and tends toward zero if \( \Delta T \) disappears (Sec. 24.3.6). In this case the system is said to be on the verge of a nondangerous stability boundary. The momentary unstable state corresponds to the establishment of a small unstable region within the domain of the initial stable focus.

On the other hand, when \( L \) is positive, the system is not reversible; a small accident may throw the system into a limit cycle whose amplitude is not infinitely small with \( \Delta T \) (Sec. 24.3.6). The stability boundary is then said to be dangerous. Instability

¹ “Povedenie dinamicheskikh sistem v blizi granits ustojchivosti,” Gostekhizdat, Moscow, 1949.
² Ibid., pp. 60–62.
³ Ibid., pp. 117–122.
corresponds to the disappearance of the initial stable domain bounded by an unstable limit cycle.

27.1.5. Poincaré's Theorems. In Fig. 27-3 is shown a closed curve $C$ (which is not necessarily a trajectory) in the phase plane. Through every point $M$ of $C$ there passes a trajectory whose tangent at $M$ has a slope $m$. We shall now consider its rotation when $M$ follows the course of the entire curve $C$ in the direction of increasing $t$. If in these conditions the tangent makes $n$ revolutions ($n$ being either a positive or a negative whole number), then $n$ is known as the index of the closed curve $C$.

It can be shown (Theorem 1) that $n$ is linked with the nature of the singular points located within $C$:

a. If there are no such singular points, $n = 0$.

b. If there is a center of a focus, $n = 1$.

c. If there is a saddle, $n = -1$.

d. If there are several singular points, $n = N - N'$, $N$ being the number of centers, nodes, and foci, and $N'$ the number of saddle points.

Moreover (Theorem 2), in the particular case that $C$ is a trajectory,

$$n = +1$$

27.1.6. Interest. These theorems can help in finding the possible periodic solutions of the system, which give in the phase plane closed trajectories (limit cycles). Thus they play a fundamental role in the study of limit cycles. A notable consequence of Theorem 2 is that any limit cycle encloses at least one singular point; if it encloses several, they are of uneven number, with one less saddle point than the number of other singularities. These considerations, unfortunately, apply only to systems for which the phase-plane representation is adequate, i.e., for systems of the second order.\(^1\) For systems of higher order, such geometrical representations can be extended by topological reasoning: the approach then becomes essentially analytic. Furthermore, the condition that the functions $P$ and $Q$ should be continuous means that this approach cannot be directly applicable, at least in the form just quoted, to on-off control problems. This has made it necessary to extend these results to broader classes of functions $P$ and $Q$. This has been done in different

ways by workers in different countries. In France, B. Hamel\textsuperscript{1} has extended some theorems of existence to the case of discontinuous control.

27.2. LIMIT CYCLES, STABILITY

27.2.1. Criteria for the Existence of Periodic Solutions. There is a theorem of existence and a theorem of nonexistence of periodic solutions.

1. Theorem of Existence. If there exist two closed curves $C_1$ and $C_2$ such that (1) $C_1$ encloses $C_2$ (Fig. 27-4), (2) on any point of $C_1$ the trajectory enters $C_1$, and (3) on any point of $C_2$ the trajectory diverges from $C_2$, then one can affirm that there exists a limit cycle in the region bounded by $C_1$ and $C_2$.

This method can be applied to Van der Pol’s equation, curve $C_1$ being at infinity and curve $C_2$ being reduced to the origin. Other examples can be found in the literature.\textsuperscript{2}

2. Theorem of Nonexistence (Bendixon’s criterion). If within the phase plane the quantity

$$ h = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} $$

has a constant sign, there cannot be a closed trajectory within the phase plane.

Proof. If $V$ is the phase velocity vector whose components are $dx/dt = P$ and $dy/dt = Q$, the integral taken along a closed curve $C$

$$ \int_V \mathbf{dM} = \int_P dy - Q dx $$

has a value equal to that of the double integral in the area enclosed by $C$, namely

$$ \iint \left( \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dx \, dy $$

If, therefore, $h$ has a constant sign, this integral is not zero. Consequently, $C$ is not a trajectory; for if it were, we should have $V \times dM = 0$.

27.2.2. Stability of a Limit Cycle. The notion of stability relative to equilibrium may be generalized in the case of a steady state of the periodic type. There is stability if, after disturbance, the point $(x,y)$ resumes the same limit cycle under the same conditions of velocity.

\textsuperscript{1}“Théorie des systèmes à plusieurs degrés de liberté à commande discontinue,” \textit{C.E.M.V.} no. 27, Service Technique Aéronautique, Paris, 1950.

More precisely, a limit cycle \( x = x_T(t), \ y = y_T(t) \) is said to be stable if, given any positive number \( \varepsilon \), it is possible to find a number \( \eta \) such that, for all initial conditions complying with

\[
[x(0) - x_T(0)]^2 + [y(0) - y_T(0)]^2 \leq \eta
\]

the resulting movement of the point \((x, y)\) will satisfy

\[
[x(t) - x_T(t)]^2 + [y(t) - y_T(t)]^2 < \varepsilon
\]

for all positive values of \( t \). If, moreover,

\[
x(t) - x_T(t) \to 0 \quad y(t) - y_T(t) \to 0
\]

for infinite \( t \), the limit cycle is said to be "asymptotically stable."

Consider now the parametric equations of a limit cycle with period \( T \):

\[
x = f(t) \quad y = g(t)
\]

(27-10)

Starting from initial conditions,

\[
x = f(t) + u \quad y = g(t) + v
\]

(27-11)

Substitution gives (the prime indicating the derivative with respect to time)

\[
f' + u' = P(f + u, g + v)
\]

\[
g' + v' = Q(f + u, g + v)
\]

Hence, neglecting terms of higher order than the first (Sec. 27.1.4), it is seen that the incremental variables \( u \) and \( v \) are solutions of

\[
u' = u \frac{\partial P}{\partial x} + v \frac{\partial P}{\partial y}
\]

(27-12)

\[
v' = u \frac{\partial Q}{\partial x} + v \frac{\partial Q}{\partial y}
\]

Now, it is possible to apply a general theorem which states that the solutions of any linear differential system with periodic coefficients (period \( T \))

\[
u' = au + bv
\]

\[
v' = cu + dv
\]

may be written in the following form, using \( \varphi \) and \( \psi \) functions of period \( T \):†

\[
u = Ae^{h_1} \varphi_1(t) + Be^{h_2} \varphi_2(t)
\]

\[
v = Ae^{h_1} \psi_1(t) + Be^{h_2} \psi_2(t)
\]

(27-13)

with

\[
h_1 + h_2 = \frac{1}{T} \int_0^T (a + d) \, dt
\]

System (27-2) has a periodic solution; differentiating gives

\[
x'' = \frac{\partial P}{\partial x} x' + \frac{\partial P}{\partial y} y'
\]

\[
y'' = \frac{\partial Q}{\partial x} x' + \frac{\partial Q}{\partial y} y'
\]

† The authors are grateful to Prof. L. Gaulthier, of the Faculté des Sciences of Nancy, who pointed out to them the present approach.
of which one solution is \( x' \cdot f', \ y' = g' \). It may be deduced that \( h_1h_2 = 0 \), whence, for example

\[
h_1 = \frac{1}{T} \int_0^T \left( \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dt \quad h_2 = 0
\] (27-14)

There is stability if \( h_1 < 0 \), and instability otherwise.

The interest of this result is twofold. First, the case in which \( h_1 > 0 \) provides a new proof of Bendixon's criterion. Second, Eqs. (27-13) and (27-14) show the role played by the divergence \( (\partial P/\partial x + \partial Q/\partial y) \) in the manner in which the point \((x,y)\) approaches the limit cycle. Whatever the system and whatever the disturbance, a limit cycle is always approached (or gone away from) at an exponential rate the time constant \( 1/h_1 \) of which characterizes the intrinsic, or structural, stability of the limit cycle under consideration.

27.2.3. Liapunov's Stability Criterion. A method for determining the stability of a steady state has been developed by A. Liapunov.\(^1\) The method applies to general systems governed by Eq. (27-1), but we shall consider only the case in which there are two variables [Eq. (27-2)].

Liapunov's theorem is as follows: If \( x = 0, y = 0 \) is a singular point, and if it is possible to find a function \( V(x,y) \) complying with the conditions (a) \( V(x,y) \) is positive for all values of \( x \) and \( y \) except that it may be zero for \( x = y = 0 \) and (b) \( dV/dt \) is never positive, then the equilibrium is stable. If, moreover, \( dV/dt \) is never zero (except possibly for \( x = y = 0 \)), the equilibrium is asymptotically stable.

\(1\) See Bibliography. Applications of Liapunov's method have been recently developed by Russian authors, especially M. Ajzerman, "Teorija avtomaticheskovo regulirovanija dvigatelej: uravnenija dvizhenija i ustojchivost'," chap. 13, Gostekhizdat, Moscow, 1952, and A. Ljotov, "Ustojchivost' nelinejnykh reguliruemyh sistem," Gostekhizdat, Moscow, 1955. An elementary but clear presentation and discussion of Liapunov's method will be found in W. Hahn, "Behandlung zweier Stabilitätsprobleme mit der zweiten Methode von Liapunov," in "Nichtlineare Regelungsvorgänge," pp. 51–66, Oldenbourg, Munich, 1956. Our presentation and the examples that we quote are adapted from these three authors.
\[
\frac{\partial V}{\partial x} = \left[ \left( \frac{\partial V}{\partial x} \right)^2 + \left( \frac{\partial V}{\partial y} \right)^2 \right]^{1/2} \cos (Ox, MN)
\]

\[
\frac{\partial V}{\partial y} = \left[ \left( \frac{\partial V}{\partial x} \right)^2 + \left( \frac{\partial V}{\partial y} \right)^2 \right]^{1/2} \cos (Oy, MN)
\]

where \( MN \) is the normal to the \( V = \text{const} \) curve (or, in the more general case, surface) with the direction from inside to outside taken as positive. Hence

\[
\frac{dV}{dt} = \left[ \left( \frac{\partial V}{\partial x} \right)^2 + \left( \frac{\partial V}{\partial y} \right)^2 \right]^{1/2} V_s
\]

where \( V_s = P \cos (Ox, MN) + Q \cos (Oy, MN) \) is the normal projection of the velocity of the point \( M \). If \( dV/dt \) is never positive, this proves that any phase trajectory will cross the surface \( V = C \) from the outside to the inside; that is, the point \( M \) necessarily passes from a surface \( V = C \) to a surface \( V = C' \leq C \). Stability immediately results, since the condition \( x^2 + y^2 < \varepsilon \) (Sec. 9.1.1) will be satisfied for all \( t > 0 \) if the initial conditions \((x_0, y_0)\) are located inside a \( V = C \) curve so chosen as to be totally included within the \( x^2 + y^2 < \varepsilon \) region.

**Example 1.** Consider a one-degree-of-freedom system in which the restoring force is a function of the velocity,

\[
\frac{d^2x}{dt^2} + a \frac{dx}{dt} + b \left( \frac{dx}{dt} \right)^2 + x = 0
\]

or, in the form of Eq. (27-2),

\[
\frac{dx}{dt} = y \quad \frac{dy}{dt} = -ay - by^2 - x
\]

Suppose that \( a \geq 0 \) and \( b \geq 0 \) and exclude the case in which both \( a \) and \( b \) are zero. It is easily seen that the function

\[
V(x, y) = x^2 + y^2
\]

complies with Liapunov's assumptions, since its time derivative

\[
\frac{dV}{dt} = 2 \left( x \frac{dx}{dt} + y \frac{dy}{dt} \right) = -2(ax^2 + by^2)
\]

is negative everywhere except at the origin. Therefore, the equilibrium position at the origin is asymptotically stable. In the present case this is easily understood, since \( V(x, y) \) is the square of the distance of \( M(x, y) \) to the origin and the \( V(x, y) = C \) curves are circumferences centered at the origin. The quantity \( dV/dt \) indicates how the point \((x, y)\) moves toward or away from the origin. When \( a \geq 0, b \geq 0 \), all phase trajectories tend toward the origin, that is, the equilibrium is stable. It can be easily seen that the \( a < 0, b < 0 \) case corresponds to an unstable equilibrium, all phase trajectories moving away from the origin. If \( a > 0, b < 0 \), \( dV/dt \) is negative only if \( 0 < y^2 < -a/b \); stability is assured only for initial disturbances that do not cause the \((x, y)\) point to lie outside a horizontal strip in the phase plane.

**Example 2.** Study of a Type of Nonlinearity. Consider a system governed by equations of the form of (27-7) and assume that a nonlinear term exists only in the first equation. Assume furthermore that \( a, b, c, \) and \( d \) are independent of \( s \):†

\[
\frac{dx}{dt} = ax + by + f(x)
\]

\[
\frac{dy}{dt} = cx + dy
\]

† These equations can represent the phase-plane behavior of a nonlinear one-variable regulator when \( d = 0 \).
Consider the case in which the function \( f(x) \) is zero for \( x = 0 \), but is never zero nor infinite for \( x \neq 0 \), that is,
\[
 h|x| < |f| < H|x| \tag{27-16}
\]
where \( h \) and \( H \) are positive numbers. In other words, it is assumed that the non-linear element (Fig. 27-6) has no dead zone and that its characteristic is always included between two linear characteristics \( w = hx \) and \( w = Hx \).

![Fig. 27-6.](image)

Liapunov's first-degree approximation is governed by the equations (27-8) and is stable if (Sec. 27.1.4)
\[
a + d < 0 \quad ad - bc > 0
\]

Let a function of the form
\[
 V = \lambda_0(ax^2 + 2\beta xy + \gamma y^2) \tag{27-17}
\]
be taken as a tentative Liapunov function. Its time derivative is
\[
 \frac{dV}{dt} = \alpha(ax^2 + bxy) + \beta[cx^2 + (d + a)xy + by^2] + \gamma(cxy + dy^2)
\]
This expression is identical with \(-(x^2 + y^2)\) if the coefficients \( \alpha, \beta, \) and \( \gamma \) are so chosen that
\[
 \begin{align*}
 \alpha c + \beta b &= -1 \\
 b\alpha + (a + d)\beta + c\gamma &= 0 \\
 b\beta + d\gamma &= -1
\end{align*}
\]
If the solution of these linear numerical equations is chosen for \( \alpha, \beta, \) and \( \gamma, \) then the \( V(x, y) = \text{const} \) curves are ellipses (it can be checked that \( \alpha\gamma - \beta^2 > 0 \)) and the \( V \) function satisfies Liapunov's conditions for asymptotic stability.

Consider now the system obtained by adding an additional linear term \( kx \) to the first of the equations (27-8):
\[
 \begin{align*}
 \frac{dx}{dt} &= ax + by + kx \\
 \frac{dy}{dt} &= cx + dy
\end{align*} \tag{27-18}
\]

The time derivative of a Liapunov function of the form (27-17) is
\[
 \frac{dV}{dt} = -(x^2 + y^2) + akx^2 + bkyx
\]
which is always negative if $\beta^2k^4 + 4\alpha k - 4 = 0$ has two real roots $k_1$ and $k_2$ and if

$$k_1 < k < k_2 \tag{27-19}$$

If this condition is fulfilled, the Liapunov condition is satisfied and the system (27-18) is asymptotically stable. The same holds good for the nonlinear system of Eq. (27-15), since $f(x)$ can be written in the place of the additional term, provided the condition (27-19) holds.

Thus it is established that the conditions

$$k_1 < h < H < k_2$$

are sufficient for the system to be stable. Thus, use of Liapunov’s condition has made it possible to establish this without having to specify the nonlinearity $f(x)$ more precisely than by means of Eq. (27-16).

The method can be extended (by using matrices) to the case in which there is an arbitrary number $n$ of degrees of freedom and there are $n$ additional nonlinear terms.

The above examples show how Liapunov’s criterion enables one to study the stability of a great number of nonlinear servo systems for wide classes of nonlinearities. One of the essential advantages of the method is that it is rigorous and makes possible the specification of the domain in which stability is guaranteed.¹ Its major disadvantage is common to all algebraic criteria (Sec. 9.2.7): as soon as the complexity of the system increases, the method leads to extremely complicated expressions which are very hard to manipulate and interpret.

There is no absolutely general rule for finding Liapunov functions $V(x; y)$, which is a matter of special cases and of skill on the part of the engineer. Generally, functions of the quadratic form are tried, because their signs can easily be discussed. A. Lurje² has found a systematic method (the Lurje transform) for finding Liapunov functions for a wide class of problems, and the Lurje transform has been generalized by A. Ljotov.³ In our “Méthodes Modernes d’Étude des Systèmes Asservis,” chap. 14, Dunod, Paris, 1959, a summary of these questions is presented.

Those technical applications of the method which have been carried through to the numerical conclusion are not numerous. In those we know of⁴ the method leads to very severe conditions, yielding extremely small stability domains. This is very likely to be a consequence of the manner in which the problem is stated. The assumptions are very general (applying to wide classes of nonlinear functions), hence somewhat vague, and the required specifications (stability whatever the disturbances

¹ On the other hand, Liapunov’s theorems (Sec. 27.1.4) merely indicate whether the equilibrium position itself is stable or unstable (stability “in the small,” or “local” stability).
⁴ The only ones known to the authors will be found in M. Ajzerman, op. cit., pp. 454–480, and in E. Pestel, “Anwendung der Liapunovschen Methode . . . auf ein technisches Beispiel,” in “Nichtlineare Regelungsvorgänge,” pp. 67–85, Oldenbourg, Munich, 1956.
may be) are too severe. It can be reasonably thought that Liapunov's criterion can lead to results of interest if it is applied to problems which are more precisely defined from the technical viewpoint, i.e., to nonlinearities present in a narrow class of existing servomechanisms. The same is true if emphasis is laid on finding the stability domain, i.e., what disturbances will not throw the system into instability.

Figures 27-7 and 27-8 illustrate these remarks. In the case shown in Fig. 27-7, stability is guaranteed so long as the operating point lies within the ellipse that is tangent to the vertical $x_1$ where the nonlinear characteristic leaves the stable region $(h,H)$. In Fig. 27-8, which corre-

![Diagram](image)

Fig. 27-7.

Fig. 27-8.

sponds to the case of a nonlinearity involving a threshold, it is possible to state that large disturbances will lead to residual errors which are smaller than $x_1$. Stability cannot be guaranteed (there can be, for example, a limit cycle inside the ellipse tangent to the $x_1$ vertical), but in any case an upper boundary for the error has been found.

On the other hand, Liapunov's criterion merely guarantees mathematical stability, which is by no means a sufficient condition for satisfactory performance. It must be possible to set forth performance criteria for nonlinear servomechanisms by studying the time behavior of the Liapunov function $V(t)$ and its time derivative $dV/dt$. In this application, Liapunov's direct method may constitute a promising research field.
CHAPTER 28
LINEARIZATION. FINAL REMARKS

Summary
1. Linearization.
2. Concluding remarks.

28.1. LINEARIZATION

28.1.1. General. The word linearization is used for different concepts which may not be equivalent:

a. In its most common sense, to linearize a system is to write its equation in linear form, which amounts to assuming that the principles of proportionality and superposition hold. In most cases, linearizing a system consists of replacing series expansions of the variables by their first terms. A typical example is small-signal approximation (Sec. 2.2.1). More precisely, Liapunov has indicated to what extent a system can be represented by a linear scheme (Sec. 27.1.4).

b. In a more special sense, to linearize a system is to treat it as if it were linear, that is, to apply to it concepts that are, strictly speaking, applicable only to linear systems. For example, the generalization of the concept of a transfer function for nonlinear systems (Sec. 24.1) is a linearization.

c. In another special sense, a nonlinear system can be linearized if it is adapted in some special way to behave like a linear system. The following paragraphs are devoted to the linearization of on-off systems, the adaptation consisting of adding a suitable periodic function of time to the system input.

28.1.2. First Example: Vibrating Relay. Consider a two-position relay controlling a motor $M$ with two directions of rotation, and suppose that the relay characteristic is ideal. If the input voltage is sinusoidal and of zero mean value ($e = e_1 \sin \omega t$), the output voltage, and consequently the motor torque, is a square wave with zero mean value (Fig. 28-1).

If now the input voltage is a sinusoidal function whose mean value is different from zero, $e = e_0 + e_1 \sin \omega t$, then it can be seen from Fig. 28-2 that the output, and consequently the motor torque, is a periodic function whose mean value is $r_0$. Because the intervals where $e$ is positive or negative are not equal, $r_0$ is not equal to zero.

![](image)

Fig. 28-1.
A simple computation shows that the mean value of \( r \) is \( r_0 = 2M e_0/\pi e_1 \), provided that \( e_0 \ll e_1 \). Thus, the mean value of the output \( r_0 \) is proportional to \( e_0 \).

If the frequency \( \omega \) is high—for example, 100 to 200 rad/sec—and if \( e_0 \) varies slowly with time, say at a few cps, it is clear that \( r_0(t) \) will remain proportional to \( e_0(t) \).

Thus by means of a relay, which is essentially a nonlinear system, a linear system of input \( e_0 \) and output \( r_0 \) has been obtained. The method has consisted in taking as relay input the variable \( e_0 \) and superimposing on it a sinusoidal function of time whose frequency and amplitude are high relative to \( e_0 \). The analogy with modulation is striking. In particular, if \( e_0 \) varies rapidly, we may have a lag which is practically as long as the period \( 2\pi/\omega \).

A variation of this system, used in sweep generating circuits, consists in forcing the relay to oscillate by placing it in a "ringing" or damped oscillation circuit (Fig. 28-3).

28.1.3. Second Example: Vibrating Spoilers. Vibrating spoilers\(^1\) (Sec. 22.3.3) are controlled to vibrate continually at a frequency high

(5 to 20 cps) relative to the natural frequency of the aircraft or missile in which they are incorporated (Fig. 28-4).

In the case of roll control, if the command is zero, the spoilers on both wings flap symmetrically. A rolling control different from zero implies

\(^1\) These flapping spoilers differ from the ordinary spoilers described in Sec. 22.3.3, not only in their principle of operation, but also in their construction and their position in the neighborhood of the main wing torque.
displacement of the mean position of the spoiler on one side and an opposite displacement of the mean position on the other side: because of the high flapping frequency, the torque is proportional to the displacement of the mean positions. In this example, also, a proportional control has been obtained by using a nonlinear element, the spoiler.

28.1.4. Generalization. This method of linearization can be generalized in the following manner: if to the input \( e(t) \) of a nonlinear element we add an arbitrary sinusoidal function of time \( e_1 \sin \frac{2\pi Ft}{T} \), then if an odd system is considered [one in which the characteristic function \( r(e) \) is symmetrical with respect to the origin], the mean value of the output is proportional, in magnitude and direction, to \( e \).

If, on the other hand, an even system is considered, the output component, which is at the frequency \( \omega \), has an amplitude which is proportional, in magnitude and direction, to \( e \).

![Figure 28-5](image)

For an odd and discontinuous element, in particular for an on-off element such as those mentioned in Secs. 28.1.2 and 28.1.3, the proof is simple: it is sufficient to write (Fig. 28-5)

\[
\begin{align*}
r &= +r_m \quad \text{when } e_0 + e_1 \sin \frac{2\pi t}{T} > 0 \\
r &= -r_m \quad \text{when } e_0 + e_1 \sin \frac{2\pi t}{T} < 0
\end{align*}
\]

The difference between the signal durations \( +r_m \) and \( -r_m \) is equal to \( 4\tau, \tau \) being given by \( e_1 \sin (2\pi T/T) = e_0 \), whence, approximating the sine function by its argument (which is permissible, since \( e_0 \ll e_1 \))

\[
\tau = \frac{Te_0}{2\pi e_1}
\]

the mean value of \( r \) is

\[
r_0 = M \frac{4\tau}{T} = \frac{2Me_0}{\pi e_1}
\]

Other linearizing functions are possible. They must be periodic and close to their tangent in the appropriate zone. In particular, a saw-
tooth function allows of accurate linearization. Additional data can be found in the literature.¹

28.2. CONCLUDING REMARKS ON NONLINEAR SYSTEMS

28.2.1. A Short Survey of Nonlinear Servo Theory. As has already been pointed out, the nonlinear field is so vast that very few methods can be considered general: there are nonlinear problems, but there is no general nonlinear theory. The disadvantage of this for the designer of nonlinear control systems is that the results obtained for a particular class of systems cannot, in general, be extrapolated to include systems of very different types. The preceding chapters have outlined some methods which are among the more general ones and have discussed their respective interest in the servo field. Other methods are outlined in the works listed in the Bibliography.

As these discussions show, the methods and the systematic use of electronic computers make analysis of most nonlinear servomechanisms possible. But, so far as synthesis of nonlinear control systems is concerned, it must be admitted that we are now at approximately the same point as we were with the synthesis of linear systems in 1939. It is exceptional for the designer to be able to proceed in a straightforward manner from the specifications to the design. The considerable amount of research work at present being carried out in the nonlinear field² will very likely lead to important progress in the next few years in the synthesis of nonlinear servo systems. Such progress will probably be the result of a twofold trend:

a. An increasing number of particular problems are under study, and the corresponding results, expressed in condensed form, will make it possible for engineers to acquire experience with broader classes of problems.

b. The more general mathematical nonlinear theories are being applied to control problems, and concepts of interest in the automatic-control field are thus being elaborated.

28.2.2. Practical Conclusion. For the time being, when studying a nonlinear system, general methods such as Poincaré’s topological method are especially useful in order to arrive at an understanding of the functioning of the system, and numerical methods (using computing machines) for arriving at quantitative solutions. This is, in fact, the position in practice. For nonlinear problems which the authors have encountered, the following approach is usually employed: (1) The problem is first trimmed by applying an approximate method—generally approximation


² Unfortunately, a great part of this research work is carried out as part of classified projects in different countries, which prevents publication of many results of interest.
by the first harmonic. (2) The problem is then attacked simultaneously from the point of view of Poincaré’s topological method and the resolution of numerical cases, using electronic computers.

These various methods enable nonlinear servomechanisms at the present time to be quantitatively analyzed at the stage of the preliminary design.
PART FOUR

COMPONENTS OF SERVO SYSTEMS

CHAPTER 29

ERROR-SENSING DEVICES

Summary

1. General remarks on sensing devices and noise.
2. Classification.
3. Impedance sensing devices.
4. Generators.
5. Electronic sensing devices.
8. The choice of a sensing device.

In this part of the book the different fundamental components of a servo system, i.e., the sensing device, the amplifier, and the servomotor, will be studied from the viewpoint of their influence on the over-all performance of the system. (The design of the compensating network has been dealt with in Chaps. 10, 18, and 19.)

The present chapter is devoted to the error-sensing device. The choice of this component is most important in the preliminary design. In research problems, choosing a certain type of error-sensing device comprises, generally speaking, one of the major decisions of the whole project, since the very principle of the method used is strictly dependent on this choice. In all servo problems, the error-sensing device is the component which requires the greatest care on the part of the designer. The reason is that, whereas imperfections in the forward path of a closed-loop system may be compensated to a certain extent, imperfections in the sensing of the error will inevitably impair the basic performance of the whole control system. In particular, the limitation due to noise is especially critical at the error-sensing stage, this being due to the different amplifications that take place in the system after the error-sensing device.

29.1. SOME REMARKS ON SENSING DEVICES AND NOISE

29.1.1. Definitions. A sensing device is an instrument which transforms one physical variable into another, relating the two quantities by a simple law and using only a very small part of the mean power delivered by the servomechanism. The sensing device may include the summing element, or differencer, or it may not (Fig. 29-1). As shown in Fig. 29-1a, the sensing device measures two variables simultaneously. One is the servo input, the other is the servo output. It produces a signal proportional to the difference between the two variables, hence its name, error-sensing device. Alternatively, in Fig. 29-1b, the sensing element is indicated separately from the summing element. Its input
is the controlled variable. Its output, however, could be an electric signal proportional to the rate of the servo output, which then would be fed to the summing element.

![Diagram](image)

**Fig. 29-1.**

-Sensing element (gyro or rate gyro)

Reference, \[ \text{Amplifier} \quad \text{Control surfaces motor} \quad \beta \quad \text{Airframe} \]

Rate gyro \[ \text{Detector} \quad \text{Amplifier} \quad \text{Control surfaces motor} \quad \beta \quad \text{Airframe} \]

**Fig. 29-2.**

As an example, consider the angular deviation from a fixed reference, or the angular rate, of an aircraft. The device, a gyro, is fastened to the airframe (Fig. 29-2a), and the reading of the instrument is proportional either to the spin-axis gimbal deviation or to the angular rate of the airframe around some reference axis (Fig. 29-2b). In the latter case (rate gyro) the absolute angular velocity is measured. These instruments will be described in Sec. 29.6.

To characterize an error-sensing device, one usually plots the output-vs.-input curve; it is termed the *static, or gain-characteristic, curve*. It must be understood that the dynamic behavior should also be known. In the case of a linear system, this can be done by specifying the transfer function.

Error-sensing devices are precision instruments and need careful attention for their design and manufacture.

**29.1.2. Requirements Stemming from Information-theory Considerations.** It is known that the amount of information is always reduced by its transmission through a system. Since the sensing device is the first element in the forward path of the loop, it must be chosen or designed with a noise level as low as possible. The noise level will include all
the following types: noise from internal sources (such as thermal noise in resistances, shot noise in vacuum tubes) and noise from external sources, like that from the power supply of the sensing device, etc.

It is necessary to reduce these disturbances to as low a level as possible. For example, a phase-lead network (Sec. 19.2) behaves like a differentiator over a certain frequency band. It then differentiates the input, without discriminating between the message and the noise; this results in an increase in the noise level (see, for example, Fig. 29-11).

Consequently, it is necessary to consider the basic noise theory from a more elaborate and more practical point of view than was done in Chaps. 12 and 22. Some of the following sections will cover a field slightly outside the subject of error-sensing devices. On the other hand, the definition of threshold (dead-zone) saturation, and saturation-to-threshold ratio, which have been considered in Chaps. 11, 12, and 22, will not be reviewed here.

29.1.3. Noise Characteristics. On the microscopic or macroscopic scale every physical system is subjected to noise, if the latter is defined as an unwanted random disturbance with zero mean value which affects every variable of the system. This noise is not necessarily noticeable on a macroscopic scale, but it is always present (for example, as brownian motion).

A system without electrical components, a hydraulic servo, for example, is also subjected to microscopic noise, e.g., the pressure forces from the molecules on the enclosure. Yet this noise is not usually detectable, because of the constant presence of quasi-microscopic coulomb friction, which causes a dead zone. Contrariwise, most electrical systems have no dead zone. The noise has its origin in the threshold that corresponds to the charge of one electron. Thus, threshold and noise source have the same energy level, and the noise can be detected without too much difficulty. It is essential to note that the noise amplitude is independent of the control input.

In an electronic amplifier, the noise and the message are amplified by the same factor, provided the amplifier is operated in a linear region. For example, consider the gain characteristic shown in Fig. 29-4. The region OA is linear, while over AB the gain decreases as saturation is approached (BC). An input message consisting of large intermittent pulses will be more attenuated than a continuous small-amplitude noise. It can be said that the signal-to-noise ratio is decreased in such a nonlinear process. Evidently, if the amplifier is always operated in its linear region, the signal-to-noise ratio remains constant, provided the amplifier does not itself have any internal noise source. Generally speaking, the presence of nonlinearities decreases the quantity of information contained in the input signal; although this is more easily seen
in the case of saturation, it is true for any nonlinearity. If, for example, the message is the sum of two sinusoids of frequency \( \omega_1 \) and \( \omega_2 \), the output will involve unwanted cross-product terms of the form \( p\omega_1 + q\omega_2 \), where \( p \) and \( q \) are integers. The case of a dead zone (Fig. 29-5) is less evident. It seems that, in the absence of signal, the noise is not transmitted if its amplitude is smaller than \( OA \), except for the values greater than \( OA \). However if the input signal—noise and message—is of amplitude \( OB \), it can be shown that the signal-to-noise ratio will be decreased, particularly if it was high at the input.

Everything remarked for the amplifier applies to all components, particularly to the various error-sensing devices, which will now be analyzed. The effects of microscopic noises in the operation of certain electrical devices will be examined first. It will then be seen that microscopic noise is not at all negligible with respect to macroscopic noise.

Both types of noise can cause limitations; they constitute two different aspects of the same problem. Finally, the last type of noise to be considered will be called dynamic noise: it depends on the input and on the specific components used in the device.

29.1.4. Microscopic Noise. Three types of microscopic noise will be described. By definition, they are all generated by the discontinuous nature of electricity. They are intrinsically related to all electrical phenomena and cannot be decreased below their respective theoretical levels.

It is useful to note that certain physical phenomena have a theoretical limit, due either to disturbances whose level cannot be eliminated or to the actual physical operation of the device. This chapter and the two following ones will be concerned with disturbances usually found in servomechanisms, and the theoretical performance limit of electric motors will be briefly mentioned (31.1.3). It is very important to be aware of these theoretical limitations; they can be considered a criterion.

1 For example, it can be shown that the maximum resolution for an optical microscope is \( 1.22 \lambda /2n \sin \alpha \), where \( \lambda \) is the wavelength, \( n \) the index of refraction, and \( \alpha \) the aperture. Present-day microscopes achieve \( \alpha = 80^\circ \). Therefore, it is useless to try to improve the other parts of the microscope, since present proportions are close to the theoretical limit.
that enables one to judge what fundamental limitations to the improvement of an actual system exist.

Three particular aspects of electrical noise will be considered; they are all generated by the discontinuous nature of the electrons: (a) thermal noise of electrons in resistances, (b) shot noise in electronic tubes with thermionic emission, and (c) flicker noise, noted particularly in oxide-cathode tubes. For a more detailed analysis, the reader is referred to Blanc-Lapierre and Fortet,\(^1\) one complete chapter of whose work is devoted to the physical characteristics of noise.

a. THERMAL NOISE. The thermal agitation of electrons in conductors has been studied by Nyquist. The mean-square value of the random voltage measured across a resistance at an absolute temperature \(T\) is

\[
e^2 = 4kRT
\]

where \(k\) is Boltzmann’s constant: \(k = 1.37 \times 10^{-14}\) erg/deg.

This equation can be deduced from the equipartition theorem, which can be stated in the following manner: For a network in thermal equilibrium the mean value of the random energy, defined on a microscopic basis, is equally shared between the inductances and capacitors. Its value is

\[
\frac{1}{2}Li^2 = \frac{1}{2}C\dddot{v}^2 = \frac{1}{2}kT
\]

Further analysis of the mean-square voltage across the complex impedance \(Z = a + jb\), where \(a\) and \(b\) are, in general, functions of frequency, yields the following expression for the spectral density:

\[
\Phi(\omega) = 4ka(\omega)T
\]

To characterize this random process completely, it can be shown that \(e(t)\) has a gaussian amplitude probability density, and the mean time interval between collisions of the electrons is of the order of \(10^{-12}\) sec.

Briefly, resistors behave like noise generators. The power which they deliver over a bandwidth \(\Delta \omega\) is

\[
P = kT \Delta \omega
\]

At room temperature, \(e(t)\) has an rms value of about 10 \(\mu\)V, for a resistance of 1 megohm.

The expression just obtained is the absolute minimum value of noise. Impurities in the resistance material, nonuniform temperature distributions, etc., are factors which can increase the noise level above the minimum value. Grid resistances in amplifiers are usually large. If the stage has an appreciable gain, it effectively acts as a noise generator.

b. SHOT NOISE. The plate current of a vacuum tube is actually the total effect of the electrons, each of charge \(e\), arriving on the plate. The particles are emitted from the cathode according to Poisson’s law. Let \(i\) be the plate current. The deviation of the current is found to be

\[
[(i - \bar{i})^2]^{1/2} = 2\bar{e}t
\]

while the spectral density is uniform. For a plate current of 100 ma, the deviation is of the order of $10^{-9}$ amp. A current of $10^{-9}$ amp is generated by the emission of one electron every $10^{-14}$ sec.

For the analysis of noise in networks, it is often useful to replace the shot noise of vacuum tubes by an equivalent thermal noise. The resistance $\rho$ which must be added to the grid resistance (Fig. 29-6) to simulate the shot noise is termed the *noise resistance* of the tube. Then, the idealized representation of a vacuum tube (Fig. 29-6) is replaced by the more exact equivalent of Fig. 29-7. For a pentode $\rho$ is about 10,000 ohms.

c. **Flicker Noise.** This phenomenon is analogous to, and is usually superimposed upon, shot noise in vacuum tubes. It is caused by some atomic or molecular irregularities in the emitting cathode. It is characterized particularly by its frequency spectrum, which is not uniform, but decreases in proportion to $1/\omega^2$.

![Fig. 29-6.](image)

![Fig. 29-7. Noise resistance $\rho$.](image)

**29.1.5. Macroscopic Noise.** This type of noise depends, by definition, on the design and manufacture of the sensing-device components. Consequently, it can be improved by appropriate techniques; this fact distinguishes it from microscopic noise. For example, one of the noise sources in a hydraulic system is the high-pressure oil supply. The periodic motion of a piston pump will cause a pressure disturbance; an identical effect is to be observed in a gear-type or centrifugal pump. If the high pressure is supplied by an accumulator, the ducts and the valves will generate turbulent flows, again causing pressure fluctuations.

In electrical or electromechanical systems, the voltage (or current) supplies are always fluctuating. The voltage from a generator can vary randomly for many reasons: bad contact of a collector ring or brush, variation in electromotive forces in the rotor, etc. In dry cells and storage cells, chemical reactions render the electrodes heterogeneous, causing variations in the voltage and the internal resistance. This effect should be distinguished from the microscopic noise generated by the discontinuous flow of the ions to the electrodes.

As a final example, consider a noise which is on the borderline between the macroscopic noise just described and dynamic noise, i.e., noise due to motion, which will be briefly discussed in the next section. Any device and, more generally, any component of the servo loop which uses a signal-modulated carrier will introduce some unwanted signal. It will be shown in Chap. 33 how the message can be obtained from the modu-
lated carrier and how it is difficult to eliminate all but the message by filtering. Evidently the disturbance consists of a periodic signal whose harmonics are known for each type of demodulator. Nevertheless, this periodic signal can be considered as noise, even if it is not a random function, since it is an unwanted disturbance of the message.

A great number of other examples could easily be quoted. It is important to note that macroscopic noise is generally introduced in a servo loop at a high-energy-level point. On the contrary, microscopic noise is found in every element of the loop with a low energy level. For example, the maximum gain of an amplifier is limited by its background noise, while the accuracy of the over-all servo system may also be limited by the backlash between motor and load. These considerations are very important for error-sensing devices, which are usually electromechanical and generate both microscopic and macroscopic noise. They are precision instruments, and they must be carefully designed and manufactured.

29.1.6. Dynamic Noise. Finally, dynamic noise, or noise due to motion, deserves special attention. By definition, it is intrinsically linked to the motion, or to the variation, of an input variable, and this input is necessary for its specification. In most cases, it can be expressed as a function of the input by a set of algebraic equations. Although they are not random functions of the input, which may itself be a random variable, these disturbances are considered as noise because they are unwanted modifications of the variable concerned. An extremely complex analysis of the phenomenon is usually necessary to understand this noise, and intricate sets of equations do not easily yield a solution.

A simple example of dynamic noise is the case of the wire-wound potentiometer (Figs. 29-8 and 29-17). It can be shown that, when the arm is moving on the windings, voltages are induced in the switched-over turns and unwanted transients are picked up by the arm. For a simple case, the induced signal could be computed, but how could the results be used? It would probably be better to obtain experimental data and, using some statistical interpretation, evaluate together all the secondary effects, such as induced voltages, bad contacts, and small variations of temperature.

Before concluding, the linearizing effect of noise should be pointed out. As an example, consider the design of a device telemetering the weights of steel ingots for blooming. An accuracy of 0.2 per cent was required, but the 200-turn potentiometer used in the sensing device had a resolution of only 0.5 per cent. Nevertheless, the effective accuracy was found to be within 0.2 per cent. This is explained by the linearizing effect of the noise. In fact, the rollers over which the ingots were brought were
not perfectly balanced. The lack of balance generated a quasi-sinusoidal oscillation, on the roller base, which accounts for the linearizing effect of the noise in this particular case.\(^1\) Note that noise cannot always be assumed to be a linearizing factor; certain conditions must be met by its spectrum and probability density. These details will not be considered here.

\[\text{Voltage}\]
\[\text{Position}\]

**Fig. 29-9.**

\[\text{Voltage}\]
\[\text{Output}\]
\[\text{Input}\]
\[\text{Position (time)}\]

**Fig. 29-10.**

### 29.1.7. Examples

1. **Low-resolution Potentiometer.** The relation between the arm position and the steady-state output voltage is shown in Fig. 29-9. Threshold, saturation, and saturation-to-threshold ratio can be easily seen from the figure.

   The stepped form of the output is often considered as noise. To define the noise level of the potentiometer, let us consider an input signal \(e_i\), shown in Fig. 29-10 by the dotted line. Note that time does not appear explicitly in the relationship. The output voltage \(e_o\) is the stepped curve, which is essentially discontinuous. The potentiometer noise will be defined as the mean-square error between \(e_o\) and \(e\),

\[
\frac{1}{2} \int_0^\infty (e_o - e_i)^2 \, dx
\]

If induction effects are assumed negligible, the mean-square error is constant over the input interval \(X\), regardless of the time involved in traversing it.

A more detailed analysis would produce two conclusions. First, it is not necessarily true that the noise level will decrease if the number of turns is increased. Second, if the potentiometer is followed by a differentiating element, the output of the latter will be much more discontinuous than the input. It can be said that the differentiating element amplifies the discontinuous noise more than the signal. A typical output voltage is shown in Fig. 29-11.

![Fig. 29-11. Effect of differentiation.](image)

Obviously, the output of a potentiometer should not be differentiated. If, for example, the static accuracy and the bandwidth specify the introduction in the loop

\(^1\) The authors are grateful to M. Fouassin, of the Aciéries d'Ougrée Marihaye, Belgium, for this example.
of a lead-lag (differentiating-integrating) network, it is better, because of noise-level considerations,\(^1\) to pass the signal through the lag network first. Other considerations, such as impedance matching, for example, indicate that the passive-network realization is easier if the differentiating part of the network is placed before the integrating part (see Sec. 18.4). The designer must choose the best compromise for the particular problem.

Even though the noise level and the accuracy are two different characteristics, they are implicitly linked by the mean-square equation.

2. Electronic Amplifier. Let a supposedly noise-free input signal be amplified. The output includes a signal, related to the input by the transfer function \(F(s)\), and a noise component \(n(t)\) (Fig. 29-12). The noise is generally assumed to be linearly superimposed on the message. The noise level is characterized by \([n^2(t)]^{1/2}\) and may depend on the input, a correlation between \(n(t)\) and \(e_i(t)\) being possible. If the input signal is the sum of a noise \(n(t)\) and a message, the output will involve the response to \(n(t)\).

3. Pick-off or Measuring Device. A measuring device cannot distinguish between the message and the noise. Although the noise has zero mean value, it affects the accuracy of a static measurement, if its low-frequency power spectrum is not negligible. The reading will oscillate around a mean value, but the accuracy is better than \([n^2]^{1/2}\). The noise can sometimes increase the sensitivity by eliminating static friction. For example, the instrument panel of a jet aircraft, where there is very little vibration, is sometimes mounted on mechanical vibrators to decrease the static friction of the instruments. This is called the dither effect.

29.1.8. Conclusions. The number of examples could be multiplied, but the above three are sufficient to indicate how noise characteristics of systems are difficult to analyze, and how the diversity of noise types makes it necessary to define them precisely for each system analysis.

It is also necessary to distinguish between the notions of threshold, saturation and saturation-to-threshold ratio, and the bandwidth. Furthermore, most of the variables are random time functions. Nevertheless, it is possible to compute the necessary bandwidth to pass the useful signal with a minimum error, the mean-square error, for example.

This short study of noise has been a necessary introduction to the design and the choice of the servo components, particularly of the errorsensing devices. Typical sensing devices will be described in the following sections.

29.2. CLASSIFICATION OF SENSING DEVICES

Sensing devices can be studied in different ways. One way is to classify them according to the variable they measure, another is to classify them according to the physical laws they involve. A third approach consists in classifying the devices according to the elements which always constitute them:

\(^1\) Recall that lead-lag networks imply no contradiction, since the lead and the lag parts cover different frequency bands.

\(^8\) See footnote in Sec. 12.3.4.
a. The *pick-off*, which senses the physical variable: temperature, pressure, angular velocity, etc.

b. The *transducer*, which transforms the sensed variable into another quantity (mechanical, electrical, etc.) which is easier to transmit

From the information-theory point of view, the transducer constitutes the transmission channel, which should be used in an optimum manner.

According to this classification, certain devices, such as a potentiometer, which detects a relative position, can be analyzed directly, or they can be considered as formed by the two elements mentioned above. A strict classification will not be used, since it would only complicate the description of the devices. The classification chosen is the following:

a. *Impedance sensing devices* measure the variation of resistance, inductance, capacitance, inductive coupling, etc.

b. *Generator sensing devices*: tachometers, a-c generators, thermocouple, photocells, crystals, etc.

c. *Electronic sensing devices*: only recently developed and still in the experimental stage, are classified according to the variable sensed.

d. *Position sensing devices*, usually constructed with components of class a or b.

e. *Angular-velocity sensing devices*. Two types must be considered: (1) Those measuring a relative speed between two variables. They are built with elements of class b. (2) Those measuring an angular speed relative to an inertial reference. They are essentially gyroscopic instruments cascaded with transducers of class a or c.

f. *Angular-acceleration sensing devices*, which measure the acceleration of their base with respect to an inertial reference. The accelerometer pick-off is used with a transducer of class a or c.

g. *Pressure, temperature, sensing devices, etc.* Their transducers are used with a special pick-off: manometer, thermocouple, etc.

The types of sensing device should not be defined too strictly; such would only result in needless complexity. For example, certain passive

![Diagram of a spin axis gyroscope](image)

**Fig. 29-13. ECA rate gyro (see Sec. 18.2.2, par. 3).**

components are often used to modify a transducer output: a gyroscope, the ECA 84, includes a mechanical phase-lead element, introduced between the pick-off (the gyro) and the transducer (the potentiometer) as shown in Fig. 29-13. To conclude, let us insist on the necessity for completely specifying the sensing device before starting the design of a servomechanism.

The plan that will be followed in the next sections is a combination of the two classifications described above (a, b, and c and d, e, and f).
29.3. VARIABLE-IMPEDANCE SENSING DEVICES

29.3.1. Variable-resistance Sensing Device. The variable-resistance sensing device is the type most commonly used. From the equation $R = \rho l/s$, where $\rho = \rho_0(1 + at + bt^2 + \ldots)$ it is seen that resistances may vary for many reasons, the main one being variation in length. For the time being, consideration will be limited to only this type of sensing device; the factors which produce a variation of resistance will be studied later.

29.3.2. Potentiometric Sensing Device. To eliminate temperature variations, use is often made of circuits including a potentiometer as shown in Fig. 29-14. The two voltages $V_a$ and $V_m$ are measured, and their ratio is taken. This circuit is very useful when employed in conjunction with a second similarly connected potentiometer. Many linear or angular-position sensing devices are based on this principle. They are also used in rate-gyro transducers, pressure transducers, etc.

The potentiometer is, however, a very delicate device and is seldom satisfactory because the voltage it produces is discontinuous. Helically-wound potentiometers deliver an accurate continuous voltage, but they are delicate instruments, the slider being displaced on a helicoidal shaft.

Sensitivity. It is obviously impossible to detect an input variation unless it produces a change in voltage. This minimum variation corresponds to the pitch of the winding on the potentiometer. For example, a potentiometer of which the available core length is 10 cm, the resistive wire diameter is 0.08 mm (rarely less than 0.02 mm), and the coil pitch is 0.1 mm would offer a maximum of

---

1 As was mentioned in Sec. 29.1.6, there are cases where a better sensitivity is possible, because of the linearization effect of noise; in practice, however, this phenomenon should not be too greatly relied upon.
1,000 different voltages. It would then be impossible to have a resolution of better than $1/1,000$ of the voltage or of the corresponding displacement as shown in Fig. 29-17.

*Threshold.* The minimum displacement that can be indicated would also be $1/1,000$ of the total possible displacement.

*Saturation.* The saturation corresponds to the slider mechanical stop.

*Range of Measurement.* The saturation-to-threshold ratio is also 1,000. These 1,000 different steps will all be equal if the wire is properly wound and homogeneous.

*Noise Level.* See Sec. 29.1.

![Fig. 29-17](image)

![Fig. 29-18](image)

**29.3.3. Strain gauges.** 1. *Description.* Variation of the length or cross section of a resistive wire is made use of in strain measurements. This method is gaining importance in the aeronautical industry structures (wind tunnel and flight tests). Note that the strain-gauge constitutes both the sensing element and the transducer. The principle is as follows: A fine resistive wire bent as shown in Fig. 29-18 is rigidly fixed onto a surface the elongation of which is to be measured. The variation of the wire resistance is, therefore, proportional to the elongation or stretch of the surface under deflection.

In shape, strain gauges resemble postage stamps. They are constructed of wire of 15 to 20 microns in diameter, rigidly glued to a thin sheet of paper. The mounting of these gauges is difficult, because they must take on exactly the shape of the surface on which they are set. In addition, before they can be used, the glue must be perfectly dry.

2. *Strain-gauge Sensitivity.* The ratio of the strain-gauge resistance variation to the variation in length is called the gauge factor. This sensitivity is given by

$$\frac{dR}{R} = 1 + 2\sigma + \frac{dr}{r}$$

where $\sigma = \text{Poisson's ratio for the material}$

$r = \text{specific resistivity of the wire material}$

$dr = \text{variation when the material is stretched}$

This is actually an empirical formula, since there is some uncertainty about the validity of Poisson’s ratio for extremely thin wires. Moreover, the sign of $dr$ is unknown.
3. Static Measurements. These are the measurements that are performed on a material in which the stress is constant or slowly varying, as opposed to dynamic measurements which take place when the material is subjected to rapidly varying stresses such as those due to structural vibrations. In practice, the strain gauge, including a thermal compensator, is mounted on a Wheatstone bridge which is balanced at zero before the measurement is made. Two zero adjustments (Figs. 29-19 and 29-20) are usually performed. One ($P_1$) is the adjustment before the stress is applied, while the other ($P_2$) is used to zero the deflection due to a stress. The adjusting knob of $P_2$ can be calibrated to read the elongation or stretch directly.

![Fig. 29-19.](image)

![Fig. 29-20.](image)

4. Dynamic Measurements. These measurements concern phenomena varying at frequencies from 1 to 2,000 cps, thus covering the usual range of frequencies encountered in the study of servomechanisms. In this case the difference voltage is amplified electronically; it can then be fed directly into an oscillograph or into the amplifier stage of a servo system.

29.3.4. Resistance Variation with Temperature: Thermistors. The resistivity variation with temperature is of the order of a fraction of 1 per cent per degree for most materials; for example, 0.07 per cent for iron and 0.04 per cent for copper. This variation is, however, much smaller for certain alloys and is almost zero for manganin (84 Cu, 12 Mn, 4 Ni) and constantan (60 Cu, 40 Ni). Conversely, some thermistors have a negative exponential increment with respect to temperature, varying, for example, from 60,000 ohms at 0°C to 400 ohms at 200°C.

29.3.5. Capacitance Sensing Devices. The capacitance $C$ of a capacitor is $C = kA/4\pi d$, where $A$ is the surface area, $d$ is the distance between capacitor plates, and $k$ is the dielectric constant of the material.

Small surface and plate-separation variations can be used to measure small or large displacements with hardly any power input. This is the essential difference between this type of sensing device and the microphone type, which will be described in a subsequent section.

Capacitance variations are also utilized in connection with strain gauges and for sensing purposes in servomechanisms. Two different circuits are used for capacitor-type sensing devices. They are described below.
1. **Sauty Bridge Circuit.** The variable capacitance is so mounted that it forms a Sauty bridge (Fig. 29-21) with one fixed capacitance and two resistances. The difference voltage \(AB\) is amplified and demodulated as explained in Chap. 33.

2. **Frequency Modulating Circuit.** If the sensing capacitor is connected in parallel with the \(LC\) branch of an oscillator, every increment in the capacitance causes a frequency variation. If the resultant signal is passed through a demodulator, a negative or positive voltage will result, depending on whether the frequency obtained is smaller or larger than the natural frequency of the system.

3. **Variation of \(k\).** The variation of the dielectric constant of a capacitor permits interesting measurements. One application of this is the electronic fuel gauge shown in Fig. 29-22. Similar sensing devices can of course be used in servomechanisms, for example, when controlling the level of a liquid in a reservoir.

![Diagram of Sauty Bridge Circuit](image)

![Diagram of Electronic Fuel Gauge](image)

29.3.6. **Variable-inductance Sensing Devices.** In this class of device, the signal sensed appears as an inductance variation. As is true of capacitances and resistances, the inductance variation could be the result of changes in the mechanical dimensions of the unit. It is preferable, however, to act upon the iron core, either by altering its position or through saturation.

1. **Adjustable-core Inductance.** The adjustable-core inductance is one case where the current-carrying part of a sensing device is fixed. It is very sensitive; at frequencies of a few thousand cycles it is possible to detect micrometric displacements of an iron slug into a winding. This mobile slug can be made light and of low inertia.

2. **Saturable-core Reactors.** An iron-core inductance can be considered to have a definite value if it is carrying relatively light currents. When this current is increased, the inductance decreases; with sufficient current, the iron core can become saturated, yielding zero inductance at greater
current intensities. If alternating currents are used, at saturation there will be distortion of either the current or the voltage, depending on the associated circuits.

There are available magnetic materials, such as permalloy, sensitive enough to be saturable by the earth's magnetic field. This property is used in airplanes, where devices using these characteristics constitute a reference for gyros. It is also feasible to obtain inductance variation by circulating direct current in an additional winding wound on the same core. Care must be taken, however, to prevent the arrangement from acting as a transformer loading into a low resistance. This subject will be developed further in the section on magnetic amplifiers.

3. Variometers, or Synchronous Tranducers. Classified as such are the instruments which transform an angular position into a magnetic-field direction. These instruments vary in principle, design, and construction, depending on the designer and the intended use of the instrument. After producing various models, manufacturers now seem to have standardized on a unit resembling a synchronous motor with a two-phase rotor and a three-phase stator. Synchronous-motor principles have consequently been adopted, more or less.

Important developments since World War II on synchronous transducers, and also their introduction into fields other than servo service, justify further detailed study. They are currently used as remote indicators. They might also be used to control a light load, provided the load torque, inertia, and friction are very small.

29.3.7. Synchronous Transducers. 1. Principle of Operation. Figure 29-23 shows the wiring connections used on these systems; the stators of both machines are paralleled, and their rotors also. An alternating voltage is fed to the rotors. If the rotor of one unit is free to rotate and the rotor of the other unit is turned to a definite position, the rotor of the first unit will adjust itself to the same position.

The stator winding of a synchronous motor develops a rotating field of constant speed, of value dependent on frequency. Under zero torque conditions, the angle between this rotating field and the poles is zero.
When a load torque $C$ is applied, the rotor takes on an angle $\alpha$ with respect to the field, with or without oscillation. $\alpha$ is a function of the torque, as shown by the expression $\sin \alpha = KC$, which can be written $\sin \alpha = C/C_m$, where $C_m$ is the maximum possible motor torque, i.e., the stalling torque. If $\alpha$ increases, $C$ will also increase; so that

$$\frac{dC}{d\alpha} = C_m \cos \alpha$$

which is positive for $\alpha < \pi/2$. This explains the fundamental characteristic of synchronous motors. $C_m$ is a function of the frequency and of the reactance of the stator windings.

Assuming the system of Fig. 29-23 to be fed by direct current, this synchronous-motor theory could be applied without further discussion in the case of a particular driver-unit speed. Alternating current is used, however, because it allows for correct alignment of the two rotors even when they are at rest.

To recognize synchronous-motor properties in this application, it is necessary to analyze the alternating-current field. It must be considered as being composed of two rotating fields, constant in value and symmetrical with respect to the direction of the stationary field. Their intensity is half that of the maximum field, as shown in Fig. 29-24. (Further details on these problems will be given in Sec. 31.5.1.) If the driver unit turns at an angular speed $\Omega$ and its two imaginary fields at $\omega$ and $-\omega$, the composite result will be two fields turning at $\Omega - \omega$ and $\Omega + \omega$. Because of the rotating fields, two torques $C_1$ and $C_2$ will be developed in the rotor. It is apparent that the resultant of these two torques causes rotation of the rotor as if in a field rotating at a speed $\Omega$. These units can be delta-connected (Fig. 29-23) or star-connected (Fig. 29-25).

2. Use as Remote Indicators. The application of synchros for remote indication is simple. The driver unit is attached to the indicator shaft whose reading is to be transferred, while the receiver unit is left free to take the same angular position. Because friction is almost absent on these units, there is faithful receiver duplication of transmitter positions. Sources of error nevertheless exist:

a. Precisely speaking, there is friction in the receiver, and the rotor will not rest in the position corresponding identically to that of the field, but will take an equilibrium position where the synchronous torque produced is equal to the friction torque.

b. When the source of power to the driving unit is low or when it has some internal impedance, some error will occur in the signal appearing at the transmitter output.

c. There are also errors due to tolerances in the construction of these instruments. To check this, the two units, connected as in Fig. 29-23, may be mounted on a jig
with the two shafts along the same axis. One shaft is held fixed, while the jig holding the stators is rotated about the shaft axis. Errors will be indicated by the displacement, if any, of the free rotor; this error is generally smaller than 1°, a negligible quantity in remote-indication applications, since reading errors are already comparatively large. To reduce the errors, two units are often used, one on direct drive as above, to do the rough positioning, and the other unit, driven at a higher speed (usually of the order of 16 times), to do the fine signaling. With a precision of 1° for each synchronous drive, the system will permit a net accuracy of 1/16°. Its application is evidently more difficult, since it is more complicated and has speed limitations.

In certain applications, it is desired to synchronize the movement of two identical objects when the disturbing forces are small. Two identical and sufficiently large driving units, one being tied to one object and the other to a second object, are connected in opposition. The causes of speed differences will thus be opposed.

3. *Remote-control Application.* This is the most interesting application of these instruments; a simple case is illustrated in Fig. 29-25. After amplification, the receiving-unit output voltage is applied to a motor driving the receiver rotor supplying this voltage. The motor rotates in a direction corresponding to the phase of this voltage and stops at a point where the latter becomes zero. Figure 29-25 also shows the voltage variation as the rotor is turned through 360°, the starting point being the position of the rotor when it is displaced 90° from the stator field.

Two stop, or equilibrium, positions appear to hold for the rotor, that is, 0 and π; however, only one of these is stable. Now, suppose the system to be under false equilibrium at 180° and let the driving unit be moved slightly. A voltage will appear across the receiver rotor, causing the associated motor to turn the unit until it reaches correct and stable equilibrium at 0 or 2π, assuming, of course, good design and movement without oscillation.
In applications of this type, errors are limited to construction tolerances, friction effects being eliminated through the use of an amplifier and a strong auxiliary motor. As mentioned earlier, if still greater precision is required, a pair of machines may also be used in this particular application. One of these is used for coarse alignment, the other for fine alignment.

4. Related Devices. In order to perform the addition of several angles (algebraically), use is made of instruments similar to those described above except that the rotors and the stators are three-phase wound. Evidently (Fig. 29-26), this makes possible the introduction of an angle other than that established by the stator windings. Such a device is also used to transmit a signal which must include an angular correction, as in aerial fire control; further, it may be used as a phase-adjusting device if one set of windings is fed from a three-phase supply.

Arrangements using two-phase stators and single-phase rotors can be considered as of the same family. Their output is proportional to the sine or the cosine of the rotor angle. In fact, they transform polar coordinates to cartesian coordinates. They are commonly called "resolvers"—or, sometimes, phase changers.

5. Magnesyn Device. These instruments have annular stator windings, and taps are taken at 120° from the end terminals. The rotors are permanent magnets, and alternating-current voltage is fed to the stators.

Such a remote-control indicator comprises two units tied in parallel across the line (Fig. 29-27). The magnet creates zones of saturation, causing current distortion or unbalance; there will be equilibrium only when the magnets are correspondingly positioned. These instruments are extremely simple and can be built in very small size, but their torque is only a fraction of that obtained from wound rotors. They are perfect when used as remote indicators but almost useless when used as sensing devices in servo systems.
6. More Complicated Instruments. The Patin transmission is an example. Its transducer consists of a circular potentiometer with a d-c voltage applied to two sliders on opposite sides and with three taps placed 120° apart, as shown in Fig. 29-28. The receiver stator is three-phase wound; the rotor is a permanent magnet. Again in this case, the magnet takes a position determined by the resultant field of the three windings; and this, in turn, depends on the positions of the sliders at the transducer.

The operation of this device differs from that of those previously mentioned in that it is not reversible; in other words, the receiver cannot act on the transducer. It is, however, subject to brush-contact trouble.

Undoubtedly a variety of combinations could be conceived along these lines, using potentiometers, permanent magnets, synchronous machines, or even commutator-type machines. Designers, however, seem to favor the synchronous type increasingly, and they are developing units of more universal use and of better, or even very good, precision.

Following this line of thought, the difference between a synchronous or induction motor and a synchronous variometer should be noted by the reader. The motor is rugged and insensitive to phase unbalance; on the other hand, the variometer imposes the requirement of precise phase balancing, as well as identity of characteristics, on all the units working together.

Selssyns, or instruments of the same type, are precision instruments and should always be treated as such.
7. **Magnetic Bridges.** This is a kind of moving-iron differential transformer as shown in Fig. 29-29. When the rotor is centered, the induced voltages in the windings A and B are equal and opposite, making the output voltage equal to zero. If the rotor is moved from this position, the induced voltages will no longer be equal, and the direction and amount of unbalance will be indicated by the phase and intensity of the output voltage.

![Fig. 29-29. Magnetic bridge.](image)

**29.3.8. Synchronous Motors Used as Syncho Detectors.** In all of the previously described systems, the available torque is very small and is always too small to drive a load of inertia more than a few per cent above that of the rotor. Since the friction torque always causes errors, it must be as small as possible.

In general, aside from remote indication, the selsyn\(^1\) transmission is not directly used. However, the principle of the selsyn transmitter and of the selsyn receiver is used in the transmission of unlimited angles (Fig. 29-30).

![Fig. 29-30. Conventional synchro detector.](image)

The rotor of the selsyn receiver is driven by a motor and thus assumes a position such as to nullify the emf induced in it. The feedback control system generally used consists of a two-phase motor, the control field of which is connected to the rotor of the receiver selsyn through an amplifier \(A\). The other field is connected to the a-c power line (it is important to note that the second field must necessarily be connected to the same reference as the rotor of the selsyn emitter). The torque available at the motor \(M\) is determined by the characteristics of the load; taking into

\(^1\) We use the expression *selsyn* for simplicity. Obviously any similar apparatus can be used.
account the rotor inertia in the selsyn $R$, the power required is delivered by the amplifier $A$ and the field $I_2$.

Alternate Method. The principle of the feedback control system is based on the production of an alternating field the direction of which, with respect to a stated reference, is at an angle equal to the angle to be transmitted. A three-phase stator is not necessary and can be replaced by a two-phase stator (Fig. 29-31).

![Fig. 29-31. Two-phase synchro detectors.](image)

It should be noted that such an alternative method does not reduce the number of leads, but the three motors $S_1$, $S_2$, and $M$ can be two-phase motors, and identical if so desired.

29.4. GENERATORS AS SENSING DEVICES

29.4.1. The Commutator-type Dynamo. This is the instrument most frequently used when it is required to transmit signals proportional to (or related to) velocity. The voltage generated by a dynamo is proportional to its angular velocity, when the flux is constant. This rule holds for a wide range of speeds; however, when load is applied (assuming separate excitation), this linearity ceases and the voltage relation is given by the group of curves pertinent to the dynamo.

If the dynamo works into a resistive load, Fig. 29-32 shows the resulting characteristics. Note that, for a constant load resistance $R$, the different operating points fall on a straight line $OM$. The information needed is determined from the curves in Fig. 29-32; there is no reason why the line $I(\Omega)$ should be straight (Fig. 29-33). It is seen that the curve of Fig. 29-33 approaches more closely the dotted straight line as load currents diminish. Now, let us look more closely into the generator output cur-
rent. At an average value, as shown for corresponding operating points on the curves, there appears a superimposed noise caused by the discontinuity of voltages on the commutator bars. Figure 23-34 shows the voltage build-up on a commutator-type dynamo. The noise is relatively more pronounced at the lower speeds and, near zero speed, the information given becomes useless. This considerably reduces the speed range within which the dynamo delivers acceptable signals. Tachometer-type dynamos differ from the industrial types in that their armature reaction is low and their commutators are of special design; if the dynamo is intended for low-speed operation, its commutator must have as many bars as possible.

Speed Regulators. In this common application, the speed remains practically constant, fluctuating slightly by a few per cent about a mean value; commutator noise is then of a constant and rather high average frequency. If \( \omega = 50 \text{ rps} \) and there are 25 commutator bars, the fundamental frequency will be 1,250 cps; this may easily be filtered out, since the passband required to transmit information on the speed of a servo component falls in general much below 50 cps.

29.4.2. Commutatorless Dynamos. There is a type of industrial dynamo (used in electrochemistry to produce currents of the order of 20,000 amp at low voltage) which has no commutator bars; this is the Poirson dynamo shown in Fig. 29-35. Its principle of operation is well known: a conductor (in this case, an imaginary cylinder between the brushes \( A \) and \( B \)) cuts a magnetic flux (circulating within the rotor proper). There are, of course, several brushes located on the side surfaces of the rotor. The voltage generated by such a machine is continuous and proportional to angular velocity, by Faraday's law of induction. Unfortunately, with the small machines used for servo operation, available voltages are very weak: millivolts for service at the lower speeds; at maximum speeds, only a few volts.
Accurate servos intended for service within a wide range of speeds could be equipped with such tachometer dynamos. It would not be advisable, however, to use such a machine for regulator service.

It must not be overlooked that these dynamos must be used in conjunction with an amplifier, which adds its own inherent noise.

29.4.3. Two-phase Motors as Generators. Two-phase motors will be studied in Chap. 31. This type is conventional, except that the total ferromagnetic circuit is stationary and the only rotating part is a thin conducting "bell" of very low inertia (Fig. 29-36). The stator consists of two separate windings. When the device is used as a motor, each winding is fed with a voltage in time quadrature with the voltage applied to the other, thus creating the necessary rotating field. It is found on a unit of this type that, if one winding is fed an a-c voltage of constant amplitude and frequency, a voltage of the same frequency will appear across the other winding. The value of this voltage is proportional to the angular velocity of the rotor. This is difficult to explain theoretically on account of the complexity of the eddy currents in the rotating bell. Figure 29-37 shows the results obtained experimentally on a motor of this kind operating as a generator with an excitation of 36 volts at 500 cps. The unit tested was a two-phase Micromoteur SFENA type 11.01.† As seen from Fig. 29-37, its linearity is remarkable, while its threshold voltage is low; for zero speed \( N = 0 \) the induced voltage is 0.15 volt. The velocity range is wide—say, 20 to 30,000 rpm. Although the manufacturer's specifications limit the maximum speed to 18,000 rpm, for mechanical reasons, the unit was checked at 30,000 rpm, twice synchronous speed, with good linearity.

29.4.4. Other Electrodynamic Units. This class comprises those which generate voltage by the displacement of a coil within a magnetic field, that is, types such as the dynamic microphone whose operation is reversible—microphone or speaker. They can be used advantageously to detect low-amplitude translational velocities.

29.4.5. Electromagnetic Sensing Devices. In this case, current is produced by a magnetic-flux variation cutting a fixed coil. The flux variation can be produced in more than one way; it can be produced either when a magnetic field is set up or when a change is effected in the

†Built by the Société Française d'Équipement pour la Navigation Aérienne, Neuilly-sur-Seine.
reluctance of the magnetic circuit linking the coil. It should be noted that, for the two types of instrument just described, an emf exists across the terminals only during the movement in the first case and only during the flux variation in the second case. Following are more details on the main types of electromagnetic sensing devices used in servo application.

![Diagram](Image)

**Fig. 29-37. Micromoteur SFENA type 11.01.**

1. **Magnetostriction-type Sensing Devices.** If a magnetic metal bar is placed in a magnetic field, a flux density \( B \) will be produced in it. Let the bar be of length \( l \) and cross section \( s \). When \( B \) is made to vary by an amount \( dB \), a force \( dF \) is produced; it tends to stretch or compress the bar, or \( dF = Ms \ dB \), where \( M \) is the coefficient of magnetostriction. On the other hand, maintaining constant the mmf across the bar, should the bar be compressed or stretched to produce an incremental \( dl \), a flux-density variation will be produced, or \( dB = MH \ dl/l \). The corresponding flux variation is

\[
\frac{d\Phi}{dt} = \frac{d(nsB)}{dt} = \frac{nsMH}{l} \frac{dl}{dt}.
\]

where \( n \) is the number of turns. This flux variation generates a voltage which becomes the output of the sensing device.

Practically, this technique is commonly used for marine soundings at supersonic frequencies, as also are quartz crystals. Further applications in other fields can easily be foreseen.

2. **Electrostatic Types.** As mentioned earlier in this chapter, the capacitor is also used as a sensing device. In the application discussed, capacitance variations were used as a measure of physical displacements while excluding entirely any power consideration.

In this case, on the contrary, only one plate of the capacitor is movable. Its mass is \( m \), its elasticity is \( k \), and its damping factor is \( f \). When this capacitor is connected in series with a resistance \( R \) across a voltage source \( E_o \), it eventually becomes charged.
to the full voltage $E_0$. The current will have ceased and the voltage across $R$ will then be zero. From these given conditions, there is no difficulty in establishing the equations of the system.

2. Pieoelectrical Instruments. Certain crystals, the most common of which are quartz and double tartrate of sodium or potassium, exhibit the following property: If, suitably cut, they are subjected to mechanical stresses, potential differences appear between certain faces.

Consider, for example, the quartz crystal in its natural state: It is shaped as a hexagonal prism with the ends in the form of hexagonal pyramids, leading to the belief that quartz crystallizes in accordance with a hexagonal system. In reality, the appearance of this natural crystal is only one aspect of a rhombohedral truncated crystal. It has a ternary axis of symmetry called the optical axis and three binary axes called electrical axes. When such a crystal is cut at right angles to its electrical axis, so as to have two parallel surfaces, electrical charges of opposite signs will develop on the latter if mechanical stresses are applied on the same surfaces—that is, applied parallel to the optical axis. The charge developed is independent of the quantity of quartz involved, being expressed by the equation $q = kF$, where $F$ is in dynes, $k = 6.45 \times 10^{-8}$, and $q$ is in electrostatic units. Suppose the crystal is to be cut to the shape of a rectangular parallelepiped, two faces being perpendicular to the electrical axis and two others parallel to the optical axis. When mechanical stresses are applied to opposite faces, $q = kF/l/e$, where $k$ and $F$ are as defined above, $l$ is the crystal dimension in the direction of the stress, and $e$ is the dimension parallel to the optical axis.

The phenomena are reversible in that, if a charge is applied to the sides perpendicular to the electrical axis, the crystal will expand or contract, depending on the polarity of the charge. The same applies to tourmaline crystals, the scarcity of which accounts for the more common use of quartz. Tartrate salts are much more sensitive than quartz, but their crystals are delicate and hygroscopic. They are used in situations where their drawbacks are of no importance.

The pieoelectric properties of quartz afford numerous applications. In addition to the basic phenomenon, quartz exhibits the characteristics of a very high Q circuit when used in oscillators, and it can serve as a basis for very-narrow-band filter circuits. The most common application of quartz crystals in servos is as pressure-sensing devices, either static or dynamic, which may be used equally well for the detection of gas pressure, weight, or acceleration.

G. PRESSURE MEASUREMENTS USING QUARTZ CRYSTALS. Since the charges developed by a quartz crystal under pressure are rigorously proportional to pressure, it follows that use can be made of this property for pressure measurements. Even in the case of rapid pressure variations, when other pressure gauges would be unable to follow because they would be limited by their inertia, measurements are possible by means of quartz crystals.

Measuring the charges across the crystal has, however, been a problem. It is possible to measure slowly varying charges by using a standard electrometer, but this instrument cannot follow rapid changes; only the advent of the electrometer vacuum tube has made such measurements possible. The vacuum-tube instrument is similar to a regular triode except for the fact that it has a very high grid-cathode, or input, resistance, $10^{18}$ ohms as compared to $10^8$ ohms otherwise. The importance of this resistance is evident. When measurements are to be made on phenomena of periods several seconds in duration, the total time constant of the combination crystal and electrometer tube must necessarily be greater than the periods. If the shunt
capacitance of the crystal is of the order of 100 $\mu$F, the isolation resistance must be large—say, $10^{11}$ ohms—if a time constant of 1 min is desired.

Electrometer tubes draw only a few microamperes of current; their plate resistance is high, with an amplification factor of approximately unity. Certain regular tubes such as the European miniature triode type 4672 are usable for electrometer service. Although this tube is of an all-glass construction that results in high insulation resistance, it cannot be used without some precautions to protect this insulation: It must be clean and dry, and the use of a drying agent is recommended. The electrometer tube should always be located close to the crystal in order to avoid losses through leakage in the wiring and interference through coupling into its high impedance.

Quartz accelerometers are constructed along these same lines; they will be discussed in Sec. 29.7.

b. QUARTZ CRYSTALS IN OSCILLATORS. Consider a properly cut crystal inserted in a series circuit comprising a voltage $E$, a resistance $R$, and an inductance $L$. The holder has a capacity $C$, a mass $m$, a damping coefficient $f$. If its stiffness is $k$ and the force applied to the crystal is $F$, then, on the one hand,

$$F = m \frac{d^2x}{dt^2} + f \frac{dx}{dt} + kx + Aq$$

where $Aq$ represents the piezoelectric effect and $x$ the crystal elongation. On the other hand, in the electrical circuit the following holds:

$$E = L \frac{dq}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} + Ax$$

The term $Ax$ corresponds to the piezoelectric effect. When $E$ is zero, the crystal acts as a receiver of mechanical effort; this would be the case if it were subjected to sound pressures, i.e., acting as a microphone. When $F$ is zero, the crystal supplies mechanical effort. It could develop sound pressures, hence serve as a loudspeaker. For many purposes, quartz is used at supersonic frequencies. One application is in marine sounding.

When a coupling term is added to the above equations so as to allow the system to oscillate, an electromechanical oscillation is obtained. This is accomplished by combining with the crystal an amplifier loaded with positive feedback. The frequency stability of the combination is very high. The wavelength of the frequency of oscillation of a quartz slab is approximately 120 m/mm of crystal thickness. This dimension is to be taken between the two faces perpendicular to the electrical axis.

c. PHOTOELECTRIC CELLS. One method of using photoelectric cells as sensing devices is by sensing and measuring with them a definite quantity of light and applying the result to the operation of some device. Another application is in determining the position of mobile objects, where the information is obtained by means of a combination of two cells plus a mobile masking plate attached to the object in motion.

d. THERMOELECTRIC DEVICES. Their operation is based on the thermocouple. Consider the theoretical case of two conductors, different in nature, connected in series and joined at both ends to form a closed circuit. If the two junctions are not at the same temperature, there will be an emf present in the circuit, and hence a circulating current. Electronically, this might be explained by saying that the free electrons of both materials or metals undergo thermal agitation. At any given time, the number of free electrons traversing the contact surface of one metal
toward that of the other differs from the number traversing in the opposite direction; this results in a difference of potential between the two metals, and the voltage is called a thermoelectric voltage. The polarity of this voltage will depend on the metals used. If the colder junction of the two is maintained at 0°C, the emf present may be expressed as a third-power function of the temperature of the warmer junction:

\[ E = at + bt^2 + ct^3 \]

For small differences of temperature, two terms of the equation, or even one, will suffice. When the cold junction is not at 0°C, the emf present may be determined by calculating for each junction as referred to 0°C and taking the difference as the effective value. For metals used in common thermocouples the values of the coefficients \( a \), \( b \), and \( c \) are given in Table 29-1. In the table, \( t \) is the number of degrees above 0°C of the hot junction; \( a \), \( b \), and \( c \) are the coefficients defined above; and \( E \) is the thermoelectric voltage in microvolts, considered as positive when the current flows through the cold junction from the first metal toward the second.

The purity of the metals used for thermocouples greatly affects their performance. This is evident from the fact that impurities add their own thermal effects, thereby altering the observed voltages. Lead is the least affected, since its structure is not altered by processing and it is not subject to hardening. This is why thermoelectric emfs are almost always referred to this metal.

The thermoelectric phenomenon just described is associated with another effect supported by electronic theory, the Peltier effect. A current flowing through a thermocouple junction produces or absorbs heat depending on its direction and on the metals involved; here the same thermal effect takes place but in opposite direction.

Pure metals are not used for thermocouples; alloys or combinations of pure metal and alloy are commonly used in practice, and the following are currently used: platinum–platinum rhodium, platinum–platinum iridium, nickel chrome–nickel aluminum, iron–constantan, copper–constantan, and nickel chrome–constantan.

It is possible to use a thermocouple in a temperature regulator, one method being to connect it in opposition to a variable but stable voltage. The tendency at present, however, is to use thermometric resistance-type devices.
Thermocouples are in general use for temperature measurements. It is possible to make them so small that they will have negligible thermal effect at the point where a reading is taken. A d-c instrument graduated in degrees is used as indicator; it can be switched for successive readings across a large number of associated couples.

Thermocouples are in general use also for the measurement of alternating current at high frequencies, say up to 50 megacycles. In this application a resistance wire, or heater, intended to carry the current to be measured, is mounted close to or in contact with the thermocouple. The temperature rise will be proportional to the square of the heater current. Currents as small as a few milliamperes can be thus measured; however, all these instruments are delicate and cannot withstand much overload.

29.4.6. Pulse-width-modulated Sensing Devices. One class of sensing device transforms a mechanical displacement into electrical pulses, whose

![Diagram](image)

Fig. 29-38.

![Diagram](image)

Fig. 29-39.

width is a function of the amplitude of the input displacement. The repetition frequency must evidently be higher than the highest frequency present in the physical phenomenon under analysis. These devices are of considerable interest, especially in nonlinear "on-off" systems or in sampling systems. Their principle of operation is elementary. One such device consists of a constant-angular-speed cylinder, the surface of which is divided into two definite sections, one conducting and one nonconducting (Fig. 29-38). The transducer transforms the input variable into a displacement of the sliding rod $AB$. The width of the voltage pulses depends on the position of slider $B$ along the cylinder axis.

Note that the nonlinearity of the transducer can be corrected easily by shaping the contour of the conducting surface on the cylinder. Figure 29-39 represents two involutes of the cylinder surface: in the first case, the pulse leading-edge period is constant; while in the second case, the pulse-center period is constant.
29.5. VACUUM TUBES AS SENSING DEVICES

29.5.1. Principle. Vacuum tubes can be controlled in two ways: electrically and mechanically. The first method consists in the control of the electron flow by means of an electric or a magnetic field. This offers the possibility of easily detecting electric or magnetic fields. A system where the desired result would be the control of the field of a large electromagnet could very well be based on this principle.

The characteristics of mechanical control make the second method readily suitable for many sensing devices. Mechanical control is generally obtained by plate-cathode spacing variation in a diode. When the tube current is space-charge-limited, this current can be expressed by

\[ i = C_0 S \frac{V^{\frac{3}{2}}}{x^2} \]

where
- \( i \) = current, amp
- \( C_0 \) = Child's constant
- \( S \) = cathode surface
- \( V \) = plate voltage
- \( x \) = plate-cathode separation

These tubes are still under development.

29.5.2. Developments. An electronic compass has been devised; its operation is based on the deviational effects of the earth's magnetic field on an electronic stream.

Many types of sensing devices incorporate electronic tubes whose geometry is variable. The plates of these tubes are generally the movable elements (Fig. 29-41). Characteristic curves for tubes of this type, manufactured by la Société LEGPA, are shown in Fig. 29-42. The
relatively high value of plate current available eliminates the need for a preamplifier. It should be noted that coulomb friction is completely absent in this device, the only friction being in the "flexing" of the movable-element supports.

29.6. GYROSCOPIC SENSING DEVICES

29.6.1. Angular-rate Sensing. In this case, one must consider the angular velocity of a system with respect to fixed space. This is a frequently required condition that occurs both in aeronautical and naval applications. Several airplane and guided-missile autopilots use angular-velocity sensing devices in order to either maintain a constant pitch or supply guidance information. Use is made of the gyroscopic effect (Fig. 23-43).

Taking $H$ as the kinetic moment of the rotor and $\Omega$ as the angular-velocity vector of the gimbal, the reaction torque on the gimbal is

$$C = |H \times \Omega| = H \cdot \Omega \sin \theta$$

where $\theta$ is the angle between the two vectors.

If the angular momentum of the rotor is constant, the torque is proportional to $\Omega$, provided $\theta$ is constant. In fact, if it is desired to convert the torque into terms of displacement, and this can be easily done by means of a spring (Fig. 29-44), the angle $\theta$ is no longer constant and the
deflection is no longer proportional to the torque. One can restrict the reading to a component $\Omega_z$ of the vector $\Omega$ on the initial plane of the rotor. In this case, the scale will no longer be linear, but it will remain single-valued whatever the rotation vector may be. This nonlinearity\(^1\) can be compensated for by means of a nonlinear transducer built for this purpose (e.g., a potentiometer). The transfer function of the rate gyro first described is

$$\frac{\Theta}{\Omega} = \frac{H}{B_1s^2 + b_1s + k}$$

where $b_1$ and $k$ are the friction coefficient and the spring stiffness, respectively.

Note that the inertia of the gimbal, including the rotor, with respect to $Oy$ appears in this expression. The performance of the rate gyro will therefore be improved if the gimbal inertia $B_1$ with respect to $Oy$ is made small.

29.6.2. Deflectionless Rate Gyro. Efforts have been made to decrease rate-gyro deflections. At the present time good instruments have deflections that are less than 5°. It must be noted that, in the case of potentiometer detection, the potentiometer range is small, approximately 5 mm. Obtaining saturation-to-threshold ratios of 100 and 1,000 thus becomes extremely difficult. Use is often made of variable-inductance transmitters, which leads to the use of an auxiliary local oscillator. Another approach for assuring linear operation consists in using a variable restoring torque in such a way that the rate gyro remains at its zero position. Under such conditions the torque developed is exactly proportional to the angular velocity that is to be detected.

Adjustment of this null device is critical. Actually, it implies the existence of a closed loop inside the sensing stage. Nevertheless, a certain deflection is always present, and this constitutes the error of the servo system. For this condition, the feedback arrangement must be electrical. The diagram shown in Fig. 29-45 represents a possible method

\(^1\) In the next paragraph a type of rate gyro that does not involve this nonlinearity will be described.
of employing this type of rate gyro. The block diagram is shown in Fig. 29-46.

Rate gyros of the type just described were used as early as 1944 in the V-2 rocket. They are still being studied today, but their future employment appears to be limited mainly to that of sensing angular velocities which are subjected to considerable variations.

![Fig. 29-45. Gyro with electrical restoring torque.](image)

Fig. 29-45. Gyro with electrical restoring torque.

![Fig. 29-46.](image)

29.6.3. Choice of Parameters. A rate gyro is characterized by two parameters, namely, the friction coefficient of the dashpot and the restoring coefficient. A dashpot is generally necessary, although some rate gyros designed for measuring purposes have none. Air dashpots are most frequently used (Fig. 29-47), adjustment being made by means of a pointed screw. The restoring coefficient corresponds to the spring, or to additional resistances if magnetic restraint is used as shown in Fig. 29-48.

The choice of the two parameters depends upon the application for which the instrument is being considered. An autopilot rate gyro is not adjusted in the same way as a measuring rate gyro. In the case of an autopilot, the parameters are chosen in such a manner as to be compatible with the airplane dynamics. The above second-order transfer function represents a good first approximation of the instrument transfer function. It should be remembered that, the smaller the damping coefficient, the more pronounced is the resonance, which fact could cause serious trouble in the sensing stage.

As far as the noise level is concerned, important improvements have been made in recent years. The present trend is to utilize sleeve bearings,
whereas ball bearings were in favor in the early 1950s. This is a result of the present requirements for higher-precision gyroscopes and rate gyros, which do not permit the noise level introduced by ball bearings.

![Diagram](image)

**Fig. 29-48.**

### 29.7. ACCELEROMETERS

**29.7.1. The Problem.** For reasons of stability it is necessary to detect not only angular velocity but also angular acceleration. A composite signal due to these two corresponding quantities, and often a third quantity proportional to angular position, is fed to the amplifying system preceding the control motor. Acceleration or angular-position sensing devices would not be necessary if the angular-velocity unit gave perfect results. However, differentiations or integrations of its output signal are often unacceptable on account of thresholds, saturation, and inherent noises. It is possible to obtain a much purer signal when the sensing element responds directly to angular acceleration and the transducer is matched to the receiver.

In principle, the angular accelerometer is the simplest of these devices because it involves only the indication of the torque applied to a dynamically balanced body. The difficulties lie in the transducer, because no measuring of a force is possible without a certain displacement. To compensate the acceleration torque, a counter torque can be applied through a spring, resulting in a system of the second order involving a resonant frequency.

Another method consists in applying a variable magnetic counter-torque, which will necessitate an internal servo that may have a resonant frequency of its own.

**29.7.2. Applications.** Practical designs vary considerably with the range and application of the instrument. An accelerometer for measuring small accelerations of the order of 1 rad/sec² with a threshold value inferior to 0.1 rad/sec² and a saturation level above 10 rad/sec² must be so designed as to reduce coulomb friction to a minimum. A perfectly balanced mass is selected and suspended axially on a pair of crossed spring rods. In this manner, correct balance as well as countertorque is obtained. This type of instrument is always viscously damped (air damping usually of the "Curie" type) in order to avoid coulomb friction.
The coupling between the receiving unit and the transducer is obtained in an interesting way: the restraining force is applied through a strain-gauge wire; acceleration quantities are consequently transformed into resistance changes. It is common practice to use two strain gauges—or some other even number of them—thus avoiding temperature effects as much as possible (Fig. 29-49). In the range of 5 to 500 rad/sec², the instrument uses a cylindrical mass supported by two small rods or torsional wires (Fig. 29-50). It then satisfies the frequency-response conditions, while the damping becomes the critical problem. Eddy currents in the cylinder are sometimes used for damping purposes. Unfortunately, this damping does not correspond very accurately to pure viscous damping.

As mentioned earlier, quartz can also be used as sensing element and transducer. This arrangement consists of a quartz crystal and an electrometer tube placed in an enclosure which is rigidly fixed to the body whose acceleration is to be measured. Care must be taken to avoid shocks to, or intense vibrations of, the electrometer tube. The natural frequency of the assembly must be made sufficiently different from that of the phenomena being measured. A very wide bandpass will be required, particularly when the measurement of acceleration due to shock is required. These accelerations may be very high, in the range of 50 to 500 g, and of very short duration or sharp waveform. Through design it should be made impossible for shock to excite the assembly into resonance at its natural frequency. This instrument is more complicated than the strain-gauge accelerometer, but its natural frequency is much
higher. However, this type of instrument should be avoided as much as possible in airborne equipment because of the fragility of the electrometer and its accessories.

29.8. CHOICE OF A SENSING DEVICE

The proper choice of a sensing device is an important problem; it determines the initial value of the quantity of information sensed, which quantity is itself inferior to the quantity to be transmitted. Moreover, this information can only diminish as it progresses further along the system.

There is no general rule for the choice of a sensing device, any more than for the choice of a driving unit. The following should, however, be kept in mind:

1. Measurement range desired.
2. Required accuracy: the sensitivity of the device should be at least equal to the largest admissible error.
3. Required bandwidth: the range of frequencies that the device will be required to pass.
4. Probable power drain on the driving unit by the sensing device. This is important. The smaller the power requirement from a sensing device, the greater will be its linearity. A generator develops a voltage proportional to its speed, but its terminal voltage varies with the current it supplies. This reaction of the load on its power source occurs generally in all applications. The important criterion is that the ratio of extracted energy to available energy be small. When a 100-watt generator supplies 0.1 watt, the loading on the generator can be considered negligible.

Unfortunately, it is not always possible to use a powerful generator in small servos. The inertia of its rotor could be an important part of the inertia of the total system, thereby impeding or even preventing correct system performance.

5. Noise level and its frequency and energy distribution. The effects of high noise level originating in sensing devices may be avoided in many cases. When this noise covers a range of frequencies outside the useful frequency band, it can be filtered out without quantitatively affecting the transmitted information. Such a condition was discussed earlier when commutator-type dynamos, for use in regulators, were studied. In that case, the noise frequency extended from 0 to 50 cps, while the fundamental frequency of the noise was 1,250 cps. Evidently, filtering is no problem in such a case.

It is worth reiterating that the sensing device should not be considered alone and isolated from the rest of the chain. Poirson tachometer dynamos have a low noise level compared with that of standard commutator dynamos, but their output voltage is also very low and must be amplified. The amplifier will add distortion, time constants, and noise. A fair comparison can be made only at equal power levels, that is, in this case, after the output of the Poirson unit has been increased by amplification to the output level of the other machines.
It is impossible to select the units of a servo system correctly if each unit is evaluated independently of the others. Thus, very simple applications on a theoretical basis often conceal stumbling blocks difficult to master with material units and their limitations. It is often more important to study the limitations than the basic principles. In servo applications, noise levels and impedance matching within the chain are determining and limiting factors in the choice of the elements. It is of prime importance always to bear in mind the system as a whole.
CHAPTER 30

DETERMINATION AND DESIGN OF SERVOMOTORS

Summary
1. The problem.
2. Discussion of inertias and gear ratios.
3. Determining the transfer function of a given motor from the characteristic curves of the motor.
4. Points of comparison between various types of motor.
5. Adjustment of the characteristics and design calculations for a motor which must comply with given conditions.
6. Particular problems.

30.1. THE PROBLEM

The behavior and realization of the driving element (motor) of a servomechanism will be discussed in detail. It is sometimes possible to select readily the proper motor from those that are commercially available. It is, however, unusual to find one that is exactly adapted to the problem under consideration, but this possibility should not be overlooked a priori.

30.1.1. The Given Conditions. It will be assumed during the preliminary evaluation that the performance that must be satisfied by the output element has been estimated as follows:

a. When driven by a sinusoidal input, the output element should be able to oscillate. The amplitude usually varies with the frequency. For example, one may have to investigate the realization of a servomotor that must be capable of driving an aircraft control surface with an amplitude of 6° up to a frequency of 5 cps.

b. Besides condition a, which is purely energetic, the transfer function is specified, either by the amplitude-frequency relationship, for which minimum and maximum values are limited in a given frequency band, or, better, by the phase-frequency relationship, for which maximum phase shift at a given frequency is specified.

c. The required accuracy in the steady-state regimes. It is difficult during the preliminary stage of the project to specify more details concerning the expected performance of the driving element than was done in (a) and (b). Furthermore, when the motor alone is considered—i.e., independently of the whole loop—only imprecise data are furnished.

The above sparse data are usually the only basis that is available for the selection and adoption of the driving element of the servo. The data should first be interpreted and then submitted to the sponsor for discussion. It is generally assumed that the motor should be designed in accordance
with the specifications of the preliminary study. There frequently
exist secondary loops within the principal loop, and the addition of a
passive network to compensate the motor should be made in the second-
ary path of that motor, i.e., at the high energy level, which is not recom-
mended. It is possible, however, to think of systems where the motor is
in the main forward path and where there is provision for a tachometric
feedback. It will be assumed that such a case has been considered in the
preliminary project.

30.1.2. Motors. The transfer-function concept leads to the considera-
tion of motors as force or torque generators. When evaluating the
transfer function of a motor, an equation of the following type is written

\[ \text{Transform of force, or torque } = F(s) \times \text{transform of input signal} \]

where \( F(s) \) is a rational operator in \( s \). Position, rate, acceleration, etc.
are the result of this force, or torque, on the mechanical system; that is,
the \textit{motion} is the consequence of the applied forces, or torques.

The more detailed concept of transfer matrix does not separate the
two components of a signal, such as voltage-current, force-speed, and
pressure-mass flow, and the driving motor is no longer a simple force, or
torque, generator, and hence no longer a motion generator. However,
for any \textit{linear} system (Sec. 10.1) the ratio of the two components of the
input signal (input impedance) and the corresponding ratio for the output
signal (output impedance) are related by a homographical relation
depending only on the elements of the system. This fact is of primary
importance.

Finally, it is to be noted that, for the two possible electrical analogies
for mechanical systems\(^1\) \( V \sim F \) and \( V \sim v \), the latter is more practical
whenever electromechanical coupling effects are to be encountered. The
\textit{mechanical impedance} is defined, therefore, as the \textit{ratio of speed to force},
or \textit{torque}.

A motor can be considered, then, as a \textit{motion generator having a known
internal impedance}, in the same way as an electrical generator is a source
of voltage with a known internal impedance (see Fig. 2-5).

30.1.3. Industrial Motors. Such motors are used in industry, railroads,
etc. They can be of many types: internal-combustion, electrically
coupled diesel engines, hydraulic motors, and steam or compressed-gas
engines. In general, these motors are characterized by their design
\textit{power} for design rpm, although it is also necessary to know the torque-
speed characteristic in the vicinity of the design velocity.\(^2\) However, for
motors used on power vehicles, it is necessary to know as accurately as
possible the torque-velocity plot for any rpm. So far as this torque-
velocity characteristic is concerned, it is to be noted that: (1) the product
of these two quantities is the power of the motor and (2) the exact
position of the output shaft is of little interest. For example, the torque

\(^1\) See Sec. 2.4.

\(^2\) Although from now on only rotating motors will be considered, the considerations
are still valid for linear actuators.
developed in the armature of an electric motor does not depend on its position, at least not to a first approximation. In fact, the torque is periodic and varies only slightly about a mean value, because of the collector blades and slots of an electric motor, the dead points of the pistons of a hydraulic motor, etc.

Finally, knowledge of the characteristic at the design supply voltage is sufficient except for railway motors or, more generally, for any motor starting under load. It should be noticed, however, that such sophisticated devices as proportional clutches and shifted gears are provided for the use of motors having unsatisfactory performance at low rpm.

The present trend, however, is to consider systematically not only the characteristics of a motor but also the transient response. It is known, for example, that electric motors must be started with a voltage that progressively increases up to the design value. Except for small motors, this method is generally used—even for unloaded motors, because of the energy required to accelerate the rotor itself. This does not apply to hydraulic motors because the available torque-to-inertia ratio of the rotor is very high.

**30.1.4. Servomotors.** For such motors, the transient response is more important than the steady-state performance at design rpm. It is possible that the latter may not even be defined. While the torque-velocity curve in the vicinity of the design angular speed sufficiently characterizes a conventional motor, it is not so with a servomotor. It is necessary to know, instead: (1) the torque-velocity curve from zero to the maximum permissible angular velocity, (2) the complete set of such curves for the various values of the command signal (voltage or current, fluid flow, etc.) and (3) the characteristics of the motor under transient conditions, for example, electrical data on the field circuit of a field-controlled electric motor or the characteristics of the controlling valve of a hydraulic motor. It is to be noted that these data are sufficient to determine the transfer function of the system but not to solve the problem fully by the use of the transfer matrix.

Furthermore, the inertia of the rotating elements becomes essential. The driving motor of a servo system operates under continuously varying conditions, and a certain amount of energy is required for accelerating the rotating parts. Consequently, even if no energy is dissipated at the output—for example, when the output consists of an inertia load and a pure spring torque without any viscous friction—the servomotor has to supply some energy.

It is possible to show that the design of a servomotor is based on its capability to absorb mechanical energy as well as to generate it. It is desirable that the motor when used as a brake will have a transient performance that is just as good as that when it is used as a driving motor.

In a servo system there is always fluctuating energy (Sec. 12.3.6); the larger the inertia of the system, the greater the energy involved. Theoretically, this energy is not dissipated, but the energy that is produced by the deceleration of a rotor cannot easily be recovered. Conversely, the external power supply must deliver the energy for the acceleration, and
its capacity may not be sufficient, or more precisely, its internal impedance may be too high. Servomotors are always designed so that the inertias of the rotating parts are as low as possible.

From the viewpoint of a servomechanism, it is necessary to know: (1) the complete set of torque-velocity characteristics for all possible values of the command signal and (2) the mechanical characteristics of the rotor, that is, its inertia and, possibly, friction coefficient. For hydraulic motors, the inertia or friction coefficient should be referred to the shaft, since some elements are rotating and some have a reciprocating motion.

The knowledge of the family of characteristic curves and of the data mentioned in (2) is sufficient, but occasionally knowledge of the characteristic curves is not necessary. Finally, it should be pointed out that the laws of mechanics, \( F = ma \); of electrodynamics, \( \tau = Ki \) (torque proportional to current); and of hydraulics, \( F = S.P. \) (force proportional to pressure) are relationships between instantaneous quantities. For example, to a current \( i \) there corresponds at any time—that is, without any time lag—a torque \( C \).

30.1.5. Summary of the Chapter. First, from a simple discussion of the maximum acceleration of the output shaft, some fundamental rules will be established. As a first approximation, they will enable a determination of the proper gear ratio when the load or the motor is specified, or when both are specified. It will be necessary, however, to proceed further and to determine the transfer function of the motor. Finally, to comply with all the requirements of the preliminary design, the characteristic curves of the motor should be appropriate.

The adaptation of a motor to a servomechanism is not a problem requiring a unique solution. The considerations listed below are simply the presentation of two methods that the authors have personally used. It is the purpose of Sec. 30.5 to explain these two methods in detail. The following should not be considered as the only methods applicable to the design of a servomotor. There is no one general method, and in this chapter only some valuable approaches to the problem are given as assistance to the designer.

30.2. DISCUSSION OF INERTIAS AND GEAR RATIOS

30.2.1. Elementary Cases: Selection of the Gear Ratio.\(^1\) The problem of the inertia of the rotor or of the optimum inertia of the load to be fully adapted to the motor has been the occasion of many errors. These originate to a large extent from improper terms and assumptions. The two cases that will be considered in what follows are:

1. The load, with inertia and friction, has been specified, and the motor has been selected from those commercially available. The motor inertia is then given and the problem is to select the optimum gear ratio based on the proper criterion.

\(^1\) Only rotational motors are considered here; results can easily be extended to linear motion motors.
2. The load alone has been specified and the motor and the gears must be designed. Although the motor cannot be separated from the load, it is possible, as a first approximation, to design it by disregarding the elements preceding it in the forward path.

The case of an unspecified inertia for the load is of no interest. If, to some extent, it is possible to adjust the inertia of the load, it is obvious that this inertia should be made a minimum. In the same way, it can be shown that the best motor for a given available torque is the one having minimum inertia.

It can be stated that inertia is the curse of the servo designer, wherever it is located (driving element, gearing, or load).

To design a motor first, neglecting the load, and then adjust the gear ratio is not acceptable. It is necessary to introduce the gear ratio into the over-all system of motor, reduction gear, load, etc. (Fig. 30-1).

![Diagram](https://via.placeholder.com/150)

**Fig. 30-1.**

A very simple case, in which the motor inertia \( J_m \) and torque \( C_m \) and the load inertia \( J_o \) are known, will first be considered. Referring to the output-shaft position \( \theta_r \) and terming \( \alpha \) the gear ratio, the following equations may be written:

\[
\alpha = \frac{\theta_m}{\theta_r}, \quad \frac{d^2 \theta_r}{dt^2} = \frac{C_m}{J_m \alpha + J_o / \alpha}
\]

The acceleration of the load is, therefore, maximum for \( \alpha = (J_o / J_m)^{1/2} \), that maximum being

\[
\left( \frac{d^2 \theta_r}{dt^2} \right)_{\text{max}} = \frac{C_m}{2(J_m J_o)^{1/2}}
\]

Because the reduction-gear ratio so determined is independent of the torque supplied, this adaptation is valid for any motor, whatever the shape of its torque-velocity characteristic. This does not mean, however, that for a given inertia of the load \( J_o \), any combination of a \( J_m \) and a gear ratio \( \alpha \) so chosen that the inertia of the load referred to the driving shaft is equal to that of the motor will be optimum. In fact, two cases may arise:

- a. It is assumed that the inertia \( J_o \) is given. By defining

\[
\beta = \frac{J_m}{J_o}
\]

the acceleration can be written as

\[
\frac{d^2 \theta_r}{dt^2} = \frac{1}{\alpha \beta + 1/\alpha} \gamma^2
\]
where $\gamma_1$ is the angular acceleration of the load alone driven by the motor torque $C_m$ ($\gamma_1 = C_m/J_o$).

The actual acceleration is a maximum when $\alpha = 1/\beta^{1/4}$, the value of that maximum being

$$\left(\frac{d^2\theta_r}{dt^2}\right)_{\text{max}} = \frac{1}{2\beta^{1/2}} \gamma_1$$

The locus of the maxima is a straight line (Fig. 30-2).

\[ \frac{d^2\theta_r}{dt^2} \frac{\sqrt{1}}{2\sqrt{\beta}} = \alpha \]

**Fig. 30-2.**

Consequently, the smaller $\beta$ (i.e., the smaller the inertia of the motor), the larger the angular acceleration.

b. The inertia $J_m$ is given. By defining $\beta' = J_o/J_m$, the acceleration of the output shaft can be written as

$$\frac{d^2\theta_r}{dt^2} = \frac{1}{\alpha + \beta' \gamma_1}$$

where $\gamma_1$ is the acceleration of the motor alone when the torque $C_m$ is applied ($\gamma_1 = C_m/J_m$). The acceleration is a maximum when $\alpha = \beta'^{1/4}$, the value of the maximum being

$$\left(\frac{d^2\theta_r}{dt^2}\right)_{\text{max}} = \frac{1}{2\beta'^{1/2}} \gamma_1$$

The locus of the maxima is a hyperbola (Fig. 30-3). This hyperbola has been plotted in Fig. 30-3, where also is plotted the parabola defined by the equation

$$\alpha = \beta'^{1/2}$$

which relates the values of $\alpha$ to any given value of $\beta'$. The smaller $\beta'$ (i.e., the smaller the inertia of the load), the higher the acceleration.

In conclusion, for the usual case of a load with a specified inertia $J_o$, the motor yielding the maximum acceleration is to be so selected that $C_m/J_m^{1/4}$ is a maximum. Other features, however, such as fluctuation of the torque at low speed and maximum velocity of the motor, are also to be taken into account when selecting the gear ratio. It may happen that the optimum value so calculated is not feasible. An example is the servo motor of an autopilot driving the elevator of an aircraft. The inertia of the control surface is taken as 5,000 cgs. For many reasons, especially that of safety, the rate of displacement of the elevator is limited to 90°/sec. In addition, the servomotor is located just behind the control surface to
eliminate the unwanted effects of the elasticity and backlash of the control rods. The small space available leads to a small motor with high angular speed. Usually motors of about 6,000 rpm maximum speed are selected, leading to a reduction gear ratio of 400/1.

This problem will be discussed in detail in Sec. 30.5. It is to be noted that to increase the inertia of the elevator in order to approach some "optimum matching" would be absurd.

Note. It is very often stated that the exchange of energy between motor and load is optimum when the gear ratio is \( (J_e/J_m)^{1/2} \). Stated in this manner, the conclusion is not correct. The accelerating power transferred to the load is

\[
P_1 = J_e \frac{d\theta_e}{dt} \frac{d^2\theta_e}{dt^2} = J_e \frac{C_m \Omega_m}{J_m \alpha^2 + J_e}
\]

The exchange coefficient is maximum for \( J_m \alpha^2 + J_e \) minimum, i.e., for \( \alpha \) as small as possible and not for \( \alpha = (J_e/J_m)^{1/2} \). For the latter, the transferred energy is \( P_1/2 \).

30.2.2. Efficiency of Servomechanisms. The concept of efficiency is not widely used in servomechanisms; however, two types of efficiencies can be defined. The first is the so-called virtual efficiency. It is the inverse of the ratio of the energy actually required by the servo to that theoretically required to fulfill the two following conditions:

a. Move the mechanical elements according to the input function; i.e., continuously maintain a zero error.

b. Compensate for the disturbances when motion occurs.

This type of efficiency is quite conventional, since it depends not only on the input function but also on the type of disturbance encountered.
The second type is the actual efficiency. It is the inverse of the ratio of the total required energy to the energy required to fulfill these two conditions:

a. Move the mechanical elements according to the output function.
b. Eventually work against the disturbances if motion occurs.

This type of efficiency can be actually measured, because the energy is referred to the output variable, and the amount necessary to compensate for the disturbances can be computed.

It is also necessary to characterize the stationary consumption, that is, the energy dissipated with zero input and no disturbance. Many other definitions of efficiency can be considered. It is mandatory, however, to state clearly to what quantities the efficiency refers.

Little attention is paid to the efficiency concept in small servo systems. This is probably because the mechanical efficiencies of small motors of less than 100 or 200 watts are comparatively low, i.e., usually less than 0.5. Curves that are characteristic of d-c motors with separate excitation are shown by solid line in Fig. 30-4.

The quantity \( \Omega_M \) is the angular speed for zero efficiency, since there is no energy supplied while there is energy dissipation in the inductor. The quantity \( \Omega_0 \) is the optimum operating velocity corresponding to maximum power output. For this, \( \Omega_0 = \Omega_M / 2 \), where \( \Omega_0 \) and \( \Omega_M \) correspond to steady states. The torque at zero speed is \( C_m \) (stall).

For high-power servo systems, the situation is quite different; for a solution to the problem cannot be readily attained with a conventional motor, which usually has a very high efficiency at one, and sometimes two, velocities and an extremely sharp fall-off in efficiency at speeds away from the optimum operating speed(s). The performance that is required of a servo system would necessitate a very large decrease in the efficiency of the motor. As previously stated, the virtual efficiency depends on the input function. In other words, the spectrum of the input function should be added to the curve of Fig. 30-4.

Another aspect that is somewhat related to efficiency is the problem of irreversibilities in a servo system. Theoretically, any irreversibility in a servo loop destroys linearity. Actually, many servomechanisms having only one nonlinear element—for example, a screw gear—can be considered to be linear to a first approximation. In this respect, it can be mentioned that: (a) When motion is present, irreversibility can be considered to be a dissymmetry of the efficiency. This will be discussed later. (b) Even in a linear system, the reacting energy, i.e., the energy oscillating between the various elements of the servo system, is dissipated at a high rate and therefore cannot be recovered. This is due to two different reasons. The one is that of dissymmetrical efficiencies of a motor when it operates as a driving motor or as a brake. The other is the practical impossibility of feeding energy back to the generator through the amplifier, the energy actually being dissipated within the amplifier.

The following theorem pertaining to the problem of mechanical irreversibility can be found in a certain number of textbooks: "The efficiency of a nonreversible system is necessarily less than 0.5." To be more specific, the efficiency considered in the above-mentioned theorem is the ratio of the torque necessary to counteract without friction a force \( F \) to that necessary when friction occurs at the contact between the screw and the gear teeth (Fig. 30-5). Thus the efficiency just defined is a ratio of two torques, and not the ratio of two powers.
It thus cannot be concluded that, in a worm-gear reduction system (Fig. 30-5), 50 per cent or more of the input energy is dissipated. As a matter of fact, the friction angle \( \phi \) for lubricated surfaces depends on the lubrication and, to a large extent, on the relative velocities of the surfaces in contact. Furthermore, even if there is static nonreversibility, it may so happen that, when there is motion, the friction angle \( \phi \) has decreased to such a value that the angle no longer corresponds to static nonreversibility. In any case, the torques applied to the screw gear are *not symmetric* for a constant absolute torque delivered by the gear; rather, they depend on the sign of the transmitted rotation. Furthermore, the efficiency in terms of transmitted torques essentially varies with the operating conditions of the gear. Straight-toothed gears have equal efficiencies in both directions.

**30.2.3. Nonlinearities of the Power Stage and Random Responses.** Generally speaking, in the design of a power supply it may be of interest to consider the linear range and, whenever possible, the operating range of the supply. *Nonlinear systems*
can be usefully associated with probability considerations (random inputs). If the spectrum density of the random input function can be determined, it may be convenient to plot on the same diagram the linear range and the over-all operating range for a random input function of given spectrum. The data can be processed as follows: On a diagram (Fig. 30-6) the amplitudes of the input assumed to be sinusoidal are plotted as abscissas and the maximum amplitude of response of the system or element under consideration as ordinates.

Such a set of characteristic curves can be experimentally determined; they are usually similar to those plotted in Fig. 30-6 for the following reasons: If thresholds are disregarded, as long as the system behaves linearly the point moves along a straight line having a slope proportional to the system gain for the frequency under consideration. This family of straight lines can easily be constructed from the amplitude-frequency curve of the system. The lines are graduated in increasing values of \( \omega \), where \( \omega_0 < \omega_1 < \omega_2 \ldots \). The plots of Fig. 30-6 describe a system having an amplitude-frequency curve similar to that sketched in Fig. 30-7. At a given frequency (Fig. 30-6), saturation may occur for input amplitudes greater than a certain value and the curve will deflect downward. If mechanical stops or any equivalent devices are provided, the characteristic curve levels out. It may even happen that, at high values of \( \theta \), the characteristic curve has negative slope. If thresholds are to be considered, the curves should be limited on the left.

The linear range is defined by the contour inside which the characteristics are straight lines. A question that now arises is whether the point remains inside the contour when the input is a random function.

The power stage is usually the element having the narrowest bandwidth. On the other hand, the amplifier stage has an output with a bandwidth that is greater than that of the power stage, but the power that it can supply is limited, owing to either its own physical features or the characteristics of the preamplifier. As an example, a system which has a preamplifier followed by a small, separate, control motor is considered. The motor directly drives the control valve of a hydraulic actuator (Chap. 32), and feedback is achieved through connections between the output of the actuator and the control valve.

Because the power level of the electric motor driving the valve is low, a motor that is more powerful than necessary can be selected. The preamplifier-amplifier combination can be considered as linear in the anticipated frequency band, while the power stage is not. It is seen, then, that for a given frequency the maximum amplitude of the output signal is fixed. The manner in which this amplitude may be determined as a function of frequency for any given power supplied by the source will be shown in Sec. 30.5.

It is possible to obtain more information by adding to the \( \theta_r \)-vs.-\( \theta_c \) diagram a \( \theta_r \)-vs.-\( \omega \) diagram. Because of the limited power output, the point will certainly remain inside the nonshaded area. The contour of this domain can be determined either from experimental measurements on the amplifier and driving motor of the control valve or from the chart described in Sec. 30.5. The same procedure can be followed for other values of the output power, and it is then possible to draw the isopower curves \( \alpha P \), where \( \alpha < 1 \).

The spectrum of the random input function \( \Phi(\omega) \) previously considered can be superimposed on the same axes. To the angular frequency \( \omega \) there corresponds a
power which is proportional to $\Phi(\omega)$ and which may be directly evaluated from the above diagram. If this power is less than the power available from the motor, the corresponding amplitude of the actuator input can be evaluated and, consequently, the operating point $(\theta_\text{r}, \theta_\text{a})$ on the diagram determined. The linearity of the system can then be determined for any frequency of the input spectrum.

If the power determined from the spectrum diagram is higher than the maximum power available from the linear element, the corresponding amplitude will be given by the shaded curve, yielding a nonlinearity. Furthermore, the point on the $\theta_\text{r}$-vs-$\theta_\text{a}$ plot can correspond to a point located in the nonlinear domain.

In conclusion, if an element that is not necessarily linear is driven in the specified frequency band by a linear element, which is the general case, three plots, $\theta_\text{r}$ vs. $\theta_\text{a}$, $\theta_\text{r}$ vs. $\omega$, and $\Phi(\omega)$ vs. $\omega$, can be drawn. It may then be determined for any given random function whether the element under consideration operates linearly.

### 30.3. Determining the Transfer Function of a Motor from Its Characteristic Curves

#### 30.3.1. Conditions of Linearity of a Motor

A motor can be described by a linear transfer function only if its driving torque $C_m$ can be expressed linearly (i.e., as a partial-fraction expansion in terms of the Laplace variable) as a function of $\Omega_m$ and of the input variable $E$. If this can be done, and if $A(s)$ and $B(s)$ are rational functions, the velocity is

$$\Omega_m(s) = A(s) C_m(s) + B(s) E(s)$$

If only steady-state responses are considered, that is, $s = 0$, the equation becomes

$$\Omega_m = a C_m + b E$$

where $a$ and $b$ are constants:

$$a = A(0) = \frac{\partial \Omega_m}{\partial C_m} \quad b = B(0) = \frac{\partial \Omega_m}{\partial E}$$

The single-parameter family of curves $(C_m, \Omega_m)$ depending on $E$ is called the family of characteristic curves of the motor (Fig. 30-8). For a linear motor, i.e., one described by a linear transfer function, it consists of equidistant parallel straight lines having constant gradient with respect to the parameter $E$.

#### 30.3.2. Linearization in a Given Domain

Direct-current motors with field or armature control have, theoretically and practically, linear characteristics. On the other hand, two-phase a-c servomotors and hydraulic valve-controlled motors usually have nonlinear characteristics. In such cases it is possible to linearize these motors within a certain interval of velocity. In principle, the method consists in substituting a
tangent to the characteristic curve (Fig. 30-9). In general, the equation would have been

\[ d\Omega_m = \frac{\partial \Omega_m}{\partial C_m} dC_m + \frac{\partial \Omega_m}{\partial E} dE \]

In the linearized range, such as the crosshatched area of Fig. 30-10, it may be written as

\[ \Omega_m = \frac{\partial \Omega_m}{\partial C_m} C_m + \frac{\partial \Omega_m}{\partial E} E \]

where both \( \frac{\partial \Omega_m}{\partial C_m} \) and \( \frac{\partial \Omega_m}{\partial E} \) are constants. On the right-hand side of the equation, the first term represents the mechanical component of the system, and the second term the electrical component.

30.3.3. Determining the Transfer Function from the Characteristic Curves. The motor shaft position \( \theta_m \) is assumed to be related to the output shaft position through a reduction gear having a ratio \( \alpha = \theta_m/\theta_r \). Let \( J_m \) be the inertia of the motor and \( J_c \) the inertia of the load on the output shaft, the assumption being that the gear inertia is either negligible or included in \( J_m \) or \( J_c \). In addition, let \( f_c \) be the viscous-friction coefficient and \( k_c \) the spring-stiffness coefficient on the output shaft. The resistive torque applied to the motor shaft is, therefore,

\[ C = (Js^2 + fs + k)\Theta_m \]

where, for convenience (Sec. 10.3):

\[ J = J_m + \frac{J_c}{\alpha^2} \quad f = \frac{f_c}{\alpha^2} \quad k = \frac{k_c}{\alpha^2} \]

It is possible to eliminate the driving torque \( C_m \) between

\[ \Omega_m(s) = A(s)C_m(s) + B(s)E(s) \]

and the equation of dynamics \( C = C_m \) and to write the transfer function

\[ \frac{\Theta_m}{E}(s) = \frac{B(s)}{-kA(s) + [1 - fA(s)]s - JA(s)s^3} \]
Note. 1. The quantity \( A(0) = \frac{\partial \Omega_m}{\partial C_m} \) is usually negative in servomotors. This, however, is not a necessity.

2. If there is no spring stiffness—that is, \( k = 0 \)—it is seen that the transfer function implies an integration,

\[
\frac{\Theta_m}{E}(s) = \frac{B(s)}{s[1 - fA(s) - JA(s)s]}
\]

In general, the term \( B(s) \) is a constant, and \( \partial \Omega_m/\partial E \) and \( A(s) \) can be expanded as \( a_0 + a_1s + \cdots \). In general, the motor is so chosen that its natural frequency is higher than the maximum frequency to be passed by the servo. The \( a_0 \) term can very often be neglected and the velocity time constant can be expressed as

\[
\tau = J \left| \frac{\partial \Omega_m}{\partial C_m} \right|
\]

3. Up to now, it has been assumed that \( \partial \Omega_m/\partial C_m \) is not infinite. If the characteristic curves are straight lines parallel to the axis of abscissas, implying that the torque does not depend on velocity as is the case for torque motors, the above calculation must be revised. The initial equation can be written as

\[
E(s) = H(s)C_m(s) \quad \text{with} \quad \frac{1}{H(s)} = \frac{\partial C_m}{\partial E}(s)
\]

The transfer function is then

\[
\frac{\Theta_m}{E}(s) = \frac{1}{H(s)} \frac{1}{Js^2 + js + k}
\]

4. Whenever the characteristic curves are straight lines, it is possible to replace, in the transfer function, \( \partial \Omega_m/\partial C_m \) by \( \Omega_M/C_M \), where \( \Omega_M \) and \( C_M \) are respectively the running speed with zero resistive torque and the stall torque at zero rpm (Fig. 30-11). The equation is written in this manner in the following chapter for d-c motors (Sec. 31.4.3).

**30.3.4. Transfer Function and Transfer Matrix.** Although useful, the transfer-function concept is limited to only some special cases of the general problem in that it expresses the manner in which an input signal characterized by a voltage and a current vs. time is transformed by the motor into a response characterized by a velocity and a torque vs. time.

It has been mentioned in Sec. 10.2 that the transfer-matrix concept enables complete visualization of the over-all problem. It is not necessary then that the disturbing torques \( C_r \) be related to the output shaft \( \theta_r \) by a relationship of the form

\[
C_r(s) = F(s)\Theta_r(s)
\]  

(30-1)

The diagram shown in Fig. 30-12 has been previously presented. To it corresponds the following equation in terms of the Laplace variable:

\[
\Theta_r(s) = \frac{AB}{1 + AB} E(s) + \frac{B}{1 + AB} C_r(s)
\]  

(30-2)

If \( C_r \), usually termed the disturbing torque, is considered as the useful torque associated with the position \( \theta_r \), the output signal can be character-
ized by its two components: the useful torque \( C_r \) and the output shaft velocity \( d\theta_r/dt \). Equation (30-2) is no longer a transfer function. The simplicity of this result is due to the linearity of the equation and its consequence, the principle of superposition. This does not imply that Eq. (30-1) is valid. If it is, there exists a conventional transfer-function relationship between the two variables \( \Theta_r(s) \) and \( E(s) \).

![Fig. 30-12.](image)

Assuming that a system is linear, it is always possible to relate the two electrical components \( V_s \) and \( I_s \) of the input signal to the mechanical components \( C_M \) and \( \theta_\sigma \), of the output signal by a matrix equation of the form

\[
\begin{bmatrix} V_s \\ I_s \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} \Theta_r \\ C_r \end{bmatrix} = \begin{bmatrix} Q \end{bmatrix} \begin{bmatrix} \Theta_r \\ C_r \end{bmatrix}
\]

where \( Q \) is the inverse transfer matrix of the motor (Sec. 10.2).

Conditions for stability cannot very easily be stated mathematically. The authors know of no necessary and sufficient stability criterion. One sufficient condition, however, is that the system is certainly stable if all the characteristic values of the transfer matrix are real and positive. Some other criteria can be given and applied to special cases. They will not be outlined here (see Sec. 21.3.3, par. 3).

Using the transfer-matrix concept, it is possible to consider the motor as a torque or a motion generator, the data contained in the transfer matrix being sufficient to characterize the motor fully. The transfer matrix, however, more readily leads to the concept of a torque generator, since the equation of the motor is written as

\[
\text{Transform of torque} = F(s) \times \text{transform of input signal}
\]

where \( F(s) \) is a linear operator. Nevertheless, it is possible to consider a motor with a known internal impedance as a motion generator.

### 30.4. POINTS FOR COMPARISON OF THE VARIOUS TYPES OF MOTORS

#### 30.4.1. Graphical Comparison. The characteristic curves of servomotors are either of the type indicated in Fig. 30-13a or of the type indicated in Fig. 30-13b. The curves of Fig. 30-13a represent the characteristics of (a) d-c motors with armature control and (b) two-phase a-c motors linearized in the useful domain. The curves of Fig. 30-13b represent the characteristics of (c) d-c motors with field control and (d) hydraulic motors. For the latter, above a certain speed the characteristics always tend toward the speed axis.
Although the characteristics shown in Fig. 30-13b could be considered to be more satisfactory than those shown in Fig. 30-13a, it should be noted that:

1. The transfer functions of the corresponding motors are not the same. The exceptionally important stability considerations are not equivalent, and only detailed evaluation of the problem can lead to a final conclusion.

2. The time required to attain any given speed depends on the respective position of the curves—i.e., on the values $C_1, C_2, \ldots$ —and any comparison must take into account the actual numerical values.

In any case, to compare two motors, it is necessary to consider, in addition to the characteristics, the total inertia referred to the motor shaft. This is the inertia of the load referred to the motor shaft in addition to the inertia of the rotor itself.

![Characteristic curves of armature- and field-controlled motors.](image)

**30.4.2. Internal Impedance.** The mechanical equivalent of the electrical system described by the equation (using conventional symbols)

$$\ U = \epsilon - ZI$$

relating voltage $U$, current $I$, electromotive force $\epsilon$, and impedance $Z$, in the voltage $\sim$ speed, current $\sim$ torque analogy, is

$$\Omega = \Omega_M - ZC$$

where $\Omega_M$ is the maximum running speed for a given input signal.

The internal impedance $Z$ of the motor is then

$$Z = \frac{\Omega_M}{C_M} = \tan \psi$$

It is immediately apparent that:

a. A motor having a characteristic of the type shown in Fig. 30-13b has infinite internal impedance. It is a constant-torque motor, an analog of a constant-current generator (see Fig. 2-5).

b. For those cases in which the useful quantity is motion (position or rate) and not energy, the motor that is to be selected is one that is capable of delivering the maximum required power. It should be the one having the minimum internal impedance, i.e., the minimum angle $\psi$ (Fig. 30-14).
It is easily shown that the maximum available power corresponds to a point having the coordinates $\Omega_M/2$ and $C_M/2$. All the characteristic curves of linear motors having the same maximum available power $P$ are shown in Fig. 30-15. These curves are straight lines tangent to the equilateral hyperbola whose equation is $\omega C = \text{const}$. The best motor for a servomechanism is that whose characteristic has the minimum slope with respect to the torque axis ($\psi = 0$). The operating point at the maximum available power is the point of contact with the envelope. However, for a regulator (Sec. 13.1.7), the angle $\psi$ should approach $\pi/2$. This is also true of torque motors of the type used in steel mills, operating at a quasi-constant speed and against a widely varying load.

It should not be concluded that the best motor is one having a characteristic with a slope between $\psi = 0$ and $\pi/2$. Actually, depending on the problem under consideration, the value of $\psi$ should approach either 0 or $\pi/2$.

30.5. POSITIONING THE CHARACTERISTIC AND DESIGN CALCULATIONS OF THE MOTOR IN COMPLIANCE WITH THE SPECIFICATIONS

30.5.1. The Problem. It is necessary to interpret the requirements in order that the characteristic curve may be so positioned in the $(\Omega, C)$ plane that it corresponds to the maximum expected input signal. In the following discussion it is assumed that the driving shaft $\theta_m$ is geared to the output shaft $\theta_r$ through a reduction gear having a ratio $\alpha = \theta_m/\theta_r$. The curve $(\Omega_m, C_m)$ that is characteristic of the motor shaft can be interpreted now in terms of output shaft $(\Omega_r, C_r)$ by the following transforms

$$\Omega_r = \Omega_m/\alpha \quad C_r = \alpha C_m$$

where $C_r$ is the torque applied to the output shaft.

It is to be noted that the characteristic curves contain only information pertaining to steady-state conditions. No information is given, for example, regarding the self-inductance of electrical motors or the mechanical inertias of the rotors. Such elements, however, must be taken into consideration when interpreting servomechanism data, e.g., when deriving transfer functions. It is obvious, then, that the design of a servo-
motor must take into consideration data that are more detailed than those implied by the characteristic plot.

The solution of the problem, which is not unique, can be carried out in steps, and it may be summarized as follows:

a. Adjustment of the characteristic and preliminary evaluation of the inertia of the motor, based on the concept of either total power required or torque and speed. As in any technological problem, the first evaluation is based on previously realized solutions. From these realizations, a satisfactory solution is found by extrapolation.

b. Check that the transfer function of the motor so defined is acceptable. This checking is often limited to the evaluation of the phase at the maximum specified frequency. This may show that a slight adjustment in the characteristic is desirable, the adjustment being made according to the indications of the previous paragraph and within the limits specified by the conditions stipulated.

c. Detailed calculations and actual realization of the equipment so selected. These calculations usually lead to a second approximation in steps a and b. For a detailed account of the second step, the reader will profit by referring to the two following chapters and to books that deal in detail with motors.

There are two possible methods of adjusting the characteristics (step a in the above procedure). The first is the harmonic-response, or sinusoidal-response, method, enabling the design of the motor just within the requirements. This method is to be used only when detailed specifications describe the required performance. The second, the modulus method, is easier to use and enables faster computation. It usually leads to a motor that is more powerful than necessary, but it should be used if the specifications are somewhat incomplete. Some further consideration is given thereafter to the correspondence between transfer function and characteristic curve for a given motor.

30.5.2. The Harmonic-response Method. 1. Assumptions. It is assumed that the output element which is characterized by its inertia and resistive torque, the viscous torque being neglected for the moment, must be able to oscillate with an amplitude \( \theta_0 \) up to an angular frequency \( \omega \). The motor cannot be designed during the first approximation to meet the specified transfer function. Therefore, the usual procedure is to design it first in order that it may meet the most stringent requirement. The result is then checked and, if necessary, modified. The following symbols will be used:

- \( \theta_0 \) Amplitude of the angular position of the output shaft (load)
- \( J_e \) Inertia of the same
- \( k_e \) Spring constant of the resistive torque \( k_e \theta_e \)
- \( J_m \) Inertia of the rotor of the motor
- \( \alpha = \theta_m/\theta_e \), reduction-gear ratio

---

1 The study of servomechanisms from the viewpoint of active and reactive power, or consumed and fluctuating power, as defined in Sec. 12.3.6, is very important. The reader would profit by reading D. P. Campbell, "Power Studies in Linear Feedback Systems," MIT thesis, 1949.
Let \( C \) be the total resistive torque applied to the driving shaft. When the load oscillates sinusoidally with an amplitude \( \theta_r \) and frequency \( \omega \), \( C \) can be expressed as

\[
C(t) = -(J_m + J_e/\alpha^2)\omega^2\alpha\theta_r \sin \omega t + (k_c/\alpha)\theta_r \sin \omega t
\]

The instantaneous velocity \( \Omega_m = d\theta_m/dt \) of the motor shaft is, then,

\[
\Omega_m(t) = \alpha\omega\theta_r \cos \omega t
\]

These two equations give the expression of required torque and velocity as functions of the one parameter \( t \) and define a family of ellipses which have the axes \( OC \) and \( O\Omega \). Consequently, in the \((\Omega,C)\) plane the characteristic of the motor corresponding to the maximum input signal must remain outside the ellipse defined by the most stringent condition.

The most stringent condition in \( \theta_r \) is \( \theta_r = \theta_0 \). However, when \( \theta_r \) has been taken equal to \( \theta_0 \), the ellipse \( \omega_0 \) is not in general outside the entire family of ellipses \( \omega < \omega_0 \). The practice is to keep the characteristic outside all the ellipses \((\theta_0,\omega)\) for \( \omega \) varying in the range \((0,\omega_0)\). In other words, the characteristic should be kept outside the envelope of the ellipses depending on the parameter \( \omega \). This must be verified for all values of \( J_m \) and \( \alpha \) contemplated during the design of the motor. A set of values for \( J_m \) and \( \alpha \) is finally selected, yielding a characteristic that is satisfactory from the transfer-function viewpoint and that corresponds to the lowest power of the motor. It may be convenient to start by evaluating the inertia \( J_m \) of the motor by comparison with previous realizations.

2. Analytical Solution. For given \( \alpha \) and \( J_m \), the ellipses depending on the parameter \( \omega \) have an envelope defined by

\[
\begin{align*}
\frac{(\Omega)}{\theta_0} &= \pm \left(\frac{k_c}{J_m + J_e/\alpha^2}\right)^{1/2} \frac{\cos^2 u}{(1 + \sin^2 u)^{1/2}} \\
\frac{(C)}{\theta_0} &= \frac{k_c}{\alpha} \frac{2 \sin^3 u}{1 + \sin^3 u}
\end{align*}
\]

where \( u = \omega t \).

Figure 30-16 gives the curve in dimensionless coordinates:

\[
\frac{(\Omega/\theta_0)}{[k_c/(J_m + J_e/\alpha^2)]^{1/2}} \quad \text{and} \quad \frac{(C/\theta_0)}{\tau k_c/\alpha}
\]

This curve is graduated in values of \( u \) (in degrees) so that one envelope corresponds to the entire range of the parameters \( J_m \) and \( \alpha \).

This solution, however, does not take into account particular ellipses having no envelope. This is explained in the following paragraphs.

3. Graphical solution. Two cases are to be considered:

a. It is anticipated that one of the inertias referred to the motor shaft is negligible. For this case, usually, the inertia of the load is the one that is to be disregarded. Many examples can be given: electric servomotors driving aircraft control surfaces governed by an autopilot (see the numerical values given below), recorders using servomechanisms, etc.
If $J_c/\alpha^2$ can be considered negligible compared to $J_m$, the equations of the ellipses are:

\[
\frac{\Omega}{\theta_0} = \alpha \omega \cos \omega t
\]

\[
\frac{C}{\theta_0} = \left(\frac{k_c}{\alpha} - J_m \alpha \omega^2\right) \sin \omega t
\]

It will be shown that a family of these ellipses is easy to draw.

b. If the two inertias referred to the motor shaft have the same order of magnitude, the following approximation is valid:

\[
J_m + \frac{J_c}{\alpha^2} \approx 2 \frac{J_c}{\alpha^2}
\]

For a given value of gear ratio, the characteristic can be positioned tangentially to the envelope of the ellipses. This characteristic defines the maximum power of the corresponding motor. Let $P_\alpha$ be that value. The probable value of the inertia $J_\alpha \pm \Delta J_\alpha$ can be evaluated by consideration of existing similar motors. It is then to be verified that

\[
J_\alpha - \Delta J_\alpha < \frac{J_c}{\alpha^2} < J_\alpha + \Delta J_\alpha
\]

**Graphical Determination of the Network of Ellipses.** The solution will be given in detail for that particular case in which the inertia of the load referred to the
motor shaft can be neglected as compared with the inertia of the rotor. The validity of such an assumption must be checked at the end of the computation. The method described below is based on the use of a diagram (Chart 4 at the back of the book) and makes it possible for the ellipses to be drawn quickly and accurately and the characteristic to be positioned.

The axes of the ellipses are, in dimensionless coordinates \((\Omega/\theta_0, C/\theta_0)\),

\[
\begin{align*}
Oz & \text{ axis:} & P &= \frac{k_2}{\alpha} - J\alpha^3 \\
Oy & \text{ axis:} & Q &= \frac{\alpha}{\omega} 
\end{align*}
\]

The following units are adopted: torques, velocity, and spring constant \(k_2\) in cgs per radian and frequencies in cps (angular frequencies will not be used). All the quantities numerically expressed are given in cgs units.

![Diagram](Fig. 30-17)

**a. Determination of the Apexes \(P\) of the Ellipses.** The straight lines of the fourth quadrant are to be used for this purpose (Fig. 30-17). A proper value of \(\alpha\) is selected on the downward vertical axis \(Oy'\). One then chooses the straight line that corresponds to the frequency \(F\) in cps of the ellipse under consideration. The apexes of the ellipses \(P_1, P_2, P_3, \ldots, P_6, P_1^{100}, P_2^{100}, P_3^{100}, \ldots, P_6^{100}\) are thus determined, \(P_0\) being the apex \(P\) of the ellipse corresponding to \(F = i, \alpha = j\) (Fig. 30-18).

**b. Determination of the Apexes \(Q\) of the Ellipses (Fig. 30-19).** For a given value of the gear ratio \(\alpha\), it is possible to draw in the second quadrant the parabola whose equation is

\[
y = \frac{k_2}{\alpha} - J\alpha(2\pi F)^3
\]

Two points are easily located on the axes: the apex \(S\) of the parabola, which is on \(Oy\) and which has the ordinate \(k_2/\alpha\), and the point \(R\) where the parabola intersects \(Ox'\). It is on \(Ox\) and has the abscissa (lower graduation) \(Jm^{1/2}/2\pi\alpha k_0^{1/6}\).

It can be noted that:

1. The slope of the parabola at this point is \(4\pi (k_2 Jm)^{1/2}\) and does not depend on \(\alpha\). (It is expressed in cgs units per cps.) The tangent at an arbitrary point can be easily determined (Fig. 30-20).
Having located $S$ and $R$, it is possible to determine graphically from them the tangent at point $R$. Thus, if $R'$ is the middle of $OR$ and is projected to $R''$ on the tangent at point $S$, $RR''$ is the tangent at point $R$. If the perpendicular from $R''$ to $RR''$ intersects $Oy$ in $R'''$, then $R'''$ is the focus of the parabola.

![Fig. 30-19.](image)

![Fig. 30-20.](image)

2. The frequency $F_1$ read on the $Ox'$ axis (lower graduation) is the frequency of the mechanical resonance of the system. This, incidentally, enables an easy checking of the effect of a change in gear ratio.

3. The apex $S$ of the parabola corresponds to the stall torque. It will be shown later that this torque is necessarily the maximum available torque if the frequency $F_1$, abscissa of point $R$, is higher than the specified frequency $F_0$. However, if $F_1 < F_0$, this is not necessarily true.

To any given frequency of the $(0,F_0)$ band, read on the $OF_0$ segment, there corresponds a point of the parabola defining an ordinate, which is the $OQ$ half axis.

The points $S$ of ordinate $k_o/\alpha$ and $R$ of abscissa $J_m^{1/2}/2\pi a k_o^{1/2}$ can be graphically determined with the help of the graduated hyperbolas of the second and third quadrants. The respective scales are $k_o$ and $(k_o/J_m)^{1/2}$, and the gear ratios can be determined from $Ox'$ (upper graduation) and $Oy'$, respectively.

![Fig. 30-21.](image)

It is thus possible (Fig. 30-18) to determine the points $S_{60}$, $S_{100}$, and $R_{60}$, $R_{100}$; then $Q_{174}$, $Q_{274}$, $Q_{374}$, $Q_{474}$, $Q_{574}$, $Q_{674}$, $Q_{774}$, $Q_{1100}$, $Q_{2100}$, $Q_{3100}$, $Q_{4100}$, $Q_{5100}$, $Q_{6100}$, $Q_{7100}$, $Q_{1000}$. $Q_{i}$ is the apex of the ellipse $F = \frac{i}{i}$ and $\alpha = j$. From the points $P$ and $Q$ corresponding to the frequency $F$ (Fig. 30-21) the ellipse is drawn in the first quadrant, its axes being $\Omega/\theta$ and $C/\theta$. It is then easy to adjust the characteristic, such that it constantly maintains the ellipses to its left.

If the inertia of the load, referred to the motor shaft, is not negligible with respect to the rotor inertia, the quantity $J_m + J_r/\alpha^2$ must be used in place of $J_m$. Nothing is changed concerning the determination of the apex $P$ located on $Ox$. The quantity $Q$ becomes

$$Q = \frac{k_o}{\alpha} - \left(\frac{J_m + J_r}{\alpha^2}\right) \alpha \omega^2$$
The point \( R \) can be determined by a similar method, but the procedure is more complicated. However, it is possible to use slide-rule calculations to obtain the abscissa of point \( R \) and to make the freehand drawing of the parabola from point \( Q \) to point \( R \).

**Important Remark.** The quantity \( Q \) can actually be negative. In this case \(|Q|\) is taken on the \( Oy \) half axis, and the ellipse is drawn from that point. Actually, only one-fourth of the ellipse is drawn in one quadrant, as previously. Generally speaking, a point located in the first or third quadrant of the \((\Omega, C)\) plane represents power that must be supplied by the motor, while a point in the second or fourth quadrant represents power that must be dissipated by the motor. It is obvious that a motor designed to supply 25 watts can absorb about 25 watts and not much more. In fact, the ellipses, i.e., the isofrequency curves, are graduated in time for a sinusoidal input,

![Fig. 30-22.](image)

The point oscillates along an ellipse, completing a cycle within the time interval \( 1/2\pi F \). where \( F \) is the value (frequency in cps) of the parameter characterizing the ellipse.

If for the parabola \( F \), the abscissa of point \( R \) (Fig. 30-22), is less than \( F_0 \), corresponding to the passband of the system, the axes \( Q \) are negative \((OQ'_0)\) on Fig. 30-22) for any frequency between \( F_0 \) and \( F_1 \). The graphical solution then shows that the corresponding ellipses do not have an envelope\(^1\) and that the values of \( OQ \) can be higher than the stall torque, point \( S \) on the parabola. One may then draw ellipses that enclose the whole family of ellipses corresponding to the frequency range \((0, F_0)\). The position of the characteristic is then determined by the ellipses corresponding to frequencies higher than the resonant frequency of the motor. This result is one that should not be expected. The system is now unable to supply energy because the motor is being used at a point which exceeds its mechanical resonance.

Thus it is seen that the choice of the gear ratio may result in increasing the resonant frequency of the system above the frequency band considered as useful.

---

\(^1\) This is the reason the graphical solution is sometimes preferred to the analytical solution. To use the latter method successfully, it is necessary to check that the resonance frequency of the system is beyond the maximum specified frequency.
30.5.3. The Modulus Method. The method previously presented was concerned with sinusoidal steady-state regimes in which the applied forces were related by well-defined phase intervals. In practice, a motor must operate continuously under transient conditions and nothing can be stated a priori concerning the respective phases of the various forces. The modulus method, which will now be described, is based on the pessimistic assumptions that all the resistive torques can be arithmetically added and that each torque is at its maximum value. One of the advantageous aspects of this method, which is due to G. Newton,¹ is its simplicity, since it deals only with arithmetical sums of all terms.

Notation.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega_e$, $A_e$</td>
<td>Maximum angular velocity and acceleration of the load</td>
</tr>
<tr>
<td>$\Omega_M$, $A_M$</td>
<td>Maximum angular velocity and acceleration of the motor shaft</td>
</tr>
<tr>
<td>$J_e$</td>
<td>Inertia of the load</td>
</tr>
<tr>
<td>$J_m$</td>
<td>Inertia of the motor</td>
</tr>
<tr>
<td>$C_e$</td>
<td>Maximum resistive torque applied to the load</td>
</tr>
<tr>
<td>$C_M$</td>
<td>Maximum driving torque available from the motor shaft</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Reduction-gear ratio between motor and load</td>
</tr>
</tbody>
</table>

Actually, $\Omega_e$ and $A_e$ cannot easily be evaluated a priori in transient conditions. These quantities can be substantially higher than the corresponding values at the input of the servo loop, since they involve all the dynamics of the system. The fluctuations of the error are usually responsible for the maximum values of $\Omega_e$ and $A_e$, because of the high frequencies found in the error spectrum. This is always the case in regulator problems.

As in previous chapters, except in those dealing with random inputs, the motor will be designed with the assumption that the input is sinusoidal. The maximum value of the amplitude $\theta_e$ and of the angular frequency $\Omega_e$ are assumed given. The amplitude is either higher than or equal to the desired dynamic accuracy, and $\Omega$ will be either higher than or equal to the frequency corresponding to the bandwidth of the servo system. The equations are

$$\Omega_e = \theta_0 \omega_0 \quad A_e = \theta_0 \omega_0^2$$

just as in the harmonic-response method.

The total resistive torque applied to the load shaft is $\Gamma_e = J_e A_e + C_e$, that is, $\Gamma_e / \alpha$ applied to the motor shaft. The motor must be able to provide sufficient torque to compensate for the latter and, in addition, provide torque which will overcome the inertia of the rotor. Its maximum value is $J_m \alpha A_e$. Thus it results from the assumptions made that

$$C_M \geq J_m \alpha A_e + \frac{\Gamma_e}{\alpha} \quad \text{or} \quad J_m A_e \alpha^2 - C_M \alpha + \Gamma_e \leq 0$$

Such a condition is equivalent to the following set of inequalities:

\[
\frac{C_M^2}{J_m} \geq 4A_c \Gamma_o \quad \alpha_1 \leq \alpha \leq \alpha_2
\]

where

\[
\alpha_i = \frac{C_M \pm (C_M^2 - 4J_m A_c \Gamma_o)^{1/2}}{2J_m A_c}
\]

The first condition, that there is a real solution for \( \alpha \), involves the ratio \( C_M/|J_m|^{1/4} \), the characteristic of the motor required to balance a load acceleration \( A_c \). This condition has already been stated in Sec. 30.2.1.

The second condition defines two limits for \( \alpha \). The first is that this ratio must be large enough to enable the driving torque referred to the output shaft to be sufficient. The second is that it must be small enough to enable the rotor inertia referred to the output shaft to be unprohibitive. To these two conditions, which pertain to torques, another, which pertains to velocities, must be added:

\[
\Omega_M \geq \alpha \Omega_o
\]

where \( \Omega_o \) is specified.

*Note.* In the modulus method, the natural frequency is not as important a factor as in the harmonic-response method. Thus, it was shown in the latter that, in the absence of frictional torques, no energy is required to maintain the oscillation. Conversely, in the modulus method, the accelerating energy (work done by inertia forces) is added to the energy required to counteract the maximum resistive torque.

**30.5.4. Conditions Pertaining to the Transfer Function.** This is step 6 of Sec. 30.5.1, which describes the second step in the design of a servomotor. This step largely depends on the particular aspects of the proposed problem. The form of the transfer function radically changes if there is, in the motor stage, provision for position feedback, tachometric feedback, integration, etc. In any case, the problem can be reduced to that of an open-loop system. In general the transfer function is characterized by the maximum permissible value of the amplitude ratio or of the phase within a certain frequency bandwidth.

A particular case will now be described in detail. The motor is assumed to operate under open-loop conditions, it is linear, and it is reduced to a pure inertia. It has been previously shown in Sec. 30.3.3 that the transfer function of the motor-gear combination can be approximated by

\[
\frac{\Theta_r}{E} (p) = \frac{\partial \Omega_m/\partial E}{\tau} \frac{1}{p(1 + \tau p)}
\]

where \( \tau = J(-\partial \Omega_m/\partial C_m) \), \( E \) is the input signal, and \( \tau \) is the velocity time constant of the motor.

The quantity \( \tau \) can be expressed as a function of the maximum motor torque \( C_M \); the inertias \( J_m \) and \( J_o \) of the motor and the load, respectively; the reduction-gear ratio \( \alpha \); and the maximum angular velocity of the
load $\Omega_c = \Omega_m/\alpha$. This quantity is

$$\tau = J \frac{\Omega_M}{C_M} = \frac{\alpha J_m + J_s/\alpha}{C_M} \Omega_c$$

The following conclusions can be drawn:

1. All the other characteristics of the system being kept constant, the smaller the angle $\psi$ that the characteristic curve of the motor makes with the torque axis, the smaller the velocity time constant.

2. If $\Omega_M$ is specified, the motor acceleration is defined by the ratio $C_M/J$.

3. The inertias of the motor and the load appear in the numerator as

$$J = \alpha J_m + \frac{J_s}{\alpha}$$

Here again, as in Sec. 30.5.2, it should be determined at the outset whether the inertia of the load referred to the motor shaft $J_s/\alpha^2$ is of an order of magnitude lower than, equal to, or higher than the motor inertia. In the first two cases, $\tau$ can be expressed as $\tau = (J_m/C_M) \Omega_c$; that is, $\tau$ depends on the motor inertia through the product $\alpha J_m$. It is for this reason that, whenever a small motor is required, in autopilots, for instance, it must be so selected that it has a very low inertia. Furthermore, the accompanying gear train must have a high ratio so that the angular speed of the motor is high. If $\alpha$ is decreased, the maximum velocity decreases and the maximum torque increases. This implies that the inertia is increased. It is possible to obtain low-speed motors having very high available torque.

Such considerations pertaining to step b of the selection process of a motor could be extensively described in detail, and many additional comments could be made on particular cases. In general, the result of the discussion is, as in the case presented above, an inequality giving the maximum value of the time constant.

30.5.5. Numerical Example. In Sec. 34.3.2, the complete design calculations of a motor based on the modulus method will be carried out. The present example shows the manner in which the harmonic response method is used in the design of a motor. It describes the design of a servomotor that is to be used for driving the control surface of an autopiloted aircraft.

The specified quantities are:

- Inertia of the control surface: 5,000 cgs
- Aerodynamic resistive torque: $45 \times 10^4$ cgs
- Maximum angular displacement: 0.1 rad

It is required that, for any frequency within the bandwidth of 0 to 5 cps, the servo be capable of driving the control surface with a maximum amplitude of 0.1 rad. An armature-controlled electric motor is to be used. By referring to previous design realizations, the inertia of the motor can be assumed to be, as a first approximation, $J_m = 200$ cgs.
To retain clarity, only the ellipses corresponding to $\alpha = 100$ and their envelope have been drawn (Fig. 30-23). The natural frequency of the system is: $\omega^2 = 45 \times 10^7/20\alpha^2$. There is no damping term. The characteristic curve of the motor must remain outside the ellipses corresponding to the frequency band 0 to 5 cps. In Fig. 30-23, the ellipses $F = 1, 2, \ldots, 8$ cps have been drawn.

Fig. 30-23.

The optimum motor having the minimum power (not including the efficiencies of intermediate elements such as gears) is one having its characteristic curve tangent to both the envelope of the ellipses and an equilateral hyperbola, and remaining outside the ellipses. The optimum characteristic curve has been positioned in Fig. 30-24; it determines both the running speed and the stall torque of the contemplated motor.

As previously explained, the characteristic so determined in this first step of the design can eventually be altered in the second step of the design. The final characteristic will certainly have, however, a smaller slope with respect to the torque axis (Fig. 30-25) than that of the first approximation (Sec. 30.4.2).
It must be pointed out that, in general, the performance is specified up to a reference frequency $F_s$; in the present case, 5 cps. The envelope is then a segment of a curve (Fig. 30-26) extending from the torque axis (since for $\Omega = 0$, $P = 0$ and $Q = 4.5 \times 10^4$ in the example under consideration) to the point $P$, at which $F = F_0$. The tangent at point $P$ to the 5-cps ellipse $E_s$ and to the characteristic $F$ (straight line $\Omega_s, C_s$) corresponds to a characteristic making a smaller angle with $OC$; that is, the point of contact $A$ of the curve $F$ and the equilateral hyperbola $H_1$ is located between $R$ and $P$. Conversely, this obviously corresponds to a motor that is more powerful than it need be (in the considered conditions of Fig. 30-25), since the characteristic is tangent to an $H_1$ hyperbola.

The characteristic, however, fully determines both the running speed and the minimum power at the operating point ($\Omega_s/2, C_s/2$). This characteristic is shown in Fig. 30-24, and it is possible to calculate the power of the motor. It is quite clear that this computed power is the required power delivered at the shaft. It is necessary, then, to take into account the respective efficiencies of the motor itself, the link transmission,
Fig. 30-25.

Fig. 30-26.
gear train, etc., to evaluate the maximum power required at the input of the motor. In general, this corresponds to the power at the output of the amplifier.

30.5.6. Note. Performance Variations of a Motor with or without Load. It is frequently required to ensure that the output angular speed of a servomotor will not change by more than \( z \) per cent between the conditions of no load and maximum specified load. This statement is not complete; it must be interpreted as follows: If there is no friction or spring torque, the speed of the motor approaches a limit which is the free-running speed of the motor. For torque motors this limit speed is due to the droop of the characteristic beyond normal operating conditions. It is important to note that such motors are subjected to specified limitations when operated under open-loop conditions. For example, a minimum load must be maintained, or the command signal must be maintained within a certain region. In closed-loop operation, however, a tachometric generator is usually provided to maintain the speed to a safe value, as well as for other reasons.

Let \( \Omega_0 \) be the running speed for a given value \( \varepsilon_0 \) of the command signal, and \( \Omega_1 \) be the speed corresponding to the same signal \( \varepsilon_0 \) when the motor works against the resistive torque of the load. The load is assumed to be a known frictional torque \( C_0 \) applied to the output shaft. The above-stated requirement is

\[
\frac{\Omega_0 - \Omega_1}{\Omega_0} < \frac{z}{100}
\]

The quantity \( \Omega_1 \) can be calculated by application of the final-value theorem to the quantity \( s \Theta_r \). The result is

\[
\Omega_1 = \lim_{s \to 0} s \Theta_r = \lim_{s \to 0} s \Theta_r
\]

The quantity \( s \Theta_r \) (Sec. 30.3.4) is

\[
s \Theta_r = F_1(s)E(s) + F_2(s)C(s)
\]

If \( C_0 \) is applied as a step at time zero, \( C(s) = C_0/s \). Then

\[
s \Theta_r = F_1(s)E(s) + F_2(s) \frac{C_0}{s}
\]

The limit of \( s \Theta_r \) depends on \( C_0 \). Then

\[
\frac{\Omega_0 - \Omega_1}{\Omega_0} = C_0 \frac{\lim_{s \to 0} F_2(s)}{\lim_{s \to 0} s F_1(s)}
\]

Note. Direct calculation of the limit speed is just as easy when directly based on the equations of the mechanical system.
CHAPTER 31

ELECTRIC SERVOMOTORS

Summary
2. Effect of inductances on time constants.
3. Field-controlled motors.
4. Armature-controlled motors.
5. Alternating-current motors.

31.1. GENERAL FEATURES OF MOTORS

31.1.1. General. An electric motor generally consists of a stationary field circuit (stator) and a rotating armature (rotor). The various classifications are described below. It is customary to call \( E \) the applied voltage and \( Q \) any auxiliary source supplying constant voltage or current.

1. Direct-current Motors. a. SEPARATE EXCITATION (Fig. 31-1). In general, either the field or the armature circuit is energized by a constant-current or -voltage source, while the other is energized by the amplifier stage of the servo loop. This type of control is respectively called armature or field control.

![Fig. 31-1. Separately excited motor.](image1)

![Fig. 31-2. Series-excited motor.](image2)

b. SERIES EXCITATION (Fig. 31-2). It will be recalled that this type of motor has a torque-speed characteristic of the nature shown in Fig. 31-3. It will be seen that the motor is nonlinear, and that it is very seldom used in servo systems. A modification of this type of motor involves two separately excited fields (Fig. 31-4) in series with the armature. This motor is linear, as will be seen later, and is called a split-series motor.

c. SHUNT EXCITATION (Fig. 31-5). This type of motor is not linear, and it is of little importance in servo-system applications.

d. COMPOUND EXCITATION (Fig. 31-6). This type is a combination of types b and c; it is not used in servo systems.
2. **Alternating-current Motors.**

   **a. Universal Motor.** This is a direct-current motor, either series or partially compound-wound. The magnetic core is completely laminated to prevent eddy-current losses. The motor has no application in servo systems.

   ![Diagram of Universal Motor](image)

   **Fig. 31-3.**

   ![Diagram of Universal Motor](image)

   **Fig. 31-4.**

   ![Diagram of Universal Motor](image)

   **Fig. 31-5.** Shunt-excited motor.

   ![Diagram of Universal Motor](image)

   **Fig. 31-6.** Compound-excited motor.

   ![Diagram of Universal Motor](image)

   **Fig. 31-7.** Nonsynchronous motor.

   ![Diagram of Universal Motor](image)

   **Fig. 31-8.**

   **b. Nonsynchronous Induction Motor.** The most common is the three-phase squirrel-cage type (Fig. 31-7). It has, unfortunately, practically no application in servo systems in its classic form. One type, however, which operates from a two-phase supply, is very useful, and it will be discussed in Sec. 31.5 (Fig. 31-8).

   **c. Synchronous Motors.** Such motors are used extensively in servo systems, but as sensing devices and not as motors (Sec. 29.3.7, on selsyns).
31.1.2. Special Characteristics of Electric Servomotors. The efficiency of industrial motors is generally limited by the maximum amount of heating permitted. This determines the maximum power output in steady-state conditions. However, many motors not used in servo systems are required to drive widely varying loads, for example, machine-tool motors. The resulting speed variations depend essentially on the characteristic curves of the motor, on its inertia, and eventually on the voltage-limiting systems introduced between the motor and the power source.

The most important characteristic of an electric servo motor is not the steady-state output power, but the maximum rotor acceleration. Occasionally other considerations take precedence. For example, in aeronautics an important factor is the power per unit mass or volume. This requires that the motors turn at angular speeds of up to 10,000 rpm, and this in turn requires the use of reduction gears.

In general, d-c motors have a certain number of undesirable features. Thus, the variations of the voltage across the commutators cause the torque that is developed by the motor to vary. Furthermore, commutators are not only costly but require much maintenance and introduce noise. The effect of armature windings in slots parallel to the rotor axis is to develop a torque which oscillates about a mean value. It is, therefore, impossible to maintain the rotor between the two stable positions determined by the armature-winding pitch. To overcome this difficulty, slots for helical windings must be provided for in servomotors. Although torque modulation cannot be completely eliminated by this method, it is possible, nevertheless, to reduce it considerably. Another solution involves the use of two or more motors in parallel, coupled to the same gear train, and conveniently located with respect to each other. This arrangement also has the advantage of diminishing the total backlash, since the axes of the different gears are relatively well located.

In servo systems, a-c motors are also used. These are nearly always two-phase motors of the type discussed in detail in Sec. 31.5.2 and of power output of less than 500 watts. As for most a-c motors, operation is based on the principle of a rotating field. Such rotating fields are easily produced by alternating-voltage sources.

Whatever system is used, it is useful to remember that electric motors have a theoretical efficiency limit because of saturation.\(^1\)

31.1.3. Maximum Theoretical Efficiency. The maximum tangential stress per unit surface area which may be applied to the rotor depends solely on the value of the induction vector, or more specifically, on the normal component of the vector. Before calculating this vector, it is

---

\(^1\) In the chapter on sensing devices mention has been made of the need for knowing the theoretical limit of a phenomenon, if such a limit exists. If the real physical phenomenon is close to this theoretical limit, it is very difficult to improve the system. However, if there is a considerable difference between the practical and theoretical limits, much can be gained by trying to improve the efficiency of the system. This should be done at all times. It will be seen that this is particularly true of electric motors.
useful to recall the few fundamental electromagnetic principles which have been previously discussed.

First, induction is the quantity that is to be considered when designing a motor. The magnetic field is merely a means of obtaining a satisfactory result. In the rationalized system, where the permeability of free space is not unity, one cannot consider the magnetic-field vector $H$ and the induction vector $B$ as being equal even in free space.

The definition of the induction vector is $B = kH + \mathcal{J}$ where $\mathcal{J}$ is the intensity of magnetization. Therefore, the fundamental results are:

a. Induction flux is conservative. This is represented by $\text{div } B = 0$.

b. The normal component of the induction vector is constant when passing from one medium to another. The tangential component, however, is generally not constant under the same conditions.\(^1\)

![Fig. 31-9.](image)

![Fig. 31-10.](image)

Finally, the quantities which appear in Maxwell's and Laplace's laws are those of induction, and not those of the magnetic field.\(^2\)

Consider now a d-c motor represented only by its stator and rotor (Fig. 31-9). The induction vector is, in general, not radial. Its modulus and angular position vary along the rotor surface. Let $B$ represent the modulus of the induction vector at point $M$ on the rotor (Fig. 31-10). The magnetic pressure (Maxwell's pressure) on a surface $ds$ perpendicular to $B$ is

$$\frac{dF}{ds} = \frac{\mathcal{J}}{8\pi}$$

The tangential component $T$ of the force resulting from this pressure, which is present on the corresponding element $d\sigma$ on the rotor surface, is $T = F \sin \theta \cos \theta$. The tangential component is a maximum for $\theta = \pi/4$.

There is no flux due to the vector $B$ cutting the rotor surface. If it is now supposed that the vector $B$ is sinusoidally distributed, and its mean value is $B^s/2$, the total tangential force per unit surface is $T = B^s/32\pi$. For $B = 18,000$ gauss, the maximum value which can be attained, it will be found that $T = 3$ kg/cm². Under these conditions very large surface current densities would be required to control the

\(^1\) The components of the magnetic-field vector behave in the opposite manner. The difficulties relative to the consideration of this last vector can now be appreciated, since one must specify the direction of the magnetic-field vector.

\(^2\) See footnote of Sec. 31.2.1.
direction of the induction vector. The amount required would be of the order of 5,000 amp/cm² of rotor surface, and this is not possible at the present time.

Values of $T$ never exceed 0.3 kg/cm², even in the most advanced motors.¹

Recent progress in silicon insulation, which allows operating temperatures of approximately 200°C, should enable a notable increase in motor efficiency in the near future.

31.1.4. Steady-state Motor Calculations. 1. Description. The motor has an even $(2q)$ number of poles and the armature may be wound in any one of several ways:

a. TOROIDAL WINDING (Fig. 31-11). This winding is difficult to achieve in practice and is not widely used in servomotors; the reason is that, for comparable motors, the moment of inertia of the armature is greater than for any other kind of winding.

b. DRUM WINDING (Fig. 31-12). This type of winding is the one that is most generally used. A motor in which $q > 1$ does not necessarily have $2q$ brushes; thus, a four-pole motor may have only two brushes. The important condition is that the commutator segments be appropriately connected as shown in Fig. 31-13, which represents a two-brush four-pole motor with a toroidal winding. For a drum winding, the principle is the same.

For all cases, the armature reaction flux, that is, the magnetic field due to armature current, has a fixed direction relative to the winding. Its modulus, which is proportional to armature current, varies and is perpendicular to the field originating from the winding. Unless otherwise stated, it will be assumed that the flux is compensated, for example, by means of interpoles (Fig. 31-14) in series with the armature. The resultant flux across the armature is therefore equal to the field flux when there is no armature current.

The general steady-state equations for d-c motors are developed below.

2. Notation. Field Circuit:

\( R, L, I, E \)  Resistance, inductance, current, emf of field circuit
\( a \)  Interpole angle, i.e., ratio of air-gap distance to corresponding circumference
\( d \)  Diameter of section crossed by field flux

Armature Circuit:

\( L \)  Length of armature crossed by field flux
\( J_m \)  Rotor inertia
\( \theta_m, \Omega_m \)  Position and angular velocity of rotor (radians and radians per second)
\( C_m \)  Apparent rotor torque
\( M \)  Maximum angular velocity (no load)
\( B \)  Induction
\( \Phi \)  Induction flux
\( N \)  Number of armature conductors
\( r, l, i \)  Resistance, inductance, armature current
\( v \)  Armature voltage
\( \sigma \)  Cross section of armature conductors
\( 2q \)  Number of poles
\( 2s \)  Number of parallel paths, that is, the number of conductors in parallel on each commutator segment

Load:

\( \theta_r, \Omega_r \)  Output position and angular velocity
\( \alpha = \frac{\theta_m}{\theta_r} \)  Gear ratio
\( \gamma \)  Inertia, viscous friction, and stiffness coefficient on output shaft (load)
Total inertia on drive shaft $J = J_m + J_c/\alpha^2$

Total coefficient of viscous friction on drive shaft;

$$f = f_m + \frac{f_c}{\alpha^2}$$

3. General Considerations. Saturation of any of the components is undesirable and is to be avoided as far as possible. In general, the tendency toward saturation is most pronounced in the armature teeth or, more precisely, in the area between the slots of the armature core.

The field flux can be calculated from the relations given in Sec. 31-2. However, it must be remembered that in small motors the leakage flux may be as high as 20 per cent of the total useful flux. Therefore, it is necessary to increase the magnetizing ampere-turns produced by the field by the same amount. Induction in the air gaps may vary from 3,000 to 8,000 gauss, which leads to an induction of almost 20,000 gauss in the armature slots.

The equations of a two-pole ($q = 1$) motor with two parallel paths ($s = 1$) will now be developed. The $N$ conductors make up $N/2$ turns. Let $\beta$ represent the relation between the area occupied by the conductors and the cross section of the armature $\pi d^2/4$. Consequently,

$$\beta = 4 \frac{N\sigma}{\pi d^2}$$

The quantity $\beta$ depends therefore upon the geometry of the armature, and to a first approximation it is equal to 0.3.

4. Steady-state Equations. The torque due to current $i$ is:

$$C_m = \frac{1}{10} \frac{i}{2} \mathcal{L} B \times 0.7 \times N \frac{d}{2} = 0.0175N \mathcal{L} Bi$$

This equation may be derived from the Biot-Savart law by noting that the current in each conductor is $i/2$. This is due to the fact that the current is divided into two equal quantities on the commutator segment in contact with the brush. The brush emf is

$$e = 0.0175N \mathcal{L} B \Omega \times 10^{-8} \text{ volts}$$

To determine the armature resistance, two parallel circuits, each consisting of $N/4$ turns, are considered. The resistances of one turn and of the armature are respectively

$$R_1 = \rho \frac{2(\mathcal{L} + d)}{\sigma} \quad r = \frac{N}{4} \rho \frac{\mathcal{L} + d}{\sigma}$$

The armature resistance is determined by successively extrapolating from a series of estimated values. Some of the quantities that are to be estimated are:
a. The ratio \( L/d \) (length over diameter) of the armature (seldom greater than 3)
b. The ratio \( \beta = 4N\sigma/\pi d^2 \) giving the wound surface of the armature
c. The heat, ventilation, and viscous-friction losses

From this series of estimations the armature efficiency, i.e., the power used divided by the power supplied, can be approximated. After the initial estimation of these values is made, the armature efficiency is calculated. If this value differs greatly from the previously estimated value, a second set of estimations is made in the same manner. For the different methods of calculating the parameters of a motor, it is suggested that reference be made to a book especially devoted to design.

Finally, it is worth noting that in servomotors much larger current densities may be used than in industrial motors. For the latter, values of 5 amp/mm² are used. In servomotors, values of 10 and often 15 amp/mm² are used.

31.2. EFFECT OF INDUCTANCE ON TIME CONSTANTS

31.2.1. Magnetic-circuit Calculations. The field circuit of a motor (Fig. 31-16) comprises various components and is called a magnetic circuit. Since \( \text{div } \mathbf{B} = 0 \) (Sec. 31.1.3), the induction flux lines form a closed path. It is useful to introduce the mmf \( \mathcal{F} \) due to the different windings encountered in the magnetic circuit,

\[
\mathcal{F} = \sum NI = \int_C \mathbf{B} \cdot d\mathbf{l}
\]

where \( N \) is the number of turns in which there is current \( I \) and \( C \) is a path cutting the turns. Finally, it is useful to introduce the magnetic relucance \( \mathcal{R} \) of the circuit. It depends on the construction and permeability of the various circuit elements. If \( \mathcal{R} \) represents the inductive flux, then across a cylinder

\[
\mathcal{F} = \mathcal{R}\Phi
\]

For actual calculations, the reluctance $\mathcal{R}$ is treated in much the same manner as electrical resistance. Thus, if in the magnetic circuit of Fig. 31-16 the reluctances of the different segments are termed $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, the total reluctance is $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2 \mathcal{R}_3 / (\mathcal{R}_2 + \mathcal{R}_3)$. This similarity may be extended further. Thus, for a cylindrical element of length $l$, permeability $\mu$, and cross section $s$, the reluctance $\mathcal{R}$ is

$$\mathcal{R} = \frac{l}{\mu s}$$

This analogy, however, is not complete for the following reason: To maintain constant current work must be done, whereas once a field is produced, it persists without further work.

31.2.2. Energy Stored in an Inductance. An electric circuit consisting only of resistors would have a time constant equal to zero and would only produce heat. It may be proved that, to have conversion from electrical energy to mechanical work, the induction coefficient between the armature and the field must be a function of time. Therefore, it is obvious that this coefficient of induction must exist. This implies that each circuit has a self-inductance which is different from zero. In general, if $L_{ij}$ is the mutual inductance between two segments of field and armature indicated by $i$ and $j$, in which flow currents $I_i$ and $I_j$, it can be shown that the converted power (electrical to mechanical, or vice versa) is

$$P = -\frac{1}{2} \sum_{ij} I_i I_j \frac{dL_{ij}}{dt}$$

Conversely, all armatures have inductances and capacitances. A straight conductor has inductance and capacitance relative to ground, and these values are different from zero. All machine circuits whose inductance is not negligible store up energy and have, therefore, one or more time constants.

Consider the following: One of the circuits in an electric motor is assumed to consist of only one mesh. This circuit may have external voltage sources and mutually inductive emf. There is, therefore, a current $i$. Suppose that this circuit consists only of the field winding that has inductance $L$ and resistance $r$, which will be assumed negligible, and $R$ is the total resistance of the circuit. Its time constant is then $\tau = L/R$, which may be written as

$$\tau = 2 \frac{L}{R} \frac{i^2}{2} \quad (31-1)$$

1 There arises the possibility of placing another resistance $\rho$ in series with the winding to decrease $\tau$. This operation is not justified, since extra energy (active) equal to $\rho i^2$ would have to be used and this would necessitate a more powerful amplifier. In general, however, for a given type of amplifier, the time constant increases with the ratio of the average power gain, output/input, required. Therefore, the decrease in the time constant in the motor stage is lost in the amplifier stage.
This relation expresses two very important points:

\( a. \) The time constant is equal to twice the ratio of energy stored in the inductance to the power used in the circuit.\(^1\)

\( b. \) For a circuit with a given time constant, it is impossible to use energy without storing energy at the same time.

In the preceding example the purpose of the field circuit is to produce a flux \( \Phi \). In fact, the winding creates a flux \( \Phi_0 \) in the core and a flux \( \Phi_l \) outside the core. The latter is termed magnetic leakage (Fig. 31-17).

![Fig. 31-17.](image)

It is now possible to express the stored magnetic energy as a function of the magnetic flux. Let \( \Phi_1 \) represent the part of the flux \( \Phi \) that acts on the rotor and \( \Phi_2 \) the part that corresponds to the magnetic leakage. Then, \( \Phi = \Phi_1 + \Phi_2 \). If \( k \) is a proportionality coefficient depending on pole geometry, \( \Phi = k \mu s N \). The value of self-inductance is

\[
L = 10^{-5} k \mu s N^2
\]

The energy stored in the inductance is, therefore, \( W_1 = \frac{1}{2} 10^{-5} k \mu s N^2 \) joules. The elimination of \( N \) gives

\[
W_1 = \frac{1}{2} 10^{-8} \frac{\Phi^2}{\mu \epsilon s} \text{ joules}
\]

Eq. (31-1) becomes therefore

\[
\tau = 2 \frac{10^{-5} \Phi^2 / 2 k \mu s}{R i^2}
\]

\(^1\) In the steady state, the power dissipated equals power supplied.

\(^2\) If the solenoid comprising the pole winding is long compared to its diameter, \( k = 4 \pi / 10 \ell \), where \( \ell \) is the length of the solenoid. In general, this condition is not considered, and the coefficient \( k \) depends on the pole geometry (windings and magnetic cores).
This expression states that: (a) For a given total flux (utilized plus losses), the time constant of the field circuit decreases as the energy used increases. (b) For a given useful flux and a given power, the time constant decreases as the leakage flux decreases.

By introducing the average length of a turn and the current density, it is noticed that the time constant decreases, as current densities increase. There is, however, a limitation to this rule that is due to the heating effect. But it must be remembered that the field of a motor is more easily cooled than the armature.

Thus far only the field circuit has been considered. It is evident that these conclusions apply also to the armature. However, the armature inductance is generally much smaller than the field inductance, and the flux originating in the armature (armature reactance) may be compensated (interpoles).

Conversely, it is difficult to define the armature circuit exactly. It is very difficult to calculate the leakage flux, although this flux represents a considerable percentage (about 10) of the total flux, even in the more advanced machines. It is obvious, therefore, that much attention must be devoted to the design of the magnetic circuit of a servomotor, since its time constant directly depends on the total flux created.

Finally, it can be said that the magnetic field due to a current \(i\) is the gradient of a potential whose expression is the product of \(i\) and the solid angle through which the circuit producing the field can be seen from the point under consideration. This rule is sometimes useful in determining the fields. It must also be remembered that all problems of field design (equipotential lines, lines of force) may be either solved or greatly simplified by the utilization of analog computing methods, i.e., by the use of electrolytic tanks, resistor networks, etc.\(^1\)

From these considerations it is apparent that the design of a magnetic circuit of any machine is very difficult. In the preliminary stage the designer must consult specialized books on the technology of rotating machinery, and use well-developed methods to obtain minimum values of time constants.

\[\text{31.2.3. Calculation of Inductances.}\] The calculation of the inductances of the different motor circuits is always a very complicated problem; a direct method under operating conditions is always preferable. Consider the simple example of a separately excited motor (Fig. 31-1). The field resistance is a known quantity and may be accurately measured. If the field inductance is measured when the motor is stationary, a value of the time constant equal to \(\tau_1 = L/R\) is found. However, under operating conditions, it will easily be seen that the time constant \(\tau_2\) is greater than \(\tau_1\). This is due to eddy currents which may be considered as flowing in a short-circuited winding located in the field circuit (Fig. 31-18). This circuit acts as the short-circuited secondary of a transformer. It has an inductance \(L_1\) and a resistance \(R_2\). Both quantities are small in modulus, but their ratio \(L_2/R_2\), which has the dimension of time, may be large. These explanations illustrate satisfactorily the experimental results obtained.

\(^1\)See the references quoted in Sec. 9.3.7. Also see M. Pélegrin, "Machines à calculer électroniques," Dunod, Paris, 1959.
The analysis given in Sec. 31.4.2 shows that the time constant $\tau_1$ is to be replaced by $\tau$ such that $\tau = \tau_1 + L_2/R_2$.

In general, for the calculation of the inductances the reader is referred to books on motor design. However, suggestions will be given here\(^1\) for calculating armature inductance.

![Diagram of armature inductance](image)

**Fig. 31-18.**

The magnetic energy stored in an armature of inductance $L$ in which flows a current $i$ is $W_1 = \frac{1}{2}Li^2$ joules. The density of magnetic energy is proportional to the square of the induction

$$W_1 = \frac{B^2}{8\pi} \quad (31-2)$$

The problem is now reduced to that of calculating $B$. Suppose that the conductors are evenly distributed on the rotor surface (Fig. 31-19). The flux lines form a closed path inside the pole. If there is no saturation, the induction $B$ at point $x$ is approximately

$$B = \frac{4}{10} n_i \frac{1}{\Delta d} x \quad (31-3)$$

where $i =$ armature current

$\Delta d =$ airgap length (corrected later to compensate for the slots)

$a =$ interpole angle

$x =$ coordinate

$n =$ total number of conductors of armature

and $n_i =$ number of reduced conductors, defined as follows: If $2q$ is the number of poles and $2s$ the number of parallel paths (that is, the number of conductors in parallel between commutator segments), $n_i = n/2q \cdot 2s$. Replacing $B$ of Eq. (31-2) by its value from Eq. (31-3) and integrating along a pole of length $b$ (Fig. 31-19) yields the expression of the magnetic energy stored in the airgaps and allows determination of the induction coefficient. This is found to be proportional to $n_i b^4/a^3 \Delta d$.

**Hopkinson's Coefficient.** It is customary to characterize leakage by a coefficient. Suppose that there are two windings on a pole. If the flux $\Phi$ produced by the first winding acts upon all the turns of this winding and only a fraction $\Phi/\nu$ (where $\nu > 1$) of this flux acts upon a turn of the second winding, the coefficient $\nu$ is called Hopkinson's coefficient.

Instead of the relation \( M = (L_1L_2)^{1/2} \), where \( M \) is the mutual inductance and \( L_1, L_2 \) are the self-inductances of the windings, one has the expression

\[ M = k(L_1L_2)^{1/2} \]

where the coupling coefficient \( k \) is given by \( k = 1/(\nu_1\nu_2)^{1/2} \). The quantities \( \nu_1, \nu_2 \) are then the Hopkinson coefficients of the windings.

### 31.3. FIELD-CONTROLLED MOTORS

#### 31.3.1. Principle

A separately excited d-c motor has the following characteristics: (a) the control voltage \( E \) is applied to the field coil, (b) a constant current is fed to the armature. The relation between the angular position of the motor shaft, \( \theta_m \), taken as the output, and the control voltage \( E \), taken as the input, is to be studied. The constant armature current requires a source with infinite internal impedance, because an emf proportional to \( d\theta_m/dt \) is induced in the armature. In practice, a constant current can be achieved in one of the following ways:

a. By the use of a ballast resistor \( R \) (Fig. 31-20). The voltage drop in this resistor must be large compared to the maximum back emf induced in the armature. The efficiency of the motor for this condition is necessarily low, and the armature current is only approximately constant. A variation of this method is shown in Fig. 31-21, in which a bridge rectifier is connected to an a-c voltage source and feeds the armature. A capacitance (or inductance) is connected in series, and its impedance \( 1/Cj\omega \) (or \( Lj\omega \)) is made large compared to the armature resistance. The advantage of this arrangement is that no energy is lost in the ballast resistor. Another circuit achieving the same result is given in Fig. 31-22.

b. By the use of a tetrode or pentode (Fig. 31-23) the plate current of which is independent of the plate voltage over a wide range. This method is used only when the power that is involved is of the order of 10 watts or less (Fig. 31-23).

c. By the use of a metadyne. This device is a rotating machine utilizing armature reaction; it will be studied for another reason in
Chap. 33. The property that is of interest here (Fig. 31-24) is that a metadyne having a constant voltage fed to one pair of brushes delivers a constant current at the other pair of brushes. This property will be proved in Chap. 33.

![Fig. 31-22.](image1)

![Fig. 31-23.](image2)

![Fig. 31-24.](image3)

31.3.2. Transfer Function. Using the notation of Sec. 31.1.4 (redefined in Fig. 31-25), the equations of the system are

\[
E = RI + L \frac{dI}{dt} \quad \Phi = k_1 I
\]

\[
C_m = k_3 i_0 \Phi \quad \text{with} \quad k_3 = \frac{N}{2 \pi} \times 10^{-3}
\]

\[
C = J \frac{d^2 \theta_m}{dt^2} + f \frac{d\theta_m}{dt} + k_4 \theta_m \quad C_m = C
\]

\(^1\) It should be noted that the use of the voltage \(E\) as input variable implies a knowledge of the internal impedance of the source. This impedance is either lumped with \(R\) and \(L\) or is entirely neglected if the impedance of the field coil is large compared to it.
If the initial conditions are assumed to be zero, the motor transfer function is obtained by Laplace transformation. The following properties should be stressed: The transfer function of the field coil is

\[
\frac{\Phi}{E} = \frac{k_1}{R + Ls}
\]

This can be written as

\[
\frac{\Phi}{E} = \frac{k_1}{R(1 + \tau_1 s)} \quad \text{with} \quad \tau_1 = \frac{L}{R}
\]

which shows the advantage of a low field-coil inductance.

The transfer function of the rotor is

\[
\frac{\Theta_m}{\Phi} = \frac{k_2 i_0/J}{s^2 + (f/J)s + k_4/J}
\]

This can be written as

\[
\frac{\Theta_m}{\Phi} = \frac{k_2 i_0/J}{s^2 + 2\zeta \omega_n s + \omega_n^2}
\]

where \(\omega_n^2 = k_4/J\) and \(2\zeta = f(Jk_4)^{1/2}\).

With the above notations, the transfer function of the motor with a load \(C\) is

\[
\frac{\Theta_m}{E} = \frac{k_6}{(1 + \tau_1 s)(s^2 + 2\zeta \omega_n s + \omega_n^2)}
\]

where

\[
k_6 = \frac{k_1 k_3 i_0}{JR}
\]

The corresponding loci may now be drawn. By writing

\[
\frac{\Theta_m}{E} = \frac{k_6/\omega_n^2}{(1 + \tau_1 j\omega)[(j\omega/\omega_n)^2 + 2\zeta j\omega/\omega_n + 1]}
\]

and noting that \(u = \omega/\omega_n\), one obtains

\[
\frac{\Theta_m}{E} = \frac{k_6/\omega_n^2}{(1 + \tau_1 \omega_n j u)[(ju)^2 + 2\zeta ju + 1]}
\]

A particular case of frequent occurrence is that of the restraining torque (torque proportional to \(\Theta_m\)) being zero, \(k_4 = 0\). The transfer function of the motor then becomes

\[
\frac{\Theta_m}{E} = \frac{k_6}{s(1 + \tau_1 s)(1 + \tau_2 s)}
\]

where

\[
k_6 = \frac{k_1 k_3 i_0}{JR} \quad \text{and} \quad \tau_2 = \frac{J}{f}
\]
The latter represents the time constant of the rotor. The corresponding loci have been described in Sec. 13.1.3 (Figs. 13-4 and 13-5). The time constant of the rotor is smaller for a smaller $J$. Hence, the importance of reducing the inertia of a servomotor as greatly as possible is evident.

*Remark.* It was mentioned in Sec. 31.2.2 that measured values of time constants are larger than computed values ($\tau_1$). The field poles of d-c motors often are not laminated, whereas high-quality servomotors should be. In both cases, however, eddy currents are created when the flux varies either by a variation of input signal $E_1$ or by a variation of load through armature reaction. These eddy currents have an effect equivalent to that of a short-circuited winding on the pole (Fig. 31-26). Let $\mathcal{R}$ be the reluctance of the magnetic circuit, $N_1$ the number of turns of the control winding, $N_2$ the number of turns of the short-circuited winding equivalent to the eddy-current effect, and $R_1L_1$ and $R_2L_2$ the resistances and inductances of these windings (Fig. 31-27). One then has,

$$L_1 = \frac{N_1^2}{\mathcal{R}} \quad L_2 = \frac{N_2^2}{\mathcal{R}}$$

Let $M = N_1N_2/\mathcal{R}$ be the mutual inductance and let $N_1/N_2 = n$. The equations for the flux $\Phi$ then become

$$E = (R_1 + L_1s)I_1 + \frac{L_1}{n}sI_2$$

$$0 = \left(\frac{R_2 + L_1}{n^3}s\right)I_2 + \frac{L_1}{n}sI_1$$

$$\Phi = k_1\left(I_1 + \frac{I_2}{n}\right)$$

Elimination of $I_1$ and $I_2$ gives the transfer function

$$\frac{\Phi}{E} = \frac{k_1 R_2}{R_1 R_2 + L_1 R_2 s + (L_1 R_1 / n^3) s}$$

Or taking

$$\tau_1 = \frac{L_1}{R_1} \quad \tau_2 = \frac{L_1}{n^3 R_2} = \frac{L_2}{R_2}$$

$$\frac{\Phi}{E} = \frac{k_1 R_2}{1 + (\tau_1 + \tau_2)s}$$

The time constant of the field winding is now $\tau_1 + \tau_2$, and the equations for the armature are not changed. The motor transfer function becomes then

$$\frac{\Theta_m}{E} = \frac{i_0 k_1 k_2}{R_1 \omega \sqrt{[1 + (\tau_1 + \tau_2)s](s^2 + 2\omega_n s + \omega_n^2)}}$$
31.3.3. Characteristic Curves. These characteristics are valid for the steady state.\(^1\) For this condition the inductances have now no influence and the above equations become \(C = k_1k_2\phi E/R\). The corresponding curves (Fig. 31-28) are straight lines parallel to the speed axis. Above a certain speed, the torque decreases because of such effects as saturation of the magnetic circuit and variation of the armature current.

31.3.4. Advantages and Disadvantages. \(a\). The characteristic curves show that this type of motor produces a torque that is independent of speed during steady-state operation. Electric motors operated in this way are thus similar to hydraulic motors. The latter, however, have a considerably lower inertia for the same nominal rating and are therefore superior in performance. Electric motors also have a constant ratio \(C/J\) over a large range of speeds, and they can therefore be used with advantage in servo systems for tracking moving targets. In fact, these systems require an acceleration (positive or negative) at the output shaft about an average speed which, although varying slowly, can take any value within a given range of speeds.

\(b\). The field-coil time constant is, in general, large compared to the time constants of other types of motor (motors with armature control, two-phase motors, hydraulic motors).

\(c\). The power required in the control field is low in general, and the amplifiers required can therefore be simplified. It has been shown previously that a control by tetrode or pentode can decrease the time constant of the field coil.

\(d\). The need for a constant-current source is an important drawback.\(^2\) Furthermore, because of the low total efficiency, it is sometimes difficult to remove the heat generated in the armature.

31.3.5. Variation 1: Series Motor with Double Field Coil (Split-series Motor). The motor has two field windings and there is a difference between field winding and pairs of poles. It has a double field winding if each pole has two independent windings. Each winding is permanently connected, and the armature is in series with the windings as indicated in Fig. 31-29. The amplification stage must be push-pull-operated; that is, the relation between the input signal \(e\) to the amplifier and the output currents \(I_1\) and \(I_2\) must be such that one has, at least for steady-state conditions,

\[
I_1 - I_0 = I_0 - I_2 = Ae
\]

\(^1\) The friction \(f\) has been lumped with the load; the equation given above represents the internal torque. Part of this torque is used in the motor itself (air-friction loss).

\(^2\) An important example of this type of control is found on Italian naval ships built before 1940. A constant-current supply fed all the armatures of the servos, and each had field-coil control. Metadyynes were used as control elements.
where \( I_0 \) is constant. The intensity in the armature is \( I = I_1 + I_2 = 2I_0 \). If the two circuits are identical except for the phase of winding, which is opposite, the total flux is the difference of the two fluxes. Thus,

\[
\Phi = \Phi_1 - \Phi_2 = 2A_k I_0 \cos \theta
\]

To a first approximation, this motor is equivalent to a motor with field control. A push-pull amplifier is required, but the constant-current source is automatically realized. This scheme is frequently used in low-power, medium-performance servos.

**Note.** If the amplifier delivers two voltages \( V_0 + E \) and \( V_0 - E \), the motor is not linear any longer.

**31.3.6. Variation 2: Motor with a Simple Field Winding, Constant Voltage Applied to the Armature.** This method is simpler to carry out because it is easy to feed a constant voltage to the armature. For these conditions, the operation of the motor is not linear. One then has the following equations (with \( k_4 = 0 \), zero load torque).

\[
E = (R + Ls)I \\
C = k_1k_4 I_1 = (J_s + f)\theta \Theta_m \\
\Phi = k_1 I \\
v_0 = (r + la)i + k_2 \theta \Theta_m
\]

Using the same notation as before, \( v_0 \) is the constant armature voltage, \( r \) and \( l \) are the resistance and inductance of the armature, and \( k_2 \) is the constant of back emf. From these equations, one obtains

\[
\frac{k_1k_2v_0}{(R + Ls)(r + la)} E + \left[ \frac{-k_1k_2k_3E}{(R + Ls)(r + la)} - (J_s + f)s \right] \Theta_m = 0
\]

The characteristic curves are given by the equation

\[
C = \frac{k_1k_2v_0}{Rr} E - \frac{k_1k_2k_3E}{Rr} \Omega_m
\]

The no-load speed is such that \( \Omega_m = v_0/k_2 \); it is independent of the input voltage \( E \). The curves (Fig. 31-31) are straight lines through a fixed point. Similar characteristics will be obtained for a two-phase motor (see Sec. 31.6).

**31.4. ARMATURE-CONTROLLED MOTORS**

**31.4.1. Principle.** The field winding (Fig. 31-32) carries a constant current \( I_0 \). This presents no problem in steady-state operation, since no
back emf is induced in the field winding. The armature reaction produces a field the direction of which is fixed with respect to the coils and the magnitude of which depends on the armature current. This field is 90° out of phase with respect to the field flux. The control voltage \( v \) (input variable) is fed to the armature. These quantities are shown in Fig. 31-32. The field flux \( \Phi_0 \) is constant.

**31.4.2. Transfer Function.** Using the same notation as in Sec. 31.1.4, the following equations hold for the motor:

\[
v = (r + ls)i + k_2 sl \theta_m \quad C = k_3 \Phi_0 i
\]

\[
C = J \frac{d^2 \theta_m}{dt^2} + f \frac{d \theta_m}{dt} + k_4 \theta_m \quad k_2 = 10^{-8} \frac{N}{2\pi} \Phi_0
\]

\[
k_3 = 10^{-8} \frac{N}{2\pi} \quad \text{(that is} \quad k_2 = k_3 \Phi_0)\]

From these equations, the transfer function for a motor with armature control is obtained as

\[
\frac{\Theta_m}{v} = \frac{k_3 \Phi_0}{Js^3 + (Jr + fl)s^2 + (k_4d + rf + k_2k_3 \Phi_0)s + rk_4}
\]

**PARTicular Case.** With no restoring torque at the output (\( k_4 = 0 \)), the transfer function becomes

\[
\frac{\Theta_m}{V} = \frac{K}{s \left(1 + \frac{2r}{\omega_n} + \frac{s^2}{\omega_n^2}\right)}
\]

\[
K = \frac{k_3 \Phi_0}{k_2k_3 \Phi_0 + rf} \quad \omega_n = \left(\frac{k_3k_2 \Phi_0 + rf}{Jl}\right)^{\frac{1}{2}} \quad 2\zeta = \frac{Jr + lf}{[Jl(k_3k_2 \Phi_0 + rf)]^{\frac{1}{2}}}
\]

If also \( f = 0 \), then since \( k_2 = k_3 \Phi_0 \), one obtains

\[
K = \frac{1}{k_2} \quad \omega_n = \frac{k_2}{(Jl)^{\frac{1}{2}}} \quad 2\zeta = \frac{r}{k_2} \left(\frac{J}{l}\right)^{\frac{1}{2}}
\]

The transfer loci are given in Figs. 8-16, 13-4, and 13-5. The two time constants \( \tau_1 \) and \( \tau_2 \) occur only if \( \zeta > 1 \). However, when the motor is used in a servo link, a negligible phase shift is desired in the speed control (except for integration). Then \( \omega_n \) should be large compared to these frequencies. The phase shift introduced by the motor is then given by the time constant

\[
\tau = \frac{2\zeta}{\omega_n} = \frac{Jr + fl}{k_2k_3 \Phi_0 + rf}
\]

For \( f = 0 \), this equation simplifies to \( \tau = Jr/k_2^2 \).
The passband for the motor is characterized by $\omega_n$. Other things being equal, it is advantageous to minimize $J$ and $l$ in servo systems.

**31.4.3. Characteristic Curve.**
The static characteristics between motor torque and speed of rotation $\Omega_m$ will be considered for all values of armature voltage $v$. In steady-state operation,

$$v = ri + k_2 \Omega_m$$

$$C_m = k_3 \Phi_0 i$$

From the above, one obtains the curves (Fig. 31-33) showing the maximum torque (starting) $C_M$ and the maximum speed (no load) $\Omega_M$:

$$C_m = C_M - \frac{C_M}{\Omega_M} \Omega_m$$

$$\Omega_M = \frac{v}{k_2}$$

$$C_M = \frac{k_3 \Phi_0 v}{r}$$

with

$$C_M = \frac{\alpha J_m + J_c/\alpha}{C_M} \Omega_c$$

For the case $k_4 = f = 0$, that is, zero load torque and no viscous friction on the output shaft (Fig. 31-34),

$$\tau = \frac{2r}{\omega_n} = \frac{J \Omega_M}{C_M} = \frac{\alpha J_m + J_c/\alpha}{C_M} \Omega_c$$

**31.4.4. General Case for Opposing Torque.** The equation for torques on the output shaft becomes

$$C = J \frac{d^2 \theta_r}{dt^2} + f \frac{d \theta_r}{dt} + C_r(t, \theta_r)$$

where $\alpha = 1$, for simplification.

It is possible to express $\Theta_r(s)$ in terms of $v_e(s)$ and $C_r(s)$:

$$\Theta_r(s) = \frac{k_3 \Phi}{(r + ls)(Js^2 + fs) + k_3 k_5 \Phi s} v_e(s)$$

$$- \frac{r - ls}{(r + ls)(Js^2 + fs) + k_3 k_5 \Phi s} C_r(s)$$
The transfer matrix is obtained from the two equations (Fig. 31-35)

\[ v_a = \frac{(r + l_s)(J_s + f) + k_2k_3 \dot{\Omega}_r}{k_3 \Phi} + \frac{r + l_s}{k_3} C_r \]

\[ i_a = \frac{J_s + f}{k_3 \Phi} \Omega_r + \frac{1}{k \Phi} C_r \]

### 31.4.5. Comparison between the Two Types of Motors

Numerical values indicate that, in general, natural frequencies \( \omega_n \) are greater for armature-controlled than for field-controlled motors. However, this difference should not lead to hasty conclusions.

1. From the energy standpoint, the field of an armature-controlled motor does not deliver any power; it could as well be replaced by a permanent magnet. The mechanical energy received at the shaft is therefore inferior to the electrical energy supplied to the motor by the amplifier. In this respect, it may be worthwhile to recall the formula expressing the efficiency of an armature-controlled motor. Proceeding from the equations derived previously:

\[ v = ri + k_2 \Omega_m \quad C_m = k_2i \]

The electrical power \( P_E \) supplied to the motor is found to be

\[ P_E = \frac{v^2}{r} - k_2 \frac{v}{r} \Omega_m \]

while the mechanical power received by the shaft is

\[ P_m = \Omega_m C_m = \frac{k_2 \Omega_m}{r} (v - k_2 \Omega_m) \]

The energy efficiency, defined by the relation

\[ \eta = \frac{\text{mechanical power on the shaft}}{\text{electrical power supplied}} \]

is then expressed by

\[ \eta = \frac{k_2^2 \Omega_m}{v} \]

that is, with \( \Omega_M = v/k_2 \)

\[ \eta = \frac{\Omega_m}{\Omega_M} \]

Curves representing \( P_E, P_m, \) and \( \eta \) versus \( \Omega_m \) are reproduced in Fig. 31-36. The energy efficiency of the motor is 0.5 in the vicinity of the maximum mechanical power. An armature-controlled motor behaves essentially as a transformer of electrical energy into mechanical energy, with an efficiency inferior to 1.

2. Consider now a field-controlled motor. Let us add the equation of the field circuit to those previously derived:

\[ V = ri + l \frac{di}{dt} + v \quad \text{with} \quad v = \text{back emf of the armature} \]
The steady-state equations read as follows,

\[
E = RI \quad \Phi = kI \quad C_m = k'i_0\Phi \\
V = ri + v \quad \text{with} \quad v = k_2\Omega_m \quad \text{and} \quad k_2 = k'i_0
\]

The mechanical power on the shaft is expressed by

\[
P_m = \Omega_mC_m = \Omega_mk'i_0\Phi_i_0
\]

while the control power \( P_E \), the control voltage across the field being constant, is

\[
P_E = \frac{E^2}{R} = \text{const}
\]

The power \( P_e \) in the armature circuit is

\[
P_e = V i_0 = ri^2 + k_2\Omega_m i_0
\]

which contains a constant term \( ri^2 \) and a term proportional to the speed.

\[\text{Fig. 31-36. Armature-controlled motor.}\]

The ratio

\[
\eta = \frac{\text{mechanical power on the shaft}}{\text{electrical control power (field)}}
\]

can exceed unity:

\[
\eta = \frac{P_m}{P_E} = \frac{k'i_0\Phi_i_0R}{E^2} \frac{\Omega_m}{\Omega_m}
\]

The mechanical energy does not come only from the control energy; the greater the speed, the smaller that part is. The motor behaves then as an energy amplifier associated with an energy transformer (electrical or mechanical).

3. As a consequence, a direct comparison of the time constant of field-controlled and armature-controlled motors is of no value, for such a comparison would apply to physical systems having different roles. The only valid comparison could be made between (1) a field-controlled motor and (2) an armature-controlled motor plus a power amplifier. As will be seen in Chap. 33, the presence of an amplifier introduces milliseconds.
31.5. ALTERNATING-CURRENT MOTORS

31.5.1. Introduction. The operation of almost all a-c motors is based on induction phenomena due to rotating fields. One exception, however, is the low-power (0 to 100 watts) low-efficiency universal motor which is of no interest in servomechanism applications.

![Diagram of Squirrel-cage-type motor](image)

Fig. 31-37. Squirrel-cage-type motor.

It is well known that a field varying sinusoidally along a fixed direction, say \( H = H_0 \sin \omega t \), may be considered to be the result of two fields of magnitude \( H/2 \) rotating in opposite directions with angular speeds \( \omega \) and \(-\omega\). This concept enables the operation of a single-phase induction (Fig. 31-37) motor to be easily explained. Once brought up to a sufficiently high speed in one or the other direction, the motor will develop only small slip and therefore high torque with respect to one of

![Diagram of single-phase induction motor](image)

Fig. 31-38.

the rotating fields, and a large slip and therefore negligible torque with respect to the other field.

The single-phase induction motor is not self-starting because no net torque can be developed at a speed that is less than the synchronous speed. For this reason the motor cannot be used in the power stage of a servomechanism.

Next to the single-phase induction motor, the simplest induction motor is of the two-phase type, the operation of which may be explained as follows (Fig. 31-38): Two currents, out of phase by \( \pi/2 \), are fed to two field circuits, 1 and 2, which are assumed to be identical and out of phase
by 90°, so as to produce the two orthogonal magnetic fields $H_1$ and $H_2$:

$$H_1 = H_0 \cos \omega t \quad H_2 = H_0 \cos \left(\omega t + \frac{\pi}{2}\right)$$

The resulting field is represented by a phasor having a magnitude $H_0$ and an angular speed $\omega$.

Note. When low-power motors (up to 500 watts) are under consideration, a single-phase supply may be used. One field coil is fed directly while the other is fed through a capacitance that is in series with it. As a result, the current in the latter inductor leads by $\pi/2$ (Fig. 31-39).

Consider now (Fig. 31-40) a two-phase motor in which a device $A$ enables control of the magnitude of the rotating field. The fields $H_1$ and $H_2$ are again orthogonal and out-of-phase in time by $\pi/2$. The magnitude of the field produced by inductor 2 is $H_0$, but that produced by inductor 1 now has a magnitude $kH_0$, where $k$ is a parameter usually found between 0 and 1.† Let the stationary fields $H_1$ and $H_2$ be resolved into rotating components as was done for the single-phase motor (Fig. 31-41). These components may be added to produce two fields rotating in opposite directions with speeds $\omega$ and $-\omega$. Their magnitudes are respectively

$$\frac{H_0}{2} (k - 1) \quad \text{phasor } OD.$$  
$$\frac{H_0}{2} (k + 1) \quad \text{phasor } OS.$$  

This shows, therefore, that an elliptical rotating field may be resolved into two circular fields rotating in opposite directions. Moreover, for the case of Fig. 31-40, the magnitudes of these two fields are linear functions of the gain of the control device. One sees, therefore, that such a motor could profitably be used in a servomechanism, since (a) it is likely to be self-starting because the magnitudes of the two fields may not be

† When the control device $A$ is also a phase inverter, it may be negative. For simplicity it will be assumed that $k > 0$. 
equal and (b) the induction phenomena due to the two fields, as apparent from the torque acting on the rotor (Lenz’s law), have a nonzero resultant which may be controlled by a single device $A$ acting on a single phase.

The above fact shows that the use of a polyphase (three- or six-phase) supply is no longer advantageous, since this would require several devices similar to $A$. However, the possibility still remains. A three-phase supply requires three conductors when carefully balanced, and four when not. The magnetic field produced by three balanced inductors, which are out of phase by 120° in time and in space, is a single rotating field. In order that such a motor be useful, it is necessary to use three perfectly balanced and synchronized control devices. In addition, if the magnitudes of the fields produced by the three inductors are unbalanced, the three stationary fields may not be easily resolved into rotating fields, and the linear interrelation will then no longer be present. Moreover,

![Diagram](image)

Fig. 31-41.

the load on the supply becomes unbalanced for such a condition, and this is a feature which is undesirable.

The above introduction has shown that for servomechanism applications, the most satisfactory a-c motor is the two-phase motor operating from a single-phase supply (this again is true when low powers, e.g., up to 500 watts, are involved). The discussion has, of course, been restricted to the field of servomechanisms. In all other fields, the three-phase squirrel-cage motor is the one that is most frequently encountered except for electric traction. However, recent achievements by the French railroad system have shown that the industrial 50-cps supply may be preferable, with either rectifiers or universal motors.

Note. It can be shown that the armature reaction field rotates at the same speed as the primary field, whereas the former is stationary in a d-c machine.

31.5.2. Two-phase Motors. 1. Construction. It is apparent from what has just been said that it is possible to use two-phase induction motors in servomechanism applications. However, classical induction-motor design technique, with copper bars\(^1\) inserted in a steel rotor, results in a large inertia, whereas a motor that is to be used in a servomechanism

---

\(^1\) Or aluminum cast in a perforated steel rotor.
must have minimum inertia. To achieve this objective, shell or bell-type rotors are often used. They are supported by a cantilever with a double air gap. Inside the bell there is a fixed soft-iron cylinder which is attached to the motor frame (Fig. 31-42).

The realization of the shell is not easy, for it must have very low inertia and a fixed resistivity and it must transmit the drive without distortion. In some cases, the shell is obtained by means of an electrolytic deposit. As in all motors, the clearance between the moving and the stationary parts is limited only by mechanical considerations, allowance being made for uneven heating and possibly gyroscopic torques. In such motors, then, the air gap must have at least twice the thickness of the gaps found in conventional motors of equivalent size. To this must be added the thickness of the bell, which is never ferromagnetic. The power of these motors is always small compared to that of the industrial-type motors.

2. Transfer Functions. The transfer function of a motor may be obtained from its characteristic curves if they are parallel and equidistant (constant-gradient) straight lines. It will be seen later that, while the first condition may be satisfied for a relatively wide speed range, the second is more difficult to obtain. The gradient is a function of the control factor $k$, and linearization can be accomplished only in a small region $(C,\Omega)$ (Fig. 31-43).

In such a region, the transfer function of an unloaded motor may, as before, be expressed by

$$\frac{\Theta_m}{E} = \frac{|\partial\Omega/\partial E|}{s(Ts + 1)}$$

with $T = J \left| \frac{\partial \Omega m}{\partial C} \right|$

where $J$ is motor inertia.
3. Notes. a. Because of the low rotor inertia, it may be assumed that \( Ts \ll 1 \) for most of the frequency range. The motor then behaves like a perfect integrator.

b. As mentioned in Sec. 29.4.3, these motors make very good a-c tachometers. One of the field coils is fed from a reference supply, and the other generates a voltage of the same frequency and of amplitude which is strictly proportional to the speed of rotation. This occurs over a range of speeds that is almost twice that which the tachometer would have as a two-phase motor. It must be realized, however, that this voltage is small, but it is much easier and simpler to amplify an a-c voltage than a direct voltage.

4. Characteristic Curves. The shape of the characteristic curves of actual motors will now be established. First, consider in detail the various fields found in a motor. For reasons already given, the analysis will be limited to a machine with a single rotating field.

Let \( 2g \) be the number of poles and \( \omega/q \) the angular speed of the rotating field, \( \omega \) being the frequency of the a-c supply. In addition, let \( \Omega \) be the angular speed of the motor, which is always less than \( \omega/q \). The angular frequency \( \omega' \) of the currents induced in the rotor is

\[
\omega' = \frac{\omega}{q} - \Omega
\]

It is customary to define the slip \( g \) by the expression

\[
g = \frac{\omega/q - \Omega}{\omega/q}
\]

and therefore \( \omega' = g\omega/q \).

The rotor currents in turn produce a field called the armature-reaction field. Its angular speed is \( g\omega/q \) with respect to the inductor (stator). It is \( \omega'' = g\omega/q + \Omega \) with respect to the rotor, which rotates at speed \( \Omega \). Consequently, with respect to the stator \( \omega'' = \omega/q \). Therefore, the armature-reaction field rotates at the same speed as the primary (stator) field, regardless of the slip.

31.5.3. Additional Considerations Regarding the Real Characteristic Curves. In practice there are several parameters, and by proper choice one may, as will be seen, construct a motor that is linear over a sufficiently wide range of \( C \) and \( \Omega \). Let it now be assumed that it is desirable to determine the shape of the characteristic curves for an induction motor with a single circular rotating field.

Let \( \omega = \) supply frequency

\( 2g = \) number of poles

\( \Omega = \) rotor speed

\( \omega' = \omega/q = \) synchronous speed

\( g = \) slip defined as \( g = \frac{\omega/q - \Omega}{\omega/q} \)

The above remarks lead to the concept that motors may be considered as transformers. Then let

\( V_1, I_1, R_1, L_1 = \) primary voltage, current, resistance, and inductance

\( I_2, R_2, L_2 = \) secondary current, resistance, and inductance

\( M = \) mutual inductance
One has, therefore,

\[ V_1 = R_1 I_1 + L_1 s I_1 + M s I_3 \]
\[ 0 = R_3 I_3 + L_3 s I_3 + M s I_1 \]

The resistance \( R_1 \) may in general be assumed to be negligible. The total leakage inductance referred to the secondary is then

\[ N_3 = L_3 - \frac{M^2}{L_1} \]

The real power is, therefore,

\[ P_1 = V_1 I_1 = \left( \frac{M}{L_1} \right)^2 V_1^2 \frac{R_2/g}{(R_2/g)^2 + N_3^2 \omega^2} \]

The real power transferred to the secondary,\(^1\) equal to \( P_1 \) since \( R_1 \) is assumed to be negligible, includes the joule loss \( P_3'' = R_3 I_3^2 \) and the mechanical power,

\[ P_3'' = \frac{P_3 I_3^2 (1 - g)}{g} \]

But the speed of rotation is \( \Omega = \omega (1 - g)/q \), so that the torque \( C \) may be expressed as

\[ C = \frac{\alpha R_3 I_3^2}{\omega g} \]

The total real power, which may be checked,

\[ P_1 = P_1' + P_3'' \]

is equal to the product of the synchronous speed and the torque. One may write

\[ C = \left( \frac{M}{L_1} \right)^2 V_1^2 \omega \frac{R_2/g}{q (R_2/g)^2 + N_3^2 \omega^2} \]

This expression defines the characteristic curves, since the slip and rotor speed are related as follows:

\[ g = \frac{\omega/q - \Omega}{\omega/q} \]

Taking the secondary (rotor) resistance as a variable parameter,\(^2\) the torque will be a maximum for \( R_2/g = N_3 \omega \).

It may be shown in a similar manner that the joule loss in the secondary (rotor) is equal to the product of the useful torque and the slip (or difference between synchronous and rotor speed).

\(^1\) To find \( P_1 \), one begins with the secondary voltage

\[ E_3 = \frac{R_3}{g} I_3 + L_3 s I_3 \]

\(^2\) For large motors, which must be started under load, a wound rotor and slip rings are used. The rings are connected to variable resistances (one per phase) so that the rotor resistance may be varied when the motor is operating. In general, these external resistances are inserted during starting and are then progressively removed.
In Fig. 31-44 a family of characteristic curves corresponding to an exciting voltage $V_1$ and various values of rotor resistance is shown. In Fig. 31-45, the same curves are plotted for one value of rotor resistance and several exciting voltages increasing in arithmetic progression.

![Fig. 31-44.](image)

![Fig. 31-45.](image)

![Fig. 31-46.](image)

It must be remembered that the above curves are for a motor with a single rotating field. A few conclusions will now be drawn before returning to the unbalanced two-phase motor.

a. If the rotor resistance is so chosen that $R_2/g = N \omega_0$, maximum torque will be developed upon starting. This condition is in general artificially met when starting
industrial-type motors. In a linear servomechanism, one will always choose \( R_s > N_m \omega \), so that the maximum torque will occur slightly to the left of the torque \( C \) axis.

The useful part of the characteristics between \( \Omega = 0 \) and \( \Omega = \omega/\alpha \) is then slightly curved upward. It may be, and very often is, approximated by a straight line.

b. The above hypothesis does not, however, give a linear motor because, for different values of \( V_1 \), the lines all pass through the point \( (\omega/\alpha, 0) \), and moreover their gradient is not constant (Fig. 31-46).

c. The above study has omitted losses due to friction, hysteresis, eddy currents, and windage.

Consideration of the losses in (c) improves the curves given thus far. Let it be assumed that the losses are governed only by a function of the type \( \alpha \Omega \), where \( \alpha \) is a constant (viscous friction). On starting with the curves of Fig. 31-47, new ones are obtained by means of the construction indicated and are shown by broken line.

The curves in this new family converge on a point \( E' \) located on a line parallel to the torque \( C \) axis passing through \( B \) in such a manner that \( BE = BE' \). Point \( E \) must then necessarily be below point \( B \). As a result, the point of convergence of the new family of curves is then further from the useful part of the torque axis than it
would be if viscous friction were neglected (Fig. 31-47). It is worth noting that any viscous friction outside the motor proper will yield the same effects and improve the motor linearity.

Now consider again the unbalanced two-phase motor in which the torque may be considered to be due to two rotating fields having the magnitudes $H_0(k + 1)/2$ and $H_0(k - 1)/2$ (Sec. 31.5.1) and moving in opposite directions. At start, the slip is unity for each field. The resulting torque is the algebraic sum of the torques due to each field. The magnitudes of the rotating fields are proportional to the corresponding exciting voltages, provided saturation does not occur. Since it has already been seen

\[
C_0 = \lambda \frac{H_0^2}{4} [(k + 1)^2 - (k - 1)^2]
\]

where \( \lambda \) is a constant. Thus \( C_0 = k\lambda H_0^2 \), and the starting torque is proportional to the control factor \( k \).

Moreover, the slope of these real characteristics would be zero near \( \Omega = 0 \) if the maximum torque for a balanced motor were located on the axis. The slope would then increase rapidly. To avoid this break, it is customary to position the point of maximum torque slightly to the left of the torque axis. The resulting characteristics (Fig. 31-48) have a negative slope at low speeds and a relatively broad speed range, for which the motor is almost linear. It must also be stated that the synchronous speed is never attained; it is, in general, very much above the actual speeds at which the motor will rotate when used in a servomechanism (Fig. 31-48). As an example,
the experimental characteristic curves of an unloaded two-phase motor (SAGEM two-phase motor, 400 cps, type 25-400-4) are shown in Fig. 31-49. It is seen that it is correct to assume that the motor is linear in the region indicated. The size of this region can be increased at high speeds by coupling the motor to a device having viscous friction (Fig. 31-49).

In conclusion, two-phase motors may in general be assumed linear for all voltages applied to the control phase and for speeds less than one-third of synchronous speed. The efficiency and power per pound of these motors are low. However, because of their low inertias, they have good acceleration characteristics.

31.6. NUMERICAL DATA FOR ELECTRIC SERVOMOTORS

Tables 31-1 to 31-6 present numerical data for some American\(^1\) commercial servomotors.

<table>
<thead>
<tr>
<th>Table 31-1. Two-phase Diehl Motor*</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Types FPE25-11, FPF85-18-1, and ZP143-2256-1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FPE25-11</th>
<th>FPF85-18-1</th>
<th>ZP143-2256-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class, a-c, two-phase:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency, cps</td>
<td>60</td>
<td>60</td>
<td>400</td>
</tr>
<tr>
<td>Maximum power, watts</td>
<td>5</td>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>Time constant, (\uparrow) msec</td>
<td>14.3</td>
<td>16.7</td>
<td>179</td>
</tr>
<tr>
<td>Type:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of poles</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Electrical data:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference coil:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permanent exciting voltage, volts</td>
<td>115</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>Permanent intensity, mA</td>
<td>120</td>
<td>470</td>
<td>4,250</td>
</tr>
<tr>
<td>Control coil:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum voltage, volts</td>
<td>115</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>Stall impedance (modulus), ohms</td>
<td>960</td>
<td>240</td>
<td>27.0</td>
</tr>
<tr>
<td>Real component, ohms</td>
<td>435</td>
<td>45</td>
<td>4.4</td>
</tr>
<tr>
<td>Imaginary component, ohms</td>
<td>856</td>
<td>263</td>
<td>26.6</td>
</tr>
<tr>
<td>Maximum temperature rise at 60 cps, °C</td>
<td>65</td>
<td>65</td>
<td>55</td>
</tr>
<tr>
<td>Mechanical performances:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight of the motor, g.</td>
<td>635</td>
<td>5,220</td>
<td>14,700</td>
</tr>
<tr>
<td>lb</td>
<td>1.4</td>
<td>11.5</td>
<td>32.5</td>
</tr>
<tr>
<td>Inertia of the armature, g-cm²</td>
<td>18</td>
<td>390</td>
<td>4,000</td>
</tr>
<tr>
<td>oz-in. (\dagger)</td>
<td>0.098</td>
<td>2.15</td>
<td>22.00</td>
</tr>
<tr>
<td>Theoretical acceleration when starting, (\dagger) rad/sec²</td>
<td>22,000</td>
<td>17,000</td>
<td>3,600</td>
</tr>
</tbody>
</table>

\* Diehl is the trademark for Diehl Manufacturing Co., Somerville, N.J.
\(\uparrow\) Defined by inertia/viscous friction.
\(\dagger\) Defined by stalling torque/inertia.

### Table 31-2. Ford Induction Motor

<table>
<thead>
<tr>
<th>Type</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>75-013</td>
<td>75-291</td>
</tr>
</tbody>
</table>

**Class, a-c, two-phase:**

- Frequency, cps: 60, 400
- Maximum power (on the shaft), watts: 11, 14
- At rpm: 1,800, 4,200

**Type:**

- Number of poles

**Electrical data:**

- Continuous voltage, volts: 115, 115
- D-c resistance, ohms: 34, 32

**Mechanical data:**

- Weight of motor, kg: 1.95, 1.95
- Inertia of rotor, g-cm²: 40, 51
- Torque stalled, g-cm: 1,180, 590
- Speed at no load, rpm: 3,400, 7,800

*Ford Instrument Company, Division of Sperry Rand Corporation, Long Island City 1, N.Y.*

**Note.** The company has developed a low-inertia servomotor which can be directly connected to the last-stage vacuum-tube amplifier (high voltage).

### Table 31-3. Kollsman Induction Motor

(Types 1515-0100 and 1805-0460)

<table>
<thead>
<tr>
<th>Type</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1515-0100</td>
<td>1805-0460</td>
</tr>
</tbody>
</table>

**Class, a-c, two-phase:**

- Frequency, cps: 60, 400
- Maximum power (on the shaft), watts: 0.5, 2
- Time constant, sec: 0.05, 0.04

**Type:**

- Number of poles: 2

**Electrical data:**

- Continuous voltage, volts: 15, 115
- Current, ma: 130, 210
- Resistance, ohms: 76, 103

**Phase 2 (control phase):**

- Continuous voltage, volts: 20, 115
- Current, ma: 200, 210
- Resistance, ohms: 76, 103

**Temperature rise at no load, °C:**

- 43, 80

**Motor stalled, °C:**

- 41, 125

**Mechanical data:**

- Weight of motor, g: 65
- Inertia of rotor, g-cm²: 2.44, 6.7
- Torque stalled, g-cm: 18, 135
- Speed at no load, rpm: 3,000, 11,500
TABLE 31-4. KOLLMAN INDUCTION MOTOR
(Type 2103-0410-0)

Class, a-c, two-phase:
Frequency ......................................................... 400 cps
Maximum power (on the shaft) ................................ 0.5 watts
Time constant ...................................................... 8 sec

Type, induction:
Number of poles .................................................. 6

Electrical data:
Phase 1 and phase 2:
Voltage ............................................................... 26 volts
Impedance ........................................................ 86 + j169 ohms
Current:
No load ............................................................ 137 ma
Stall ................................................................. 143 ma

Temperature rise of control phase (reference phase 26 volts):
26 volts at no load ................................................ 25°C
26 volts at stall ................................................... 30°C
8 volts at stall .................................................... 27°C

Mechanical data:
Weight of motor .................................................... 75 g
Inertia of rotor .................................................... 1.0 g-cm²
Torque stalled ..................................................... 22.5 g-cm
Speed at no load .................................................. 6,500 rpm
Acceleration at stall ............................................. 20,500 rad/sec²

TABLE 31-5. MIDWESTERN INSTRUMENTS TORQUE MOTOR
(Model 9)

Class, d-c to 250 cps (3 db):*

Power:
Quiescent, 20 ma .................................................. 2.6 watts
Maximum force, 40 ma ........................................... 5.3 watts

Type, permanent magnet:
Energizing coils .................................................. 2 each

Electrical characteristics:
Coil:
Resistance (each) ............................................... 3,300 ohms
Inductance at 1 kcps .......................................... 11 henrys

Current:
Balance, each coil ............................................... 20 ma
Full stroke* ....................................................... 20 ma†
Full force .......................................................... 40 ma†
Hysteresis, maximum ........................................... 2%

Mechanical characteristics:
Midposition force ............................................... 7.5 lb
Moment of inertia* ............................................... 64 × 10⁻⁴ in.-lb-sec²
Output radius .................................................... 0.906 in.
Resonant frequency* ............................................ 300 cps
Stroke ............................................................. 0.015 in.
Weight ............................................................. 21 oz

English Metric

<table>
<thead>
<tr>
<th>Midposition force</th>
<th>7.5 lb</th>
<th>3.4 kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moment of inertia*</td>
<td>64 × 10⁻⁴ in.-lb-sec²</td>
<td>0.074 cm-g-sec²</td>
</tr>
<tr>
<td>Output radius</td>
<td>0.906 in.</td>
<td>2.3 cm</td>
</tr>
<tr>
<td>Resonant frequency*</td>
<td>300 cps</td>
<td>300 Hz</td>
</tr>
<tr>
<td>Stroke</td>
<td>0.015 in.</td>
<td>0.038 cm</td>
</tr>
<tr>
<td>Weight</td>
<td>21 oz</td>
<td>595 g</td>
</tr>
</tbody>
</table>

* Unloaded armature.
† Differential, or in one coil.
### Table 31-6. Midwestern Instruments Torque Motor (Model 11)

#### Class, d-c to 325 cps (3 db):

**Power:**
- Quiescent, 20 ma: 1.1 watt
- Maximum force, 40 ma: 2.24 watts

#### Type, permanent magnet:
- Energizing coils: 2 each

#### Electrical characteristics:
- **Coil:**
  - Resistance (each): 1,444 ohms
  - Inductance at 1 kcps: 2.7 henrys
- **Current:**
  - Balance, each coil: 20 ma
  - Full stroke*: 20 ma†
  - Full force: 40 ma†

#### Hysteresis:
- Maximum: 3% English Metric

#### Mechanical characteristics:
- Midposition force: 4.5 lb 2.0 kg
- Moment at inertia*: 27 × 10⁻⁵ in.-lb-sec² 0.031 cm-g-sec²
- Output radius: 0.75 in. 1.9 cm
- Resonant frequency: 425 cps 425 Hz
- Stroke: 0.008 in. 0.020 cm
- Weight: 14 oz 397 g

* Unloaded armature.
† Differential, or in one coil.
CHAPTER 32

HYDRAULIC SERVOMOTORS

Summary
1. Description of hydraulic motors.
2. Summary of the fundamental laws of hydrodynamics.
3. Transfer function of a hydraulic transmission.
4. Establishing and discussing the equations of a motor controlled by a valve.
5. Servocontrols and boosters.
6. Additional data.
7. Motor data sheets.

32.1. DESCRIPTION OF HYDRAULIC MOTORS

32.1.1. General Considerations. Hydraulic motors have been used for a long time, but the theoretical investigation of their characteristics has been started only recently, mostly because of their incorporation into servomechanisms.

Actuators supplied by a fluid under pressure (oil or air) are widely used to power the autopilots of airplanes (Siemens K-12 and Alkan autopilots) or of guided missiles (V-2 and present-day missiles), gun turrets, the controlling valve of hydraulic turbines, etc.

One reason for selecting hydraulic power for actuators is that the possible power density of high-pressure oil lines is extremely high. The fluid-transmitted energy can be transformed into mechanical energy within small-size motors whereof the constituent material (usually steel) is worked at high stress values. (It may be noted that, in electric motors, iron is necessary in order to allow a proper distribution of the magnetic flux in the armature, but the useful stresses, i.e., the tangential forces, in that iron are extremely low with respect to the ultimate permissible loads.)

The fluids generally used are mineral oils, even though their viscosity coefficients are quite sensitive to temperature changes. The fluid used should not be considered as incompressible. Also, together with the compressibility of the fluid itself must be considered the compressibility of mixed or dissolved air and the effect of the expansion of the piping (flexible pipes). It will be shown that the performance of a servomechanism is evaluated with respect to "apparent compressibility," which takes into account the above-mentioned phenomena. The effect of dissolved air is particularly harmful, as can be easily shown. The compressibility coefficient of a pure liquid (containing no dissolved gases) is defined by

\[
\frac{\Delta V}{V} = \frac{1}{B} \Delta P
\]
where $\Delta V$ is the variation of a given volume $V$ under an increment of pressure $\Delta P$. It is to be noted that the bulk modulus, the coefficient $B$, has the dimensions of a pressure. Its order of magnitude is $10^{10}$ baryes (cgs).

Consider, now, that, at the atmospheric pressure $P$, a volume $v$ of air has been dissolved in the volume $V$ of liquid. For a pressure increment $\Delta P$, the total volume $V + v$ contracts by an amount $\Delta(V + v)$. The total of the respective variations of the volume of oil and air is

$$\Delta(V + v) = \left(\frac{V}{B} + \frac{v}{P}\right) \Delta P$$

It follows that, if $v = 0.001V$, the term $v/P$, decreasing for increasing pressure, becomes equivalent to the term $V/B$ for a pressure $P = 10^7$ baryes, i.e., about 10 kg/cm² ($B$ has been taken as $10^{10}$ baryes). Very often, though, the percentage of dissolved air is more than 0.1 per cent, particularly if the circuit has been operating for a certain time. The oil, therefore, must be degassed and frequently replaced.

32.1.2. Various Types of Hydraulic Motors. 1. Motors with One Master Cylinder (Linear Actuator). Assuming a power supply (high-pressure, or HP, oil), the flow delivered by the source is controlled by a distributing valve. The HP oil is directed toward one side or the other of the piston (Fig. 32-1 shows a four-way valve). The loads on the pistons $A$ and $B$ (rigidly connected together) are very small; they depend on the viscosity of the fluid and on the flow pattern. Pistons that have been designed recently have a very low resistive force with the portholes partially open, the magnitude of the force being independent of the displacement (Sec. 32.6.1).

It is sometimes considered that the combination master cylinder and valve are the motor and its power amplifier. The term amplifier is, however, limited to elements having the same type of input and output quantities (Sec. 33.1.1). In the case of an actuator, the input of the valve is a linear position and the output is a flow. Such a system will be called a linear actuator with controlling valve.

---

1 Linear here means rectilinear.
2 The distributor is not a part of the motor; it is the motor's controlling member.
2. Motors with One Single-acting Master Cylinder. Besides the double-acting device shown in Fig. 32-1 (Fig. 32-2 presents the details of the controlling four-way valve), some others are the so-called single-acting systems. On one side of the piston, a constant pressure is maintained. The valve controls the pressure on the other side of the piston exclusively (Fig. 32-3). Side B of the piston is subjected to a constant pressure equal to half the normally used pressure of the source. The pressure on side A varies between zero and \( P \), depending on the displacement of the controlling valve. As, in general, there is only one high-pressure source, supplying oil at a pressure \( P \), a constant-leak device can be used to drop the pressure from \( P \) to \( P/2 \). Such a device is shown in Fig. 32-4.

Incidentally, the connecting rod of the actuator must be calculated under buckling critical conditions. It is a beam with one clamped end, the other end being held by a ball joint. The critical case is when the
piston is at the maximum displacement toward the rod seal. The support given by the packing may have to be taken into account (Sec. 32.6.4). The obvious defects of such a device, particularly in transient conditions (i.e., whenever an oil flow has to go through the pipes) can be corrected by the use of a piston having "differential areas." Figure 32-5 represents a commonly adopted solution (unequal-area pistons) which has an evident disadvantage: the \textit{inertia} of the piston is higher than in the previous solution sketched in Fig. 32-3. On the other hand, however, there is no need for a reduced-pressure power supply.

3. Rotating Motors: Gear Motors and Vane Motors. These motors are more commonly known as \textit{pumps} (e.g., the oil pumps of most aircraft engines). Moreover, they are of the constant-displacement type (Figs. 32-6 and 32-7). Such motors are not widely used in servomechanisms for many reasons, one being the comparatively low permissible pressure supply. Accordingly, the volume must be quite large for a given maximum torque; it will be seen later that the time constant of a hydraulic motor depends on the volume of oil in the operating circuit; the larger the volume, the worse the time response.

4. Radial Piston Motors. A motor of such type is schematically shown in Fig. 32-8. Many similar designs have been effected. The cam is usually so shaped that the acceleration of the rollers (and of the connected pistons) is constant. The cam is, therefore, made of segments of parab-
olae. Unless special means are provided, the volume per revolution is constant. The oil input and exhaust are controlled by ports located near the center. They are cyclically connected to the oil tank (low-pressure, or LP, line) and to the oil collector (HP line).

Very often, such motors are used in servomechanisms, though axial-cylinder motors (see below, Sec. 32.1.3) have better characteristics. The inertia of moving parts of radial motors is high. They cannot have a high rotational speed because of the unilateral connection between cam and rollers. The speed is usually less than 1,000 rpm.

![Fig. 32-8. Radial-piston motor.](image)

The fundamental gain in reducing the volume of oil under high pressure will be demonstrated later. It may be remarked at the moment that, by reducing the volume of oil under compression, the designer reduces the torque; to deliver a given power, the speed of rotation must be increased.

5. Axial-piston Motors (Barrel-cylinder Motors). These consist of a set of small cylinders A having their axes equally distributed on a cylinder B. In each small cylinder (there is usually an odd number) moves a piston C. The connecting rod D is attached to a connecting disk E by a ball-and-socket joint F. *The plane of the connecting disk is not perpendicular to the axes of the cylinders A* (in Fig. 32-9 only two pistons have been drawn). The block of steel, in which the cylinders A are drilled, is clamped to the shaft G; while the connecting disk E is coupled to the same shaft by a Cardan joint (or any equivalent universal joint). Furthermore, the plane of the connecting disk is maintained in a constant direction: its normal is fixed, making an angle $\alpha$ with the direction of the driving shaft.

It is clear that, if high-pressure oil is admitted to the cylinders located in front of the projection plane, through the lenticular porthole I of the
valve plate \( J \) (the cylinders located behind the same projection plane being opened through the slot \( K \)), the system will rotate.

Theoretically speaking, it is necessary to evaluate the reaction moment of the connecting rods, not by reference to the main axis, but by reference to the axis bearing the rotation vector of the connecting disk. Forces transmitted through the rods are necessarily directed along them, since they are connected to the plate\(^1\) by means of spherical joints. Their moment with respect to the axis of rotation is zero, yet these forces are the driving forces. The plate should, in fact, be considered as an isolated system; therefore, this holds for the reactions of the stationary lugs maintaining the plate in a fixed direction plane. The moment of these reactions with respect to the axis of rotation is not zero.

Another version, the so-called fixed-angle motor, is sketched in Fig. 32-10. In both types of motors the distributing disk—fixed with respect to the frame—has one of its slots connected to the high-pressure oil supply, the other being connected to the low-pressure oil tank.

Finally, it will be noted that, by adjusting \( \alpha \), the angle of tilt of the connecting disk, the volume displaced per revolution can be adjusted. The use of this feature will be shown later. Of course, in such cases of variable-flow motors, the motors are usually of the type shown in Fig. 32-9, rather than of the type of Fig. 32-10.

### 32.1.3. Pumps

Each of the motors discussed above is reversible if its controlling member can be reversed. The required performance can

\(^1\) In fact, the rods are not exactly parallel to the axis of rotation during a revolution, but their angle remains very small.
be obtained either from a constant-pressure source or from mechanical energy connected to a pump having controlled characteristics. An axial-cylinder motor (See Sec. 32.1.2) used as a pump is the solution usually adopted (Fig. 32-11). So far as pressure sources of constant output are concerned, it is assumed that their internal impedance (defined as the ratio of pressure to mass flow) is zero. The pressure at the output of the source, in that case, does not depend on the flow. The complete outfit usually consists of a constant-flow pump, a calibrated relief valve, and a hydraulic accumulator (Fig. 32-12). The purpose of the accumulator is to smooth out the pressure fluctuations due to the pump. Some particular devices, being servo systems themselves, must be introduced into any important equipment. Figure 32-13 shows a device made of a calibrated spring and a damper (dashpot). The piston is subjected to the pressure available at the output of the pump; it actuates the flow-control element of the pump.

The combination of a variable-flow pump and a motor is called a hydraulic transmission. The most elaborate transmissions normally utilize barrel-cylinder pumps and motors (Fig. 32-14).

32.1.4. Controlling Devices. The distributing valve previously mentioned in Sec. 32.1.2, par. 1, has many different configurations, as summarized in Fig. 32-15:

1. The linear valve with four different ways [for example, for an actuator with double-acting main cylinder (Fig. 32-15a)]
2. The rotating valve with, again, four ways (Fig. 32-15b)
3. The three-way linear valve [for example, specified to control an actuator with a differential-area piston (Fig. 32-15c)]
Fig. 32-13.

Fig. 32-14. Hydraulic transmission.

Fig. 32-15. (a) Four-way valve.  (b) Four-way rotating valve.  (c) Three-way valve.
It should be noted that, in neutral position, the valve locks the piston (except for the leaks between piston and cylinder) if the piston in the valve exactly covers the ports connecting the motors to the oil lines.

Many combinations of valves can be found. In order to improve both their sensitivity and their linearity, the ports \( a_1 \) and \( b_1 \) are very often made wider than the thicknesses of the corresponding pistons \( A \) and \( B \) (Fig. 32-16). Further consideration will be given later to that solution. An obvious defect of such a disposition is that there is a permanent leakage through the valve. Also, there is sometimes provision for superimposing on the motion of pistons \( A \) and \( B \) (input function) a high-frequency (100 cps for instance) oscillation (called dither). The amplitude of the oscillation is very small with respect to the maximum displacement of the valve. This solution is similar to the one above: it improves sensitivity and linearity at the cost of a small permanent oil leakage.

*Leaking valves* (or needle valves) are very often utilized to control the single-acting actuators having differential-area pistons (Fig. 32-17). The control of the leakage by the position of the needle exactly determines the pressure in the chamber if the flow required by the motor is zero or is negligible. It is difficult to write the equations describing the operation of such a valve; all assumptions must be carefully detailed. The reader can best refer to specialized books. The chief advantage of the device under discussion is its easy, cheap realization. On the other hand, it not only has a permanent leak, but it is not linear (Sec. 32.4.2). Again, the reaction force on the needle varies appreciably; it depends on the needle displacement, the upstream pressure (and its fluctuations), and the downstream disturbances (those reaching the motor). The needle valve is commonly used as a "preamplifier relay" supplying the energy necessary to actuate the piston of a distributing valve, such as that shown in Fig. 32-1.
32.2. SUMMARY OF THE FUNDAMENTAL LAWS OF HYDRODYNAMICS

32.2.1. Perfect and Real Fluids. A fluid is said to be perfect when no internal friction exists, i.e., when there is no viscosity. Such a condition does not preclude compressibility. The study of these fluids is within the scope of rational mechanics, which makes it possible to predict, with reasonable accuracy, the performance of an actual hydraulic system.

However, the introduction of the additional parameter, viscosity, is necessary to describe the condition of the real fluid. The pressure along the pipe in which the flow takes place is not constant, as in the case of a perfect fluid; this fact is represented by the concept of pressure-head loss. By calculation of pressure losses in the various areas of a hydraulic circuit, the use of a viscosity coefficient can be avoided, except in determining the exact nature of the flow. An accurate calculation of the pressure losses is theoretically possible, but complex; on the other hand, it becomes simpler once a relative error of about \( \pm 30 \) per cent is admitted. This degree of approximation is obtained in practice only in circuits incorporating many elements whose pressure-head-loss coefficients are estimated once and for all, or are made up of abnormally rough pipes. Plus or minus 30 per cent accuracy is, in practice, always sufficient for a preliminary study. It is often adequate for an installation project, where pressure losses represent but a small fraction of the total dynamic-head loss.

The purpose of this section is to outline a simple method for determining pressure-head losses in a fluid flowing through a pipe of arbitrary shape. The method holds for an incompressible fluid, it being assumed that compressibility can, in the same circuit, be considered as negligible from the standpoint of the pressure-loss calculation, while not negligible from the standpoint of the dynamic characteristics of the element in the circuit.

32.2.2. Bernoulli’s Law. The flow of a perfect fluid in a pipe—that is, incompressible flow without internal friction (no viscosity)—is described by Bernoulli’s law, which expresses the constant total pressure throughout the flow path:

\[
h + p + \frac{1}{2} \rho V^2 = C^* \]

where \( h \) is the mean height of the tube of fluid (assumed to have a small cross section) above a horizontal reference plane, \( p \) and \( V \) are respectively the static pressure and the fluid velocity in the cross section considered (supposed to be of constant sectional area) and \( \rho \) is the specific mass of the fluid (the specific weight, \( \gamma = \rho g \), is often used).

Note 1. Many writers, particularly German authors, consider a pressure head. This is more representative, since it occurs in meters:

\[
h + \frac{p}{\rho g} + \frac{V^2}{2g} = C^* \]

Note 2. In high-pressure circuits, the \( h \) variations introduce only a corrective term \( \Delta h \rho g \), which is small with respect to \( p + \rho V^2/2 \). They can, therefore, be neglected.
32.2.3. The Pressure-head Loss. In practical cases, the fluids involved are not perfect, and between two cross sections the total pressure undergoes a reduction, called the pressure-head loss, namely,

$$\Delta P = -h_2 + (p_1 - p_2) + \frac{\rho}{2} (V_1^2 - V_2^2)$$

32.2.4. Distributed and Local Pressure-head Losses. An installation is generally made up of sections of tubing connecting various elements (distributors, metering orifices, nozzles, or simply elbows and fittings). In the following developments only passive elements are considered to the exclusion of pumps and motors, which add or subtract energy to and from the fluid.

Pressure-head losses are generally divided into two categories: distributed and local losses. The first appears in tubing, piping, conduits, i.e., in elements where the fluid velocity is constant along a path much greater than the cross section of the tube of fluid. The latter appears where local variations, i.e., sudden variations in area or in direction, arise.

32.2.5. Expression of Local Pressure-head Losses. It is well known that, while the transformation of pressure into velocity is fairly easy, the transformation of velocity into pressure is somewhat more difficult. From this stems the idea of representing the pressure-head loss across a restriction (Fig. 32-18) by an expression of the form of

$$\Delta P = \zeta \frac{\rho}{2} (V_2^2 - V_1^2)$$

where $\zeta$ represents the fraction of static pressure-head loss in the convergent tube which is not recovered in the divergent part of the circuit.
Note 1. The value of such an expression is confirmed by experience which shows that, within wide limits, \( \iota \) depends upon the shape of the restriction and not upon the velocity and the geometrical scale.

Note 2. By definition, \( V_1 \) and \( V_2 \) in the preceding formula are the velocities which would be reached by the fluid supposed to flow at uniform velocity \( V = Q/s \) across the entire cross section. This condition is necessary to a convenient use of the formula.

Note 3. The preceding analysis shows that, contrary to widespread opinion, there is no reason for \( \iota \) to be smaller or greater than unity. If the cross section of the restriction varies quite slowly (Fig. 32-18), an important fraction of the pressure is recovered and \( \iota \ll 1 \). If, on the other hand, the restriction is very sudden, the flow, by effect of the contraction, takes on an actual minimum cross section \( s_i' \ll s_i \); whence the actual velocity is \( V_i' > V_i \). In addition to the fact that the pressure recovery is practically negligible, a value of the coefficient \( \iota > 1 \) should be applied in the pressure-head-loss equation (where \( V_2 \), and not \( V_2' \), is used).

Note 4. In many practical cases \( s_i' \ll s_i \), and the equation becomes

\[
\Delta P = \iota \frac{\rho}{2} V_i^2 = \iota \frac{\rho}{2} \frac{Q_i}{s_i^3}
\]

where \( s_i \) is the minimum cross section.

The formula for the pressure-head loss in a restriction can be generalized for all local irregularities. For convenience, however, the following two cases should be considered.

1. Restricting Elements: Nozzles, Distributors. The purpose of these elements is to create a pressure drop or to limit a flow. Their pressure-drop coefficient will be derived by the formula

\[
\Delta P = \iota \frac{\rho}{2} V_i^2 = \iota \frac{\rho}{2} \frac{Q_i}{s_i^3}
\]

The minimum cross section offered the fluid is always small with respect to that of the tubing. Additively, its geometrical shape is well known; therefore, a good approximation is obtained.

a. For a nozzle consisting of a hole drilled in a wall, the diameter of the hole being about equal to the wall thickness and the hole presenting no important burrs: \( \iota = 1.75 \) to 2.10. A conical restriction upstream lowers the \( \iota \) value rapidly to 1. A conical variation downstream (diffuser) is quite less efficient, and only if the aperture angle is small (less than 10°). In any case, if a certain pressure-head drop is desired, the best accuracy will be obtained by means of a simple hole with a very clean edge and a value of \( \iota = 1.90 \).

b. For a distributor, the \( \iota \) value is generally between 1 and 2. It is close to 1 for a progressively shaped distributor (cone or slope) and close to 1.9 for a sharp-edged distributor. The \( \iota \) value can be estimated fairly well by considering the distributor in the position concerned (by analogy with the case of drilled holes with or without transition cones). In some exceptional cases, the flow is not governed by the simple law adhered to previously. Such an exception will be discussed later in connection with tubes.

2. Other Elements. “Other elements” include those whose purpose is not to create a pressure-head loss (connection fittings, wall crossings, taps or valves in the open position, one-way valves, etc.). It is convenient to relate their pressure-head loss to the internal diameter of the tubing on which they are installed. In a few cases, a value of \( \iota < 1 \) may be found (straight fittings). The \( \iota \) value obviously depends upon the care exercised by the manufacturer. However, the following \( \iota \) values can be admitted: 1 for an elbow, 2 to 3 for a banjo fitting, and 3 to 5 for current one-way-
type values. It is very often possible, through minute modifications, to divide the $\zeta$ value of these elements (banjo fittings, for instance) by 2.

32.2.6. Equation of Distributed Pressure-head Losses. It is interesting to extend the formula arrived at for local pressure-head losses to the case of distributed losses. However, in this case it is necessary to consider also the length of the tube, hence the proposed formula

$$\Delta P = \lambda \frac{L \rho}{D} V^2$$

where $\lambda$ = distributed pressure-head-loss coefficient  
$L = \text{tube length}$  
$D = \text{tube diameter}$  
$V = \text{velocity along the tube}$

Unfortunately, the $\lambda$ value is not as constant as the $\zeta$ value. Two types of flow are to be considered: turbulent and laminar.

For most practical purposes it can be considered that a flow is turbulent if the Reynolds number $R$ is greater than 2,500 and laminar if less than 1,500. Let it be recalled that $R = VD/\nu$, where $V = \text{flow velocity}$, $D = \text{tube diameter}$, $\nu = \text{kinematic viscosity}$ ($\nu = \mu/\rho$ where $\mu$ is the viscosity coefficient and $\rho$ the fluid specific mass). Most of the flows encountered in aircraft are turbulent (except those of lubricating oils). However, the $\lambda$ values which can be reached in a laminar flow can be very high; it is mandatory to calculate the Reynolds number whenever one deals with the low-velocity flow of a viscous fluid in a small tube (or simply a very cold fluid, momentarily viscous).

Turbulent Flow. Except in the vicinity of the transition point, the $\lambda$ coefficient is approximately constant and depends only upon the internal roughness of the tube. To a first approximation, for carefully manufactured tubes such as those used in the aeronautical industry, a value of

$$\lambda = 0.025$$

can be used. Roughly speaking, it can be said that a tube length equal to 40 diameters leads to a pressure drop of "one unit of velocity gain" ($\rho V^2/2$).

Note. It is safe to add one-half of "unit velocity gain" at the upstream extremity of the tube, if the velocity gains its increase through a sudden cross-sectional change (Fig. 32-19).
Laminar Flow. In this type of flow, it is well known that the pressure-head loss is proportional to the velocity. However, it is still convenient for purposes of homogeneity to express it by the same formula for laminar flow which occurs only rarely and locally. The $\lambda$ coefficient will then be proportional to $1/V$. In this case, the influence of the viscosity is practically negligible, and Poiseuille's law gives a very good approximation for $\lambda$, namely, $\lambda = 64/R$.

Note 1. Since the viscosity increases very rapidly when the temperature decreases, low temperatures can lead very roughly (speed-setting variations) and very rapidly ($\lambda = Ks$) to catastrophic pressure-head variations.

Note 2. Local pressure-head losses can result from a laminar flow, for instance, in the case of flow through parts subjected to a certain play. In first approximation, the Reynolds number is derived by replacing the diameter by the play.

32.2.7. Calculation of an Installation. Pressure-head losses are obviously additive. Two cases are to be distinguished when the flow is to be computed between two sources at a known pressure:

1. The flow is, in all probability, entirely turbulent:

$$\Delta P = \sum \zeta_i \frac{\rho}{2} V_i^2 + \sum \lambda_j \frac{L_j \rho}{D_j} \frac{V_j^2}{2}$$

$$\Delta P = \sum \zeta_i \frac{\rho}{2} \frac{Q_i^2}{s_i^2} + \sum \lambda_j \frac{L_j \rho}{D_j} \frac{Q_j^2}{2 s_j^2}$$

$$\Delta P = \left[ \sum \zeta_i \frac{\rho}{2} \frac{1}{s_i^2} + \sum \lambda_j \frac{L_j \rho}{D_j} \frac{1}{2 s_j^2} \right] Q^2 = P_1 - P_2$$

an expression in which the total of the terms between brackets is constant. The assumption of a true turbulence is of course to be verified.

2. The flow is laminar in certain areas. It is then more rapid to draw the curve $\Delta P = f(Q)$ with a certain number of points and to intersect by $\Delta P = P_1 - P_2$.

The problem of circuits in parallel is only slightly more complicated than for electrical circuits; it is conveniently solved by use of the equivalent theoretical cross section:

$$Q = \left( \frac{2}{\rho} \Delta p \right)^{\frac{1}{4}} 1 = \left( \frac{2\Delta p}{\rho} \right)^{\frac{1}{4}} s_{th}$$

or

$$s_{th} = \frac{1}{\zeta^{\frac{1}{4}}} s = \left( \frac{D}{L \lambda} \right)^{\frac{1}{4}}$$

The order of magnitude of practical values will be given below:

a. Computation of $\Delta P = \rho V^2/2$ (unit velocity gain). It is advisable, for practical purposes, to replace $\rho$ by $\gamma/g$, where $\gamma$ represents the specific density (this is the value usually given). For instance, if $\gamma$ is expressed in kilograms per cubic meter, $\Delta P$ is obtained in kilograms per square meter by the relation

$$\Delta P = \frac{\gamma}{20} V^2$$

$V$ in meters per second

1 Recall that 1 kg/cm$^2$ is equivalent to 14.504 psi and 1 kg/m$^2$ is $10^{-4}$ kg/cm$^2$. 
or $\Delta P$ in kilograms per square centimeter by the relation

$$\Delta P \sim \frac{\gamma \times 10^{-5}}{2} V^2 \quad V \text{ in meters per second}$$

For those hydraulic fluids commonly used in servomechanisms, $\gamma$ is about 860 kg/m$^3$; the chart shown in Fig. 32-20 gives $\Delta P$ as a function of $V$ (or reciprocally) for the velocity increase (in a convergent):

$$\Delta P = \frac{\gamma}{2g} V^2$$

b. Pressure loss in a tube. Referring to the above formula for tubes,

$$\Delta P = \lambda \frac{L \rho}{D} V^2 \quad \text{with} \quad \lambda = \frac{1}{40}$$

the pressure-head loss in kerosene flowing at 10 m/sec will be approximately 1 kg/cm$^2$ for $L = 100 \ D$; that is, for a length of 1 m and an inside diameter of 10 mm. For a 2-m/sec velocity, the pressure-head loss in the same tube would be only 0.040 kg/cm$^2$.

This explains why axial velocities are practically limited to 8 to 10 m/sec in high-pressure tubings, where a pressure drop of a few kilograms per square centimeter is permissible because it represents only a low percentage of the 50, 100, or 150 kg/cm$^2$ from the fuel pump of a jet engine, or of the 200 or 300 kg/cm$^2$ from a hydraulic pump. Conversely, in low-
pressure tubings, the limit velocity is of the order of 2 m/sec, since higher velocities are incompatible with the available pressure, especially at high altitudes.

Note that, with a given flow and pipe length, the pressure-head loss is inversely proportional to the fifth power of the tube diameter:

\[ \Delta P = \lambda \frac{L \rho}{D^2} V^2 = \lambda \frac{L \rho}{D^2} \left( \frac{Q^2}{\pi D^2/4} \right)^2 = \lambda k \frac{Q^2}{D^6} \]

c. Some values of the kinematic viscosity in square meters per second at 20°C are:

<table>
<thead>
<tr>
<th>Liquid</th>
<th>Viscosity (m²/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>1.08 × 10⁻⁵</td>
</tr>
<tr>
<td>Water vapor</td>
<td>560 × 10⁻⁶</td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.74 × 10⁻⁶</td>
</tr>
<tr>
<td>Ethyl ether</td>
<td>0.3 × 10⁻⁶</td>
</tr>
<tr>
<td>Fuel oil</td>
<td>~0.6 × 10⁻⁶</td>
</tr>
<tr>
<td>Kerosene</td>
<td>~1.7 × 10⁻⁶</td>
</tr>
<tr>
<td>Castor oil</td>
<td>~1,000 × 10⁻⁶</td>
</tr>
<tr>
<td>Glycerin</td>
<td>~650 × 10⁻⁶</td>
</tr>
<tr>
<td>Hydraulic fluids</td>
<td>~25 × 10⁻⁶</td>
</tr>
</tbody>
</table>

The kinematic viscosity is often expressed in centistokes: 1 cs = 10⁻⁴ m²/sec.

32.3. TRANSFER FUNCTION OF A HYDRAULIC TRANSMISSION

32.3.1. Writing the Equations. Reference is made to Fig. 32-9.

**Notation.**

- \( n \) Angular velocity of pump, rps
- \( d_p \) Maximum volume displaced by pump in one revolution
- \( x_p \) Percentage of above as set by position of control lever
- \( d_m \) Volume displaced by motor per radian
- \( J_m \) Total inertia referred to output shaft
- \( L \) Leakage coefficient
- \( B \) Compressibility modulus (usually termed *bulk modulus*)
- \( V \) Total volume of high-pressure oil contained in transmission circuit
- \( P \) Oil pressure at inlet port of motor
- \( \theta \) Angular displacement of output shaft (motor shaft)

The equation is established by expressing the conservation of the volumetric flow:

Pump flow = motor (received) flow + oil leakage flow + flow due to compressibility

that is to say:

\[ Q_p = Q_m + Q_L + Q_c \quad Q_p = d_p n x_p \quad Q_m = d_m \frac{d\theta}{dt} \quad Q_L = LP \]

\[ Q_c = \frac{dV}{dt} = \frac{V}{B} \frac{dP}{dt} \]

\( B \), the compressibility modulus, has the dimension of a pressure (for commonly used oils, \( B \) is about 2 × 10¹⁰ baryes).

¹ To be exactly correct, it is necessary to add a term due to the expansion of the oil circuit (pipes, etc.).
For a perfect hydraulic motor,
\[ C_m = d_m P \]
where \( C_m \) is the driving torque. If the only load on the shaft is an inertia,
\[ C_m = J_m \frac{d^2 \theta}{dt^2} \]
and the flow equation can be written as
\[ d_p n x_p = d_m \frac{d \theta}{dt} + LP + \frac{V}{B} \frac{dP}{dt} \]
Eliminating \( P \) and \( C_m \) between the three last equations yields
\[ d_p n x_p = d_m \frac{d \theta}{dt} + \frac{LJ_m}{d_m} \frac{d^2 \theta}{dt^2} + \frac{VJ_m}{Bd_m} \frac{d^2 \theta}{dt^2} \]
Taking the Laplace transform of both members for zero initial conditions yields
\[ \frac{\Theta(s)}{X_p(s)} = \frac{d_p n}{s \left( \frac{VJ_m}{Bd_m} s^2 + \frac{LJ_m}{d_m} s + d_m \right)} \quad (32-1) \]
and, the nondimensionalized equation being
\[ \frac{\Theta(s)}{X_p(s)} = \frac{K}{s \left( \frac{s^2}{\omega_n^2} + \frac{2z}{\omega_n} + 1 \right)} \quad (32-2) \]
it is found that
\[ \omega_n^2 = \frac{Bd_m^2}{VJ_m} \quad (32-3) \]
\[ \frac{2z}{\omega_n} = \frac{LJ_m}{d_m^2} \quad (32-4) \]
\[ K = K_v = n \frac{d_p}{d_m} \quad \vdots \quad (32-5) \]

32.3.2. Discussion. Equation (32-5) shows that the velocity time constant \( K \) [it is \( \lim s \Theta(s)/X(s) \) as \( s \) approaches zero] is finite: there is, therefore, a "velocity error" in the case of a constant-velocity input signal. In the case of a constant-position input signal, there is no position error because \( \Theta(s)/X(s) \) goes to infinity when \( s \to \infty \) (Secs. 15.1 and 15.2). Relative to the transfer function of Eq. (32-1) or (32-3), it is to be noted that the system has a natural frequency. Under the proper conditions, therefore, a resonance oscillation can occur. The effect of the values of the parameters \( \omega_n \) and \( z \) on the performance of the system has been previously investigated (Chap. 6) by consideration of the transfer function
\((s\Theta/X)\), that is, the function relating the output-shaft velocity to the entry position signal. The quantity normally taken into consideration is the velocity, and that is why the output-shaft-position concept is of small interest. As the above discussion is based on the assumption of a linear system, the discussions of position or of velocity characteristics are strictly equivalent. It is only because the results can be more readily utilized that the velocity is considered.

It is to be noted in Eq. (32-3), defining the natural frequency, that the inertia of the motor is in the denominator. A parallel result would be established in writing the equations of a linear\(^1\) actuator. In that case, the inertia of the piston is similarly in the denominator. This demonstrates our previous statement: an actuator with a piston having two different effective sections has a lower natural frequency than the conventional actuator, since the inertia of the piston is higher.

![Diagram](image)

Fig. 32-21.

Again Eqs. (32-1) and (32-3) are evidence that the natural frequency is inversely proportional to the square root of the volume \(V\) of high-pressure oil. It is profitable to keep this volume to a minimum. It is the total of three terms:

\[
V = \frac{2\pi d_m}{4} + \frac{d_p}{4} + \frac{v}{2}
\]

where \(v\) is the volume of the piping and of all the dead volumes between the cylinders of the pump and of the motor. On the other hand, \(d_p\) can be reduced by driving the pump at a faster rate than the motor. That motor must be compared to a linear actuator (Fig. 32-21: piston area, \(S\); piston mass, \(M\)). With the notation of Fig. 32-18, the transfer function may be written as

\[
\frac{X_m}{X_p} = \frac{\pi d_p/S}{s [(VM/BS^2) s^2 + (LM/S^2)s + 1]}
\]  
(32-6)

The smaller \(V\) is, the higher the natural frequency.

A fundamental advantage of the rotating motor is that it allows a compromise between (1) the use of high-pressure oil and (2) the reduction of the ratio \(V/d_m\). This is made feasible by the use of a higher reduction ratio.

\(^1\) Here again, linear means rectilinear.
Finally, consider the damping coefficient \( z \) [Eq. (32-4)]. It is proportional to the oil-leakage coefficient \( L_1 \), which means that a very accurately machined hydraulic transmission would have a very low damping coefficient. This is not usually of interest and does not justify such a costly means of realization.\(^1\)

One may refer to the amplitude-vs.-frequency and phase-vs.-frequency diagrams for a second-order system. The transfer functions are represented by Eqs. (32-2) or (32-6), in which the ratios \( s\Theta/X_p \) or \( sX_m/X_p \) are considered; they are plotted in Sec. 6.2.

![Diagram](image)

Fig. 32-22.

If the input signal \( X_p \) frequency spectrum is represented by \( \Phi_{xx}(\omega) \) and if \( H(s) \) is the transfer function of the hydraulic transmission, it has been established (Sec. 12.3.1) that the output spectrum is

\[
\Phi_{xy}(\omega) = |H(\omega)|^2 \Phi_{xx}(\omega)
\]

that is, if \( A(\omega) \) is the amplitude ratio of \( H(s) \),

\[
\Phi_{xy}(\omega) = |A(\omega)|^2 \Phi_{xx}(\omega)
\]

In Fig. 32-22, the curve \( A(u) \), as a function of the dimensionless frequency \( u \), has been plotted for a damping ratio \( z = 0.4 \). To an input spectrum \( \Phi_{xx}(u) \) (solid line) corresponds an output spectrum \( \Phi_{xy}(u) \) (dotted line). Figure 32-23 deals with the case \( z = 0.1 \). It is seen that to the same input spectrum \( \Phi_{xx}(u) \) corresponds an output spectrum having an energy near the dimensionless resonance frequency, which is enormously increased with respect to the input energy level. Without further discussion, it is easily seen how harmful an incorrect “filtering” by the upstream stages would be if it admitted a nonnegligible residue of harmonics located in the immediate vicinity of the resonance frequency.

\(^1\) However, at the present time, gaskets do not leak at all. A calibrated hole is therefore drilled through the piston in order to adjust the leakage coefficient to a proper value.
It can be concluded, therefore, that, even if the component of the spectrum $\Phi_\omega(u)$ located in an interval $\pm du_\omega$ around the resonance pulsation $u_\omega$ is very small, the output signal nevertheless can have about the corresponding frequency an important energy level which can completely disturb the behavior of the servo.

![Graph of $\Phi_\omega(u)$](image.png)

Fig. 32-23.

Usually, no corrective effect is expected from the driving element; it is required that it reproduce as exactly as possible the spectrum density of the input signal.

32.4. EQUATIONS OF VALVE-CONTROLLED HYDRAULIC MOTORS

32.4.1. Establishing the Equations. Reference is made to the first control device described in Sec. 32.1.4. The equations are the same, and the results are similar—qualitatively—for a linear actuator and a rotating motor. This last system will be considered here.

The pressure drop $\Delta P_1$ in the distributing valve is related to the mass flow $Q$ passing through it by $Q = k(\Delta P_1)^H$, where $k$ is a characteristic parameter of the valve that depends on the shapes of the ports and the motion of the valve [linear displacement of a spool valve (Fig. 32-15a), angular displacement of a rotating valve (Fig. 32-15b)]. It is, then, a function of time. Sometimes, the flow is assumed to be proportional to the displacement which results when all the other parameters remain
fixed. This assumption, however, is not quite valid near the zero or neutral position. If \( \Delta P_2 \) is the pressure drop between the input and the output of the motor, and if the input of the system is characterized by the function \( k(t) \), or more simply by \( k \), the following equation can be written:

\[
 k(\Delta P_1)^{\frac{1}{2}} = d_m \frac{d\theta}{dt} + L \Delta P_2 + \frac{V}{B} p \Delta P_1 \quad \text{with} \quad P = \Delta P_1 + \Delta P_2
\]

where \( P \) is the constant pressure of the valve oil supply and

\[ C = d_m \Delta P_2 \]

whence, by eliminating \( \Delta P_1 \) and \( \Delta P_2 \),

\[
k \left( P - J \frac{d^2\theta}{dt^2} \right)^{\frac{1}{2}} = d_m \frac{d\theta}{dt} + \frac{L J}{d_m} \frac{d^2\theta}{dt^2} + \frac{V}{B} J \frac{d^2\theta}{dt^2}
\]

(32-7)

Equation (32-7) is not linear and cannot be integrated exactly. It is possible, as a first approximation, to consider the flows due to leaks and compressibility as negligible. Equation (32-7) is, then, reduced to

\[
k \left( P - J \frac{d^2\theta}{dt^2} \right)^{\frac{1}{2}} = d_m \frac{d\theta}{dt}
\]

or, defining \( u \) by \( \frac{d\theta}{dt} = u \),

\[
P - J \frac{du}{dt} = \frac{d_m^2}{k^2} u^2
\]

The response as a function of time can, then, be evaluated for a given input signal \( k(t) \). The equation is a Ricatti equation which can be integrated if a particular solution is known. The use of analog-computer equipment allows easy solution of the equation as soon as the coefficients have been numerically evaluated. Only the final results are summarized below;\(^1\) if the maximum torque (acceleration of the load) required from the motor is less than the maximum theoretically available torque \( (P d_m) \), then a sinusoidal velocity of the motor shaft corresponds, essentially, to a sinusoidal motion of the valve.

32.4.2. Linearization. The above-mentioned nonlinearities are the consequence of the relationship \( Q = k(\Delta P)^{\frac{1}{2}} \), as plotted in Fig. 32-24. The smaller the curvature of the parabola, the better the linearization. Therefore, near the neutral point (here assumed to be set at the zero-flow position), no satisfactory linearization can be effected.

On the other hand, linearization near a point corresponding to a nonzero average flow \( Q_0 \) is feasible; the smaller the ratio \( \Delta Q/Q_0 \) (\( \Delta Q \) being the flow increment due to small motions superimposed on the average constant motion), the better the linearization. Linearization of a rotating motor has an obvious significance; linearization of a linear\(^2\) actuator is somewhat questionable, since a uniform motion of the piston must be assumed.


\(^2\) Meaning \textit{rectilinear}. 
Furthermore, from previous considerations, it is possible to deduce a method of linearization near the neutral point. If the ports are slightly broader than the thicknesses of the pistons in the valve (Fig. 32-16), then even at standstill there remains a finite flow. The linearization is yet effective under the condition that the motion near the zero point is of such an amplitude that never, at any time, is the flow through the ports zero.

32.4.3. Energy Balance. In a valve-controlled system, the energy carried within the high-pressure fluid is dissipated in the valve. As a first approximation, the power consumed is a function only of the velocity of the motor and does not depend on the working conditions at the output of the motor. Under such conditions, not only is the efficiency very poor, but the heat dissipated within the valve is appreciable. Furthermore, the operating conditions, as a consequence, more or less aerate the oil in the distributing valve.

This, again, tends to favor the hydraulic transmission in which no systematic energy dissipation occurs (Sec. 32.3). The average power efficiency of such a transmission is excellent (since the efficiencies in various operating conditions are excellent).

32.4.4. Case of the Leaking-Valve-controlled System. The general properties of such a device are similar to those of the valve-controlled motor, except that a leaking-valve-controlled servo operates around a nonzero average permanent flow. Consequently, (1) the leaking-valve-controlled device can be linear over a domain broader than that of the distributing-valve-controlled motor. On the other hand, (2) the permanent flow depends on the power of the associated motor and on the required performance. In particular, if a small time constant is specified, it is necessary to have provision for an important permanent leakage; actually, the permanent leakage must be larger than the flow corresponding to the maximum velocity or maximum acceleration.

32.4.5. Constant-flow Valve. Instead of a constant-pressure source, a constant-flow source is often used; then it is obvious that the controlling device (valve or needle) must present a permanent leak. For example, consider a constant-volume pump supplying flow to an underlapped four-way spool-type valve (Fig. 32-25). Under equilibrium conditions, the sleeve is at center and the pressure drop is the same on both sides of the cylinder $A$; as a result, the piston of the motor (main cylinder) does not move. If the sleeve is moved to the right or left, the constant flow pumped is divided into two different flows, creating different pressure drops, and thus the main piston is moved to the right or left, respectively.

Suppose now that the sleeve of the four-way valve is an integral part of the cylinder, resulting in a one-to-one follow-up system (unity feedback). The piston rod is fixed to the frame of the unit. The cylinder, through a suitable linkage, controls the position of the load. If a disturbance is applied to the load, the pressure builds up to the output demand and the partition of the constant flow is modified according to the disturbance.
The equations of such an open-center valve associated with a motor (cylinder and piston, for example) are based upon the pressure drop in an orifice: the flow is proportional to the area of the orifice opening and to the square root of the pressure drop across the orifice. The total flow pumped is constant. The final equations relating the output (position of the piston) to the input (position of the sleeve) are generally not linear and depend on the shape of the ports.\(^1\)

32.4.6. Remarks on Transfer-function Calculations. It has been previously mentioned that, in the case of electric motors, the transfer function can usually be evaluated without taking into account the load applied to the motor. It has also been said that such a simplification is very often not justified in the case of hydraulic motors. This point must be somewhat detailed.

If the motor considered is of the rotating type, it differs from an electric motor in that its inertia for a given available power is much lower. If the motor, on the contrary, is an actuator, the situation is quite different. As compared to the inertia of the piston of the master cylinder, the inertia of the load, transferred to the actuator rod, must be taken into consideration; it is not usually negligible.

If there are counter-torques, the condition specifying the maximum velocity in sinusoidal oscillations or the maximum velocity in steady-state conditions very often fully determines the actuator. An example is as follows:

A servocontrol actuator for an aircraft has to be designed in compliance with the following specifications:

Inertia of the control surface: 15,000 gm-cm$^2$ (cgs unit)
Maximum amplitude of the motion: 10°
Resistive torque: 8 csn-m/rad = $8 \times 10^4$ dyne-cm/rad (cgs unit) (csn stands for centisathene, i.e., approximately 1 kg force; the sathene is the mts unit of force)
Maximum frequency: 5 cps
Maximum velocity: 100°/sec
Oil-pressure supply: 30 hps (one hectopieze, or kilogram per square centimeter, is approximately 14 psi)
Compressibility modulus $B$: $2 \times 10^{10}$ cgs

(The actuating lever connecting the actuator rod to the control surface is assumed to be 3 cm.)

In such conditions, the transfer function may be written as

$$X = 1.75 \times 10^{-4}s^4 + 6.1 \times 10^{-4}s^3 + 1.71s + 0.31$$

where $x$ is the transform of the distributing-valve displacement and $y$ that of the piston displacement (Fig. 32-26); the flow through the valve is assumed to be proportional to $x$. If $q$ is the unit flow, the transfer function is

$$\frac{Y}{X} = \frac{3.22q}{(1 + 5.52s)(1 + 3.55 \times 10^{-4}s + 10.25 \times 10^{-4}s^2)}$$

In this transfer function, the time constant $\tau_1 = 5.52$ sec is strictly related to the counter-torque. If there is no such torque, the transfer function enfold a pure integration,

$$\frac{Y}{X} = \frac{0.585q}{s(1 + 3.55 \times 10^{-4}s + 10.25 \times 10^{-4}s^2)}$$

The value of the quoted time constant is somewhat surprising, as compared to the other time constants. Its meaning, however, is as follows: If the servo system, in which the actuator is the power stage, were strictly a position servo, the transfer function with a nonzero counter-torque shows that the equilibrium position would be reached 9 sec after the application of a command signal to the valve. Also, that position would correspond to a displacement of the control surface of more than 10°, which, in the present problem, does not mean anything. For a displacement of $\pm 4°$,

$^1$ It is seen that, in sinusoidal oscillation, the maximum velocity corresponds to an amplitude of less than 10°. Nevertheless, it is required that, at lower frequency, the amplitude be able to be 10°.
however, at a frequency of, for example, 5 cps, the transfer function can be considered as being
\[
\frac{Y}{X} = \frac{0.585\gamma}{s(1 + 3.55 \times 10^{-4} + 10.25 \times 10^{-6}s^3)}
\]
which means that the considered system is a velocity-control servomechanism (independent of the resistive torque).

Incidentally, characteristic parameters of a hydraulic actuator may be noted as:
\[
\omega_n = 3,120 \text{ rad/sec, or } 495 \text{ cps} \quad z = 0.0056
\]

The natural frequency is extremely high: this is due to the fact, that, in the above calculation, the compressibility modulus \(B\) utilized corresponds to pure oil; that is, dissolved air, distortion of the cylinder and of the pipes, etc., are not considered. The value \(z\) of the damping coefficient is, on the other hand, extremely low, because of the assumed small leakage coefficient \(L\): 2 per cent of the maximum flow of oil under maximum pressure. Such a coefficient corresponds to excellent machining of all the parts (Sec. 32.3.2).

### 32.5. SERVOCOMMUTORS OR BOOSTERS

#### 32.5.1. The Problem

The servocontrol, or booster, which is sometimes inadequately called a servomotor,

1 is a system which allows control of the motion of a load with an effort substantially lower than that required by direct actuation, the direct operation constantly remaining possible in case either of failure of the supply circuit or of desired disconnection of the servocontrol (declutching). These are essentially position servo systems, since obtaining a zero position error is attempted, at least for a constant position input. Other boosters are designed to give a zero error for a constant velocity input [zero velocity error (Sec. 14.3)].

The direct-operation requirement has a fundamental influence on the whole design of the device. Boosters can be considered as force (or torque) amplifiers. They must be clamped at one end to the frame and are very often coupled to the rods connecting the command member (stick, rudder pedals) to the control surface. Finally, if they are linear, their transfer function has the form,

\[
\frac{K(s + a_1)(s + a_2) \cdots}{s^n(s + b_1)(s + b_2) \cdots r},
\]

with \(n \geq 1\). It is recalled that, whenever possible, such a device has to be a minimum phase system \((a_1, a_2, \ldots \text{ positive})\). Furthermore, it has to be stable on open-loop operation,\(^2\) a condition which means that \(b_1, b_2, \ldots \) must be positive. In that case, the booster in the closed loop (i.e., with the first internal feedback path around the motor, Fig.

1 The term servomotor is used for the motor plus feedback loop (first internal feedback path around the motor) assembly. The back emf of a motor can be considered as a feedback path. This, however, is excluded here, since it only corresponds to the interpretation of the function of an element of the motor.

2 It must not be forgotten that, in a servo loop, some elements can be individually unstable (see Sec. 16.2.2).
32-27) will be stable if the transfer locus, plotted from \( \omega = -\infty \) to \( \omega = +\infty \), does not enclose the \(-1\) point.

As an example, consider first the control booster of an aircraft. The manual control of a heavy airplane (more than 50 tons) or a fast airplane (able to fly at supersonic speed) is impossible, or at least laborious. In the case of supersonic aircraft, the aerodynamic torques applied to the control surfaces can be extremely high.

*Power-assisted* steering of vehicles is another example: it is used on trucks, busses, and private cars.

![Diagram](image)

**Fig. 32-27.**

![Diagram](image)

**Fig. 32-28.**

32.5.2. Realization. The hydraulic boosters (or those actuated by compressed or depressed air) are the most widely used of servocontrols because of their simplicity, particularly in the case of rectilinear motion, and because of the small time constants easily obtained (the time constants, when air is the driving fluid, are obviously much larger than in the case of compressed oil).

Many designs have been made, but they are all based on the same principle: the distributing valve is located inside the piston of the master cylinder. The solution shown in Fig. 32-28 is that of a single-acting system, which is very simple and acceptable if very small time constants have not been specified. The high-pressure oil (or air) is supplied in \( E \) and applies a constant effort to the face \( H \) of the piston; it is also supplied, through the pipe \( D \), to the central chamber of the distributing valve.
made of a double piston $A$ connected to the control rod $G$. The face $K$ of the piston is connected to the cylinder of the distributing valve through the portholes $C$ drilled in the master piston itself. If an input signal displaces the control rod to the right, the chamber $L$ is connected to the outside through the pipes $C$ and $F$, the pressure drops, and the piston is driven to the right. The master piston, therefore, follows up the control rod to which no effort, or a small effort, must be applied. In case of failure of the high-pressure oil supply, the control rod carries the master piston mechanically; there remains a residual backlash corresponding to the relative maximum displacement of the valve with respect to the piston, usually only a few millimeters.

Note. If an autopilot is used to control an aircraft, it is interesting to use the servocontrols as the power stage of the autopilot. The block diagram of such a system is presented in Fig. 32-29. In that case, there must be provision for a device to clutch in and clutch out the autopilot. But in order to avoid any discontinuity when the human pilot takes over the control of the aircraft, it is advisable that the stick or the pedals follow up the motion of the corresponding control surfaces when
the autopilot is clutched. In that way, before disconnecting the autopilot, the human pilot follows up the motion of the stick and of the pedals, then declutches. Such a device is presented in Fig. 32-30, evidencing the principle of a solution utilized under many forms. Detailed technological description of boosters is outside the scope of this book.

32.6. ADDITIONAL REMARKS

32.6.1. Efforts on the Controlling-valve Rod. The connection of the controlling valve to preceding elements in the loop requires a knowledge of the forces applied to the valve (Fig. 32-15). These are the forces due to the inertia of the valve, friction, and the reaction of the flow of oil. Reaction forces can be calculated for a nonviscous fluid. Even in that case, these forces are preponderant (certainly the viscosity effect will further increase them).

It can be considered that, in optimum operating conditions, the efforts on the valve are about 400 g force/kw of power available at the motor shaft or actuator rod for a pressure supply of 200 hectopiezas. (Recall that 1 hectopieza, or 1 kg/cm², is approximately 14.5 psi.)

A systematic study of the shapes to be given to the pistons and the parts of a valve has been made. 1 It is shown that, for certain shapes such as those presented in Fig. 32-31, it is possible to have a negative reaction. In Fig. 32-32 are plotted the curves giving the force on the valve-vs.-piston displacement. The solid line presents the compensated piston, the dotted line is for the noncompensated piston. Figure 32-33, reprinted from the same source, shows the shape of a compensated distributing valve (the flow channels of the oil in the cavity are shown in Fig. 32-31).

1 S. Y. Lee and J. F. Blackburn, Dynamic Analysis and Control Laboratory, MIT, paper presented to the ASME session, November, 1951.
Fig. 32-33.

Fig. 32-34a.
32.6.2. Flow in a Hydraulic Pipe. It has been stated in Sec. 32.2 that the velocity of the fluid is proportional to $\Delta p$ for a laminar flow and to $(\Delta p)^{1/2}$ in the case of a turbulent flow. In servomechanisms, it is advisable to avoid any energy dissipation, not so much from an efficiency point of view, but to prevent emulsification of the oil and solution of air. The sections must be so calculated that the flow remains laminar even at the maximum specified rate. Figure 32-34 contains charts\(^1\) corresponding to

---

Figure 32-34b.

---

the standard hydraulic fluid ($\rho = 0.86$; $\eta = 0.32$ stoke) at three different temperatures: $-40^\circ C$ (hydraulic boosters of an aircraft cruising at high altitude); $0^\circ C$ (aircraft boosters in normal conditions); $50^\circ C$ (normal temperature of boosters in fixed installations). The flows are given in cubic decimeters and pressure drops in hectopascals per meter of pipes; the lines are marked in terms of internal diameters of the pipes as expressed in millimeters.

32.6.3. Comparison between Electric and Hydraulic Motors. This comparison\(^2\) is quite common in the case of industrial motors (as opposed to servomotors) and is usually reduced to a comparison of weights. As mentioned in Chap. 30, there is

---

1 These abacae are taken from J. Faisandier, "Les mécanismes hydrauliques," Dunod, Paris, 1957.

2 See, for example, ibid.
no general rule for selecting a motor, nor even for selecting the type—electric or hydraulic—of motor to specify.

Without considering the performance, it is of interest to enumerate here the main conclusions obtained for industrial motors:

a. A rotating (hydraulic) motor supplied by oil under a pressure of 200 hectopieses and delivering a power of about 10 kw has a specific weight of 0.53 kg/kw. The equivalent electric motor connected to 24-volt direct current weighs about 2.1 kg/kw.

b. For linear actuators, the advantages of the hydraulic solution by comparison with electric motors and associated screw-nut or gear-rack systems are still more obvious. For a delivered force of about 5 tons on a 0.5-m stroke and a linear velocity of 0.1 m/sec (this corresponding to 7.5 kw), the specific weight is about 1.1 kg/kw for a hydraulic jack and 25 kg/kw for an electric actuator.

c. Again, for equivalent nominal performance, it is found that the weights per unit of length of pipe or wire\(^1\) are of the same order of magnitude for high-pressure hydraulic motors (under a pressure higher than 150 hectopieses) and “high-voltage” electric motors (120 volts, for instance, but not 24 volts).

\(^1\) The electric motor is assumed to receive direct current through two conducting wires.
d. On the other hand, the line losses (pressure drops or line resistance) are higher in hydraulic pipes (the percentage in terms of pressure drops is about twice that in terms of voltage reduction).

e. Finally, it must be pointed out that it does not pay to accumulate hydraulic energy, as compared with electrical energy. In the latter case, it costs some 0.01 kg/kilojoule in electric batteries. The cost jumps to 0.6 kg/kilojoule in spherical hydraulic accumulators.

32.6.4. Stress Analysis of the Piston Rods. It has previously been mentioned that buckling is the critical condition for the actuating rod. The required mechanical accuracy and the danger of seizing yield a maximum permissible compressional stress well below critical buckling stress (Fig. 32-35).

![Fig. 32-35.](image)

The piston rod can be considered as a beam clamped at one end, hinged at the other end. The following notation is used:

- $E$ Longitudinal modulus of elasticity
- $I$ Inertia momentum (quadratic momentum) in the rod cross section (this momentum is necessarily constant in the sections of the rod presently under consideration)
- $l$ Length of the rod

The critical buckling strain is\(^1\)

$$\frac{\pi^4 EI}{\alpha^4 l^4}$$

where $\alpha = 0.7$ (for a beam clamped at one end, hinged at the other). The critical buckling stress is, then,

$$\sigma_c = \frac{\pi^4 E}{(\alpha l/\rho)^4}$$

where $\rho$ is the radius of gyration of the cross section of the piston rod. In the mts system, it can be assumed that, for steel, $E = 20,000$ myriapiaces\(^1\) and specify $\sigma_c \ll 100$ myriapiices.

In the example of Sec. 32.4.5, the calculation of the critical buckling strain leads to a section of more than 0.2 cm\(^2\) (a minimum diameter of 5 mm). It is evident that to take any additional strains, such as a bending moment at the hinge, the section ought to be at least 1.5 times that given by the calculation detailed above.

32.7. NUMERICAL DATA SHEETS FOR HYDRAULIC SERVOMOTORS

Tables 32-1 to 32-3 present numerical data for some French and American commercial servomotors.

---

\(^1\) See, for example, P. Dellus, "Résistance des matériaux," Technique et Vulgarisation, p. 175, Paris, 1956.

\(^2\) 1 myriapiere is $10^4$ piezas and about 1 kg/mm\(^2\). It is equal to 100 hectopiezas, 1 hectopieza being 14.504 psi.
### Table 32-1. Performance of Hydraulic Motor SOM-CRH* (Types PAF-1, PAF-6, PAF-12, and PAF-100)†

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>PAF 1</th>
<th>PAF 6</th>
<th>PAF 12</th>
<th>PAF 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displacement per revolution</td>
<td>cm³/rev</td>
<td>1</td>
<td>6</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>Torque:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At 100 hectopieze</td>
<td>cgs</td>
<td>$13.5 \times 10^6$</td>
<td>$81 \times 10^6$</td>
<td>$162 \times 10^6$</td>
<td>$1,350 \times 10^6$</td>
</tr>
<tr>
<td>At 200 hectopieze</td>
<td>cgs</td>
<td>$27 \times 10^6$</td>
<td>$162 \times 10^6$</td>
<td>$324 \times 10^6$</td>
<td>$2,700 \times 10^6$</td>
</tr>
<tr>
<td>Maximum continuous speed at 210 hectopieze</td>
<td>rpm</td>
<td>5,000</td>
<td>3,500</td>
<td>3,000</td>
<td>1,500</td>
</tr>
<tr>
<td>Maximum intermittent speed</td>
<td>rpm</td>
<td>8,000</td>
<td>6,000</td>
<td>5,000</td>
<td>2,500</td>
</tr>
<tr>
<td>Maximum continuous power</td>
<td>kw</td>
<td>1.4</td>
<td>5.7</td>
<td>9.5</td>
<td>40</td>
</tr>
<tr>
<td>Moment of inertia of rotating group</td>
<td>cgs</td>
<td>200</td>
<td>1,300</td>
<td>13,000</td>
<td>450,000</td>
</tr>
<tr>
<td>Weight of motor</td>
<td>kg</td>
<td>0.89</td>
<td>2</td>
<td>11</td>
<td>85</td>
</tr>
<tr>
<td>Number of cylinders</td>
<td></td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

* SOM, Paris, France.
† Hydraulic motor, axial cylinder type.

### Table 32-2. Performance of Hydraulic Motor SAMM M-172*

Class: Hydraulic motor, radial piston (see Sec. 32.1.2):

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum continuous power</td>
<td>19.6 kw</td>
</tr>
<tr>
<td>Time constant</td>
<td>10 msec</td>
</tr>
</tbody>
</table>

Type and performances:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cylinders</td>
<td>7</td>
</tr>
<tr>
<td>Oil pressure (continuous)</td>
<td>100 hpz</td>
</tr>
<tr>
<td>Torque at 100 hectopieze</td>
<td>26 m·kg</td>
</tr>
<tr>
<td>Maximum continuous speed at 100 hectopieze</td>
<td>750 rpm</td>
</tr>
<tr>
<td>Maximum intermittent speed</td>
<td>1,000 rpm</td>
</tr>
<tr>
<td>Leakage flow, at stall, at 100 hectopieze, at 50 ± 5°C</td>
<td>150 cm³/min</td>
</tr>
</tbody>
</table>

Mechanical characteristics:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displacement per revolution</td>
<td>170 cm³</td>
</tr>
<tr>
<td>Moment of inertia of rotating group</td>
<td>200,000 cgs</td>
</tr>
<tr>
<td>Weight of motor</td>
<td>20 kg</td>
</tr>
</tbody>
</table>

Recommended fluid:

- French reference: CF 5ème classe, pamphlet P1571
- Viscosity at 0°C                              | 240 cs†  |
- 20°C                                          | 50 cs    |
- 35°C                                          | 24 cs    |

Measured characteristics curve: see figure below

---

* S.A.M.M., Boulogne, Seine, France.
† cs = centistoke (kinematic viscosity).
<table>
<thead>
<tr>
<th>Notes</th>
<th>Item description</th>
<th>Units</th>
<th>Model designation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>906</td>
</tr>
<tr>
<td>1.</td>
<td>Displacement†</td>
<td>in.³/rev</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cu in./rev</td>
<td>1.56</td>
</tr>
<tr>
<td>2.</td>
<td>Moment of inertia of rotating group†</td>
<td>lb-in.²</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lb-in.-sec²</td>
<td>1.33 × 10⁻⁴</td>
</tr>
<tr>
<td>2</td>
<td>5. Torque at:‡</td>
<td>lb-in.</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lbs</td>
<td>61.4 × 10⁻⁶</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gals</td>
<td>68.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cm³</td>
<td>77.1 × 10⁻⁶</td>
</tr>
<tr>
<td>2</td>
<td>6. Power at:‡</td>
<td>hp</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kw</td>
<td>3.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hp</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kw</td>
<td>8.42</td>
</tr>
<tr>
<td>3</td>
<td>7. Acceleration at:‡</td>
<td>(rad/sec²)</td>
<td>3.4 × 10⁶</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(rad/sec)</td>
<td>5.1 × 10⁶</td>
</tr>
<tr>
<td>4</td>
<td>8. Min rev to stop from max intermittent speed‡</td>
<td>sec</td>
<td>0.189</td>
</tr>
<tr>
<td>5</td>
<td>9. Time constant‡</td>
<td>sec</td>
<td>0.0023</td>
</tr>
<tr>
<td>Notes</td>
<td>Item description</td>
<td>Units</td>
<td>906</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>6</td>
<td>Torque(\times)inertia for:§</td>
<td>lb-in.</td>
<td>1.5 \times 10^9</td>
</tr>
<tr>
<td></td>
<td>3,000 psi (210 bps)</td>
<td>sec. ²</td>
<td>4.4 \times 10^{13}</td>
</tr>
<tr>
<td></td>
<td>4,500 psi (310 bps)</td>
<td>sec. ²</td>
<td>3.5 \times 10^{13}</td>
</tr>
<tr>
<td>7</td>
<td>Vol. under compression:†</td>
<td>in. ³</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>Fixed displacement</td>
<td>cm. ³</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>Variable displacement</td>
<td>in. ³</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td>Variable displacement</td>
<td>cm. ³</td>
<td>4.66</td>
</tr>
<tr>
<td>8</td>
<td>Leakage at 1,000 psi (70 bps)†</td>
<td>in. ³/sec</td>
<td>0.0390</td>
</tr>
<tr>
<td></td>
<td>cm. ³/sec</td>
<td>0.04</td>
<td>1.26</td>
</tr>
<tr>
<td>9</td>
<td>Weight</td>
<td>lb</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>kg</td>
<td>0.95</td>
<td>1.85</td>
</tr>
<tr>
<td>14</td>
<td>Max continuous power:</td>
<td>hp/lb</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>kw/kg</td>
<td>3.81</td>
<td>3.18</td>
</tr>
</tbody>
</table>

† Pumps and motors, either fixed or variable displacement.
‡ Pumps and motors, fixed displacement only.
§ Motors only.
NOTES ON VICKERS MOTORS

1. Speed. A low inlet supercharge pressure is usually required with pumps to prevent cavitation at maximum speeds.

2. Torque and Power. The values under items 5 and 6 refer to the maximum theoretical available output torques and powers of hydraulic motors.

The maximum continuous power rating is at 3,000 psi and maximum continuous speed (item 3); and the maximum intermittent power rating is at 4,500 psi and maximum intermittent speed (item 4).

The maximum intermittent pressure to which these pumps or motors should be subjected is 4,500 psi.

Generally speaking, maximum intermittent power conditions should not be allowed to exist for more than 5 per cent of the duty cycle. The frequency and duration of maximum power conditions directly affect the life expectancy of a unit. Vickers Engineering Division should be consulted regarding these factors for any specific application.

3. Acceleration. The figures given in item 7 are accelerations at the hydraulic-motor output shaft when there is no external load. These acceleration figures are theoretical maximums. They are not obtainable in usual applications because of the existence of external loads. Vickers Engineering Division should be consulted if the application is such that the load permits accelerations in the order of one-half the theoretical figures given in the tabulation.

Since no appreciable, self-generated, opposing force (such as a counter emf) occurs in hydraulic motors, their acceleration is substantially constant over the speed range given in the tables.

In the instances where the external load is primarily inertia, the following formula may be used to calculate the theoretical maximum acceleration \( \alpha \) at the output shaft:

\[
\alpha = \alpha_0 \frac{I_o}{I_o + I_e}
\]

where \( \alpha_0 \) = acceleration figures given in item 7

\( I_o \) = inertia figures given in item 2

\( I_e \) = equivalent load inertia reflected at the hydraulic-motor shaft

In other instances where the load is composed of appreciable components of windage, friction, etc. (in addition to inertia), the acceleration will be altered, assuming a linear relationship of these components to the required forces. Then \( \alpha \) becomes

\[
\alpha = \alpha_0 \frac{I_o}{I_o + I_e} \frac{T_o \pm T_s}{T_0}
\]

where \( T_o \) = torque figures given in item 5

\( T_s \) = a summary of torques required to overcome all load components except inertia. The nature of the load determines the algebraic sign.

4. Minimum Revolutions to Stop. The minimum number of revolutions to stop from maximum intermittent speed is the angular distance the hydraulic-motor shaft will rotate before the back torque overcomes the kinetic energy of the rotating group. For these calculations the back torque is assumed to be produced at 4,500 psi, and it is assumed that there is no load on the output shaft of the hydraulic motor. The formula given in the general notes may be used to calculate the minimum number of revolutions to stop from any particular speed, using the back torque available at any given back pressure.

In the event the external load consists of inertia, friction, and windage components, the formula may be modified as follows:

\[
\text{Min rev to stop} = (\text{rpm})^2 \times \frac{\text{inertia}'}{\text{torque}' \times \frac{\pi}{60 \times 60 \times 388}}
\]
where torque' = T₀ ± Tₛ (the algebraic sign depends upon the nature of torque load)
 inertia' = I₀ + Iₛ

and I₀, Iₛ, T₀, and Tₛ are defined as in item 3.

The minimum time required to stop from maximum intermittent speed is the same as the minimum time required to reach maximum intermittent speed from standstill—when calculated on a theoretical basis as in this tabulation. This means that items 8 and 9 express two characteristics which have the same value: one is shown in seconds and the other in revolutions merely to show these features in different dimensions.

5. Time Constant. The time constant tabulated under item 9 is defined as the time required to accelerate the hydraulic motor from standstill to maximum continuous speed with no external load, using the torque available at 3,000 psi. The same value applies for maximum intermittent speed and 4,500 psi. The time constant for any other speed and torque combination may be calculated by using the formula given in general notes. It should be noted that the time constant defined by this formula is for 100 per cent of the speed under consideration, while other time constants in general use are defined for 63.2 per cent of the speed under consideration (such as the time constant of electric motors).

The time constant of a fixed-stroke hydraulic motor is also a measure of the time required for the motor to reach a specific power output. In addition, the time constant represents the elapsed time during which the motor theoretically can be stopped under conditions given in item 4.

6. Torque/inertia. The torque-inertia ratio is expressed for no-load conditions only. It is a figure of merit evaluating relative responsiveness of hydraulic motors in servo systems. The inertia units used for evaluating this ratio are pound-inches per second per second.

7. Compressibility. Approximate values of the over-all compressibility factor for the range of 0 to 3,000 psi are given below for several commonly used hydraulic fluids. (These values are essentially equal to the incremental compressibility factors at the 1,500-psi point.) They are used for response, spring-rate, and volumetric-efficiency calculations in 3,000-psi systems.

MIL-O-5606 aircraft-type hydraulic fluid:
4.7 × 10⁻⁸ in.²/lb at 90°F
5.6 × 10⁻⁸ in.²/lb at 160°F

DTE light oil:
3.6 × 10⁻⁸ in.²/lb at 75°F
4.3 × 10⁻⁸ in.²/lb at 140°F

MLO-8515:
4.3 × 10⁻⁸ in.²/lb at 76°F

MLO-8200:
4.4 × 10⁻⁸ in.²/lb at 76°F

Over-all compressibility factor = \( \frac{1}{\text{over-all bulk modulus}} \) = \( \frac{\text{(change in volume)}}{\text{(volume at atm pressure)} \times \text{(pressure)}} \) in.²/lb

8. Leakage. The leakage values given are the average values only. They represent the combined line-to-line and line-to-ump leakage, that is, all of the leakage occurring during closed-hydraulic-circuit operation.

9. Weight. Item 13 gives the weight of the unit before oil is added.
CHAPTER 33
AMPLIFIERS AND PREAMPLIFIERS

Summary

1. The general problem.
2. Vacuum-tube amplifier.
3. Relay amplifiers.
5. Machines derived from the Ward-Leonard system.
7. Magnetic amplifier.
10. Transistors.

33.1. THE GENERAL PROBLEM

33.1.1. Definitions and Remarks. The amplifying component, which can often be separated into preamplifier (or voltage amplifier) and amplifier (or power amplifier), can be considered as a power controller. An amplifier possesses two inputs and one output: a control input and a power input linked to an energy source the output of which is to be regulated (power). The output must reproduce "as well as possible" the input function, but at a higher energy level (Fig. 33-1).

The amplifier often possesses a feedback loop, which causes it to resemble a servocontrolled system. However, two special characteristics cause the amplifier to be simpler than the servomechanism of which it constitutes a link:

a. Often the input and output quantities are of the same nature. Of course, electromagnetic coupling and mechanical elements in motion may exist inside the amplifier, but these are intermediate elements, and it will be seen that their inertia does not enter as a fundamental parameter.

b. The phase shift caused by an amplifier in an open loop is relatively small compared to that caused by the motor (in an open loop).

It should also be pointed out that, in general, the method of decreasing the time constants of the motor stage by inserting resistances $R_1$ in order to change $L/R$ to $L/(R + R_1)$ is often illusory; it is, in fact, necessary to increase the steady power gain of the amplifier, which generally costs milliseconds. This will be mentioned later, following the study of machines having crossed fluxes.
Finally, the fact should be emphasized that analysis of the amplifier includes that of its power supply.

33.1.2. Impedances to Be Taken into Consideration. The impedance of the energy source must be as low as possible if a source of voltage is involved and as high as possible if a source of current is involved. If these impedances were respectively zero and infinity, the energy furnished would be proportional to the control signal.

In all actual cases the characteristics of the energy source depend on the power supplied. For example, a nonsynchronous motor, if assumed to drive a "hydraulic transmission" (Chap. 32) at constant speed, is actually subjected to speed variations (rarely exceeding 4 per cent) according to the torque required at the output shaft of the hydraulic transmission. The impedance of this energy source (torque at constant speed) is not zero. It can never be zero for any motor, because torque-speed characteristics independent of speed are never found in the transient state.

The two other impedances which should be pointed out are the input impedance and the output impedance.

We have seen what impedance matching comprises. It is necessary to match the amplifier impedance on both the input and output sides; however, it can well be imagined that in a servocontrolled system it can be very difficult to accomplish complete impedance matching from the detector on through to the load on the system. Starting at one end of the system chain where the impedance of the farthest dipole is fixed, the impedances can be matched step by step, but difficulties in physical construction may arise. Experience shows, and theory confirms, that the matching rules can be obeyed: (a) in the power stages (power amplifier, motor, load) and (b) in the stages having very low energy levels (detectors, correcting networks, voltage amplifier); but that it is often difficult to match these two stages to each other.

It must be observed that in the low-power stages (passive elements) the energy level necessarily declines from input to output sides. One of the objects of impedance matching being to transmit a maximum of energy, the importance of this matching can be clearly seen.

Similarly, in the power stages it is desirable not to downgrade the energy furnished by the source, as would be done, for example, by increasing the proportion of reactive (or fluctuating, Sec. 12.3.6) energy through a poor matching between motor and load.

The junction between the two amplifying stages cited above is accomplished either by electronic coupling (the voltage preamplifier controls the grid of a power tube; it can be assumed that the load impedance of this grid is infinite) or by an impedance-matching stage (transformer, cathode, etc.) inserted between the preamplifier and the power amplifier.

33.1.3. Reversibility. It was mentioned in Chap. 30 that a servomotor should act as a brake as well as a motor. What becomes of the energy furnished by the motor to the amplifier?

Reversibility of the system would require that the amplifier transmit

---

1 It is also desirable for the power amplifier to be matched to the energy source.
this energy toward the source. However, linearity can be maintained even if this energy is dissipated in the amplifier; in other words, even though the energy source must be considered as being a part of the servo-controlled system, the reversibility of the source-amplifier system is not necessary.

This is generally the case: the energy furnished by the motor is dissipated in heat either in the amplifier or in the "regulators" (voltage, pressure, etc.) associated with the source.

33.1.4. Threshold, Noise Level. All that has been said earlier (Chaps. 12 and 22) concerning thresholds also applies to amplifiers. However, since the amplifier is an active component, it is desirable to indicate what quantities should be taken into consideration (Sec. 33.1.1). The input threshold is the smallest entering quantity which can give rise to an output signal, or (as in the case of thyatrons) the smallest entering signal which can trigger the system. Hysteresis phenomena can then be such that even if the magnitude of the input signal diminishes, the output signal may still be controllable.

Finally, another important consideration is the noise level of the amplifier. The energy source should itself have as low a noise level as possible. This condition is difficult to attain in a rigorous manner unless dry or storage batteries are used. When an electronic amplifier is employed, the different voltages required (plate voltage, grid voltage, heating, etc.) must be very well filtered (except perhaps for heating those tubes requiring indirect heating). In particular, the a-c components at 60 and 180 cps must be less than 1 per cent of the d-c voltage.

Nevertheless, as was pointed out in Sec. 29.1.4, the amplifier has a noise level of its own which is due to the agitation of the electrons in the resistances, to thermionic emission, or to the transmission of "water hammer" if the device is hydraulic. It is this background noise which limits the gain of the amplifier. Assuming a perfectly linear amplifier, the absolute value of the noise, for a certain gain, becomes too great, that is, it becomes superior to the apparent threshold of the system element which follows the amplifier.

By apparent threshold is meant either the real threshold of the apparatus or the conventional threshold fixed by the design specifications. For example, if the amplifier drives a motor\(^1\) directly, the acceptable noise level at the amplifier output corresponds to the precision with which the output element must be positioned; the actual threshold of a motor is linked to the number of notches and segments on its commutator.

33.1.5. Classification. Amplifiers can be classified in many ways. We will briefly study direct-current amplifiers; i.e., those requiring no a-c electrical source (the supplies may take their energy from the a-c line, but it can always be assumed that the high voltage, bias, and heating voltages are obtained from dry or storage batteries).

We will discuss the vacuum-tube amplifier (triodes or pentodes); relays, both vibrating and nonvibrating; the Ward-Leonard system; and similar systems: amplidyne, Rototrol, and metadyne.

\(^1\) It is assumed that the motor exhibits a perfectly linear torque-current characteristic and that, in particular, there is an infinite number of segments on its commutator.
Next mixed amplifiers will be examined, i.e., those requiring an a-c energy source and in general, also, a d-c source, this latter being obtained independently of the a-c circuit or by "rectification" of a part of the a-c energy. The following will be mentioned: vacuum-tube amplifiers, gas-tube amplifiers (thyatrons), and magnetic amplifiers. The output current from these elements may be either alternating or direct. If it is alternating and if the power element consists of a d-c motor, it becomes necessary to rectify the output current from the amplifier. This is accomplished by means of rectifiers or demodulators, which will be discussed.

Hydraulic or pneumatic preamplifiers will also be discussed.

The choice of an amplifier depends greatly on the stages preceding it and, more importantly, on those following it. However, as has been said in preceding chapters, the choice of the type of motor is not generally fixed during the preliminary design; it may then happen later that the design and construction of an amplifier for the type of motor initially chosen turn out to be too difficult. In this case the problem must be reexamined, dealing from the outset with the combined amplifier and motor.

Above all, the final goal of the servo system should not be lost from sight. In particular, if the system is intended for aeronautical service, the use of electronic tubes may not be the best solution (owing to their fragileness, their requirement of plate voltage and of voltage regulators, their high load impedances which necessitate the use of transformers, etc.) if more rugged solutions are acceptable.

33.2. VACUUM-TUBE AMPLIFIERS

These amplifiers, well known in telephone applications, are not always sufficiently powerful for servo applications; they are then used as pre-
We refer the reader to books on radio engineering for tube characteristics and to Secs. 29.1.2 to 29.1.7 for a discussion of the various aspects of noise.

33.3. RELAY AMPLIFIERS

33.3.1. The Classical Relay. The relay can be considered a nonlinear amplifier which has a very high power gain. The amplification of the relay may be defined as the ratio between the power which can be cut off by the relay and the normal operating power of the relay windings. There exist relays for which the power amplification is as high as $10^4$ or $10^6$.

The simplicity of this amplifier is counterbalanced by the fact that the output signal takes the form of waves with relatively steep fronts, modulated in duration. Figure 33-3 illustrates the case of the polarized relay (input voltage $V_i$, output voltage $V_o$, and $S_0$ the threshold of the relay). The impedance of the source (Fig. 33-4) may be chosen for good matching with the following stage.

On the other hand, the current which corresponds to a square-wave signal such as that of Fig. 33-3 is rarely directly usable. In general, it is rectified and filtered. The average value of the current after rectification but before filtering is equal to the algebraic sum of the areas described (Fig. 33-3). The filtering is intended to smooth the resulting current and thus reduce the intensity of the harmonics to which the steep fronts give rise. This point will be discussed again later.

A few circuits may be suggested. Figure 33-5 shows (a) a motor control ($x$ volts) involving a single battery ($x$ volts), a primary single-throw relay, and two power relays; (b) the same control using a single double-throw relay; (c) control of a motor of $x/2$ volts using a battery of $x$ volts and a single simple inverting relay.

33.3.2. Vibrating Relay. In some setups the vibrating relay is employed. Its principle is as follows: A relay with several windings con-
sists of a contact mounted on a reed the vibration of which is maintained (an a-c or d-c device can be used); the movable contact oscillates between two fixed contacts which it touches for equal periods of time during each cycle. A winding (sometimes called a deflecting winding), modifies the

![Diagram](image)

(a)

![Diagram](image)

(b)

![Diagram](image)

(c)

**Fig. 33-5.**

mean position about which the reed oscillates; the contact time is then not the same on the right and on the left, and as a result the amounts of electric energy sent into each of the two circuits are not equal. These quantities can be integrated in a circuit of the type shown in Fig. 33-6,
which amounts to supplying the circuit elements which follow with direct current.

The advantage of this device, as compared to the nonvibrating relay, is that it allows quasi-continuous control of the energy delivered by the battery $B$. It remains, however, a delicate circuit element the amplification coefficient of which never exceeds 1,000 and is usually less than 100. On the other hand, it is often used in integrating circuits; the outputs consist of unsymmetrical, interrupted currents of nearly constant frequencies (Fig. 33-7). The average value of the output current can be directly extracted from the vibrating-relay output current by a motor (which acts as an integrator) having a time constant which is large compared to the frequency of the output current (Fig. 33-8).

To return to the classical relay, in addition to the characteristics of the relay, the opening and closing time delays must be known. These two quantities, which are sometimes erroneously called "time constants," are in general distinctly different, and they depend on the magnitude of the input function (control signal) of the relay, which is a nonlinear system. Similarly, it should be remarked that a relay having a time delay of $x$ msec for a given step excitation will not necessarily follow a sinusoidal current of frequency near $1/x$ cps, even if the amplitude of the sinusoidal control current is such that the energy transmitted (Fig. 33-9) to the relay per half cycle is equal to the energy transmitted by the unit step during a time equal to the delay of the relay (the two cross-hatched areas are then equal).
On the other hand, the *vibrating relay* can, under certain conditions, be considered a linear element the transfer function of which would be

\[
\frac{V_r}{V_s} = \frac{K}{1 + \tau s}
\]

The time constant \(\tau\) is of the order of the relay vibration period. It is to be noticed that this quantity is relatively large by comparison with the "delay" of the relay used directly for an oscillation frequency of 50 cps; the time constant is about 20 msec.

![Diagram of a vibrating relay](image)

**Fig. 33-8.**

![Graph showing energy delivery](image)

**Fig. 33-9.**

Generally speaking, the maximum frequency at which the relay can be made to oscillate is 400 cps, which yields a minimum time constant of 2.5 msec.

### 33.4. ROTATING AMPLIFIER. WARD-LEONARD SYSTEM

If it is desired to control a d-c motor of more than 100 watts with a good time constant and good linearity, the devices so far considered are not suitable. Controlling the speed of a motor giving, for example, constant torque becomes a difficult problem when a range of measurement (in the steady state) of over 50 is sought. Speed control to within 1 per cent is already difficult to achieve. One is led to use generators which are so arranged that their time constants (due to the magnetic energy stored in
them) are very small. We shall examine below the different types of generators.

33.4.1. Rotating Amplifier, Voltage-Voltage Type. Let us first consider a separately excited generator, of classical construction, the field (resistance $R$ and inductance $L$) of which is supplied by the input voltage $v_e$, the rotor being driven at constant speed (Fig. 33-10).

The emf $v_r$ is proportional to the control voltage $v_e$ in the steady state. In the transient state the following holds:

$$V_r = k \frac{V_e}{R + L_s} \quad (33-1)$$

However, if the circuit following the generator is a loop closed on an impedance $Z_r$, the relation (33-1) no longer holds. The steady-state power gain is about 100, but the time constant can be large if the source impedance is not high compared to that of the field.

This elementary device is rarely used as a voltage amplifier.

![Fig. 33-10. Rotating amplifier (voltage-voltage type).](image)

![Fig. 33-11. Rotating amplifier (voltage-current type).](image)

33.4.2. Rotating Amplifier, Voltage-Current Type. Let us now consider the case in which the generator is connected in a loop of impedance $Z_r$ (Fig. 33-11). Let $r$ and $l$ be the total resistance and self-inductance of the circuit including the generator armature and the load (impedance $Z_r$). The following result is readily obtained,

$$\frac{I_r}{V_e} = \frac{k_1}{rR(1 + \tau_1\delta)(1 + \tau_2\delta)} \quad (33-2)$$

where $\tau_1 = L/R$ (field time constant)  
$\tau_2 = l/r$ (armature-circuit time constant)

and $k_1/Rr$ is the gain.

The transfer loci represented by Eq. (33-2) are traced in Fig. 33-12; the transfer function is written

$$\frac{I_r}{V_e} = k_1 \frac{rR[1 + (\tau_1/\tau_2)ju](1 + ju)}{rR(1 + \tau_1\delta)(1 + \tau_2\delta)}$$

where $u = \tau_2\omega$.  

33.4.3. Rotating Amplifier, Voltage–Current Type, with Compensating Windings. Let us now study the effect of a “compensating” winding (the meaning of which will be given later) added to the field coils and supplied in series with the output impedance $Z_r$ (Fig. 33-13); in this figure and the following ones the geometric relation between the axes of the magnetic fields and the line of the brushes has been preserved. Let $M$ be the coefficient of mutual inductance (algebraic) between the two windings $A$ and $B$. It will be assumed that they are wound in such a way that in the steady state the flux due to $B$ is in the opposite direction to that due to $A$. Let $e_r$ be the emf of the armature; the equations governing the system are

$$
\begin{align*}
\nu_s + M \frac{di_r}{dt} &= Ri_s + L \frac{di_s}{dt} \\
\Phi &= k_1 i_s - k_2 i_r \\
e_r &= k \Phi + M \frac{di_s}{dt} \\
e_r &= ri_r + l \frac{di_r}{dt}
\end{align*}
$$

(33-3)

Transforming to Laplace variables and eliminating $\Phi$ and $e_r$ yields the following transfer function for $I_r/V_s$:

$$
\frac{I_r}{V_s} = \frac{kk_1 + Ms}{(R + Ls)(r + kk_2 + ls) - (kk_1 + Ms)Ms}
$$

which can be written as

$$
\frac{I_r}{V_s} = \frac{kk_1}{R(r + kk_2)} \frac{1 + \tau_3 s}{(1 + \tau_1 s)(1 + \tau_2 s) - Mkk_1s(1 + \tau_3 s)/R(r + kk_2)}
$$

(33-4)

where

$$
\tau_1 = \frac{L}{R}, \quad \tau_2 = \frac{l}{r + kk_2}, \quad \tau_3 = \frac{M}{kk_1}
$$
Comparison between Eqs. (33-2) and (33-4) shows that the term 
$1 + \tau_2 s$, in the numerator of Eq. (33-4), plays the role of phase lead. At
high frequencies the system acts as if

$$\frac{I_r}{V_o} \approx \frac{k k_1}{R(r + k k_2)} \left[ \frac{\tau_2 / s}{\tau_1 \tau_2 - M k k_1 \tau_2 / R(r + k k_2)} \right]$$

has phase tending toward $-\pi/2$ [whereas that of the system represented
by Eq. (33-2) tends toward $-\pi$]. It is for this reason that the series
winding has been called the compensating winding.

Moreover, if the second term of the denominator in Eq. (33-4) is ne-
glected, which is generally permissible since the coefficient $M$ is small, it
can be seen that the time constant $\tau_2$ due to the armature is smaller than
that which would exist in the absence of the compensating winding. It
can be considered that this winding gives rise to a reaction; the diagram
(Fig. 33-14) representing Eqs. (33-3) and those following consists, in

![Fig. 33-14.](image)

fact, of three loops. The principal loop is that shown in heavy lines.
The dotted-line loop is not, strictly speaking, a loop, since there direct
action is involved.

### 33.4.4. The Simple Ward-Leonard System.

The two preceding
devices, taken as intermediary amplifiers, are used rather rarely; on the
other hand, when directly coupled to the motor of the servocontrolled
system, they constitute a very important class of power stages. Origin-
ally, the control of a motor, through its armature, by control of the cur-
rent produced by a generator whose field was regulated, was employed
to obtain either motors which were controllable over a large speed range
while furnishing a heavy torque or motors subject to varying loads which
had nevertheless to operate at relatively constant speed (mine hoists,
cable cars, rolling mills, etc.).

The generator is assumed to be driven at a constant speed. As a
matter of fact, the power variations called for at the output shaft neces-
sarily give rise to generator speed variation, since no source with infinite
internal impedance (ratio $\Omega/C$) exists.

It is difficult to take these speed variations into account; it would in
fact be necessary to consider the two components $\omega_r$ and $C_r$ (speed and
torque) of the output signal as well as the efficiency of the system (which
is itself a function of these quantities). The use of a flywheel fixed to the generator shaft is a poor solution which, in present-day engineering, should be systematically eliminated.¹

![Fig. 33-15. Ward-Leonard system.](image)

The Ward-Leonard system constitutes a perfectly adapted amplifier-motor group which shows good performance. It is, in fact, composed of three parts (Fig. 33-15):

1. A motor, generally three phase and nonsynchronous, which constitutes the energy source. It is assumed that it rotates at constant speed and furnishes the necessary power.

2. A generator the field of which is supplied by the control voltage (input function); the armature being fixed to the shaft of the motor of (1).

![Fig. 33-16.](image)

3. A motor (strictly speaking, motor element of the servocontrolled system) the armature of which is supplied by the output current of the preceding generator. The field is supplied at constant voltage. On the shaft of this motor are fixed those elements which are to be controlled (it is assumed that they correspond to an inertia $J$ and a damping $f$).²

The corresponding electrical diagram is shown in Fig. 33-16.

¹ M. Darrieus has pointed out that the minimum required moment of inertia of a turbo-alternator, for purposes of regulation, fell to one-quarter of its formerly accepted value as soon as it became clear that the problem should be considered as a servomechanism problem. (Mémoires de la Société des Ingénieurs Civils de France, 1952, fasc. 1.)

² A restoring torque could also be added.
Let

\[ R, L = \text{resistances and self-inductance of generator field} \]
\[ r, l = \text{total resistance and inductance of armature circuit, including armatures themselves} \]
\[ \Phi_g, \Phi_m = \text{field fluxes of generator and motor} \]
\[ e_g, e_m = \text{emf of generator and back emf of motor} \]

The following relation then holds:

\[ I_s = \frac{V_s}{R} \frac{1}{(1 + \tau_1 s)} \]

where

\[ \tau_1 = \frac{L}{R}, \quad \Phi_g = \frac{k_1 V_s}{R} \frac{1}{(1 + \tau_1 s)} \]

The generator emf is proportional to the field flux (for constant armature speed), that is,

\[ e_g = k_2 \Phi_g \]

The back emf of the motor is proportional to the speed of the motor armature (field flux \( \Phi_m \) constant):

\[ e_m = k_3 d\theta_r/dt \]

The current flowing between the two armatures is, therefore,

\[ I = \frac{k_2 \Phi_g - k_3 \Phi_m \theta_r}{r + l s} \]

Let us set \( \tau_2 = l/s \). The preceding equation can then be written in the form

\[ I = \frac{k_1 k_2 V_s - k_3 R \Phi_m (1 + \tau_1 s)}{R r (1 + \tau_1 s)(1 + \tau_2 s)} \] (33-5)

Finally, the dynamical equation, applied to the motor, gives

\[ C = k_3 i = J \frac{d^2 \theta_r}{dt^2} + f \frac{d\theta_r}{dt} \]

which, after Laplace transformation and writing \( \tau_2 = J/f \), becomes

\[ I = \frac{f_s}{k_3} (1 + \tau_2 s) \theta_r \] (33-6)

Elimination of \( i \) between Eq. (33-5) and Eq. (33-6) yields

\[ \frac{\Theta_r}{V_s} = \frac{k_1 k_2 k_3}{R rf} \frac{1}{s (1 + \tau_1 s) [(1 + \tau_2 s)(1 + \tau_2 s) + k_3/s/rf]} \]

The factor in brackets in the denominator of the last equation shows that the transient response of the system may be oscillatory.
33.4.5. Ward-Leonard System with Compensating Winding. Let us imagine the effect of a series winding acting on the excitation of the generator (Fig. 33-17) and wound in such a way that its ampere-turns are opposed to those of the field winding.

![Figure 33-17](image)

To the notation of the preceding section we add \( M \), the coefficient of mutual inductance between field windings. The following relations hold:

\[
\begin{align*}
V_e + M_s I &= (R + L_s) I_e \\
E_z &= k_1 I_e - k_2 I + M S I_e + k_3 S \Theta \\
E_z &= (r + l_s) I
\end{align*}
\]

whence the transfer function

\[
\Theta_r = \frac{k_4 (k_1 + M s)}{(r + k_2 + l_s)(J s^2 + f s)(R + L_s) - k_5 k_4 (R + L_s) s - M s (J s^2 + f s)(k_1 + M s)}
\]

Comparison of this transfer function with that obtained in the preceding case (without compensating winding) shows that here the first-degree term in the numerator plays the role of phase advance. Other things being equal, a greater bandwidth can be expected with the Ward-Leonard system when it is equipped with compensating winding.

![Figure 33-18](image)

33.4.6. Other Machine Groups. If the required amplification exceeds the possibilities of the Ward-Leonard system, a first stage consisting of a rotating amplifier, generally called an exciter (Sec. 33.4.2), can be employed; the role of this device is to create the field for a Ward-Leonard group (Fig. 33-18). For stability reasons, it is necessary to introduce, again, compensating windings; we shall not present the corresponding circuits or the accompanying transfer functions, which are easily obtained in the same way as above.
33.5. MACHINES DERIVED FROM THE WARD-LEONARD SYSTEM

33.5.1. General. For servocontrolled systems having high performance, if a single stage (generator and motor) does not suffice, designers are led to using machines having crossed flux in place of a cascade of Ward-Leonard machines. Chronologically speaking, the first crossed-flux machine studied for use as an amplifier in a servocontrolled system\(^1\) was Pestarini’s metadyne. This machine\(^2\) was not greatly developed in Europe: toward 1940 the General Electric Company in the United States put on the market a machine which is very similar to the metadyne: the

![Diagram of a crossed flux machine]

**Fig. 33-19. Metadyne, or amplitidyne, machine.**

amplitidyne. We shall examine these two machines together; at the end of the discussion their differences will be pointed out.

33.5.2. Metadyne and Amplitidyne. These are essentially generators with two pairs of brushes, one pair of which is short-circuited. The generator rotor is driven at constant speed. Compensating windings may exist on one or the other of the circuits associated with the pairs of brushes (Fig. 33-19). The principle employed in these machines is that of armature reaction, which leads to crossed fluxes.

1. Schematic Diagram and Operation. A so-called constant speed motor drives the rotor \(G\) which is wound exactly like that of a d-c motor. Four brushes rub on the commutator at \(AA'\) and \(BB'\). The control field \(C\) (resistance \(R_1\), inductance \(L_1\)) is excited by the control voltage (input quantity \(E\) to which correspond a current \(i_s\) and a flux \(\Phi_s\) carried by axis

\(^1\) Crossed-flux machines have existed for some time; the earliest applications were for generating current at constant voltage in cases where the generator speed varied considerably (automobile generators) or where it even reversed direction (railroad-car lighting with a storage-battery buffer).

\(^2\) Originally constructed by Alisthm (France), about 1930, under the direction of Pestarini.
1. An emf $e_2$ is therefore generated between the brushes $AA'$; the external circuit linking these two brushes is practically constituted by a "short circuit" (of resistance $R_2$). Occasionally in this circuit there may be a winding (of inductance $L_2$) the purpose of which is to create a flux perpendicular to axis 1. As a result of the relatively high current which circulates in this circuit, the armature reaction of this generator is very high; this means that, because of the rotor winding, this current gives rise to a flux $\Phi_2$ carried by axis 2 (it may also be caused by a winding added to the external circuit).

The flux $\Phi_2$ gives rise to an electromotive force at the terminals of the brushes $BB'$. The external circuit connected to these brushes includes the load impedance, in practice, a d-c motor with armature control. This circuit also includes a compensating winding $F$—always present in this case—the role of which is to compensate the armature reaction due to the current $i_1$ which flows in the output circuit of the amplidyne (or metadyne).

It should, in fact, be carefully noted that the armature reaction due to the current flowing in the circuit $ADA'$ should under no circumstances be compensated by any device (compensating windings, offset brushes, etc.), whereas in order to obtain maximum efficiency at the second stage (circuit $BFB'$), this secondary armature reaction should be compensated. It is the percentage of compensation which differentiates the metadyne from the amplidyne:

a. The metadyne is slightly undercompensated (hence there exists a positive reaction, and a greater resulting gain, in the armature of the machine).

b. The amplidyne is, on the contrary, 100 per cent compensated.

2. Advantages over the Ward-Leonard System. It can be seen that two-stage amplification has here been obtained with the same number of rotors. Let us take up once more the explanation of the operating principle of the machine, assuming it to be an amplidyne. The emf $e_2$ appearing at the terminals $AA'$ is the same, whether or not a circuit be connected to $BB'$; all the relations written for the generator are linear, the superposition principle is applicable, and therefore two emf systems can coexist without interaction in the armature. It should also be pointed out that the voltages created by the field flux $\Phi_1$ at two points such as $B$ and $B'$ are equal; it then follows that, for the winding $C$, the circuit $BFB'$ has no influence on the emf across $AA'$.

The crossed-flux machine is also more advantageous than the use of two cascaded generators (Fig. 33-18). In fact, if the emf across the brushes $AA'$ were to supply the field of a second generator, there would be introduced a time constant, large compared to that of the circuit $ADA'$ (Fig. 33-19), the latter constant becoming practically that of the armature. Because of this fact the machine employs a flux which, in single-flux machines, must be compensated in order to avoid mechanical complications (brush advance, etc.). Herein lies the very great advantage of crossed-flux machines and, at the same time, the explanation of their low time constants and high gain.
3. Operating Equations. With the aid of the notation of the figure, we may write

\[ \Phi_\omega = k_1 I_\omega = \frac{k_1 E}{R_1 \tau_1 s + 1} \]

where \( \tau_1 = L_1/R_1 \) and \( k_1 \) is the characteristic constant of the magnetic circuit including the field \( C \).

In the case of the amplidyne, since the compensation is complete, the emf appearing at the terminals \( AA' \) is such that

\[ e_2 = N k_2 \Phi_\omega \]

where \( N \) is the constant drive speed of the rotor and \( k_2 \) is a geometric constant of the machine. Then let \( R_G \) and \( L_G \) be respectively the resistance and inductance of the armature as they were defined in Chap. 31 and \( R_2 \) be the very low resistance of the “short circuit.” (If there is a winding on this circuit, \( R_2 \) is no longer very low, and an inductance \( L_2 \) must be added.)

The current in the short-circuited turn is

\[ I_2 = \frac{E_2}{(R_G + R_2)(\tau_2 s + 1)} \]

where \( \tau_2 = L_G/(R_G + R_2) \) and the flux \( \Phi_2 \) has the value \( \Phi_2 = k_3 i_2 \) \((k_3 \) being a constant of the machine). The flux \( \Phi_2 \) induces an emf \( e_3 \) in the brushes \( BB' \) which has the value \( e_3 = N k_2 \Phi_2 \).

Let \( R_3 \) and \( L_3 \) be respectively the total resistance and self-inductance of the operational circuit, i.e., the resistances and self-inductances of the external circuit including those of the motor \( M \) plus those of the generator armature (amplidyne). Let \( e_m \) be the back emf of the motor \( M \); this is proportional to the speed \( d\theta_r/dt \) of its armature. If \( k_4 \) designates the back-emf constant of the motor, the current \( i_3 \), flowing through its armature is, in terms of Laplace variables,

\[ I_3 = \frac{E_3 - k_4 s \Theta_r}{R_3 (1 + \tau_3 s)} \quad \text{with} \quad \tau_3 = \frac{L_3}{R_3} \]

Finally, the motor couple \( C_m \) is proportional to \( i_3 \); thus, if \( k_5 \) is the proportionality constant, \( C_m = k_5 i_3 \). Suppose that the output is constituted by an inertia and a friction couple, then

\[ C_m = (J s^2 + f s) \Theta_r \]

The amplidyne transfer function, obtained by eliminating \( \Phi_\omega, i_2, \Phi_2, e_3, i_3, \) and \( C_m \) from the preceding equations, is

\[ \Theta = \frac{N^2 k_1 k_2^3 k_3 k_5}{R_1 (R_2 + R_G)} \frac{1}{s (1 + \tau_1 s)(1 + \tau_2 s)(J s + f)(1 + \tau_3 s)R_2 + k_4 k_3} \]

In the case of the metadyne, since the compensation is not complete, the expression for the flux along axis 1 is \( \Phi_\omega + k_6 b_1 \), where \( k_6 \) is a constant of the machine. We may then revert to the case of Sec. 33.4.3 (rotating amplifier with compensating winding).
The machine remains linear. The transfer function can be calculated in the same way as for the amplidyne.

4. Remark. As was pointed out in Sec. 33.1.1, any increase of the steady-state power gain requires, in general, an increase of several milliseconds in the time constant. For a machine with given mechanical properties (in particular, the rotation speed), an increase of the emf at the output terminals (BB' of Fig. 33-19) necessitates a greater primary field flux and hence requires more turns on the control coil C. Now, the inductance of a coil increases as the square of its number of turns, whereas the resistance grows directly with the number of turns; hence the time constant increases, to a first approximation, as the number of turns.

5. Second Remark. Professor A. Tustin has pointed out\(^1\) that for output powers of less than a kilowatt there exist machines having gains as high as 100,000.

33.5.3. Another Application of the Metadyne. The metadyne was originally designed for the purpose of transforming a constant d-c voltage supply into one furnishing constant current (Chap. 31). Constant d-c voltage sources are relatively numerous: d-c supply lines, where they

exist; dry and storage batteries; and a large number of classical types of dynamo. On the other hand, there usually are not to be found many constant-current sources. We have encountered a case where such a source was required: the electric motor with field control (Sec. 31.3). Large-scale servocontrolled systems in which the constant-voltage supply has been replaced by a constant-current supply have been designed. Let us consider a metadyne without control-field winding (Fig. 33-20). The brushes AA' are supplied at constant voltage; the current flowing in the armature acts like the short-circuit current which appeared in the regular arrangement. The output current is taken off as before through the brushes BB'. The machine thus obtained produces a constant current \(I_2\) which corresponds to the constant voltage \(E\). This result can be found by intuitive reasoning, but it is simpler to make use of the principle of the conservation of energy.

If the machine is not compensated, the output current \(I_2\) creates, through armature reaction, a proportional emf \(e_2 = kI_2\), and the emf across the terminals BB' is proportional (with the same proportionality constant) to \(I_1\), thus, \(e_2 = kI_1\).

The power absorbed by the machine is $e_1 I_1$, and the power furnished to the load is $e_2 I_2$, that is, in view of the relations above, $k I_1 I_2$ for the two powers. The power transmitted to the load is equal to that taken from the source; the current $I_2$ is, therefore, constant because $e_1$ is constant also (at least to a first approximation). Note that the power required to turn the rotor is very low—it corresponds to the losses in the armature.

33.5.4. The Rototrol (Westinghouse). The Rototrol is a four-pole machine employing a principle similar to that of the amplidyne or metadyne in that it is a two-stage crossed-flux rotating amplifier. The control

![Fig. 33-21. Rototrol.](image)

![Fig. 33-22.](image)

current $i_2$ creates a flux $\Phi_1$ (Fig. 33-21) in only two opposite poles. As in all machines with four brushes, a current is produced in the two brushes $P$ and $Q$. In order to compensate the armature reaction, this current $i_2$ flows through the two poles perpendicularly to the exciting poles. This corresponds to the first amplifying stage. The second stage corresponds to a classical four-poled rotating amplifier (Fig. 33-22); the armature of this amplifier is the same armature that was used in the first stage, and
the four poles are the poles already used in the first stage. The windings \textit{ABCDEFGH} shown grouped together in Fig. 33-21 are "correctly oriented" in Fig. 33-22.

To prevent interaction between the two stages, the poles are so wound that the total flux produced by a pair of opposite poles is zero; the output current then has no influence on the excitation due to the control current (Fig. 33-21). The output current of this amplifier is taken off between the point \textit{R} and the midpoint of the second-stage field coils in order that the windings of the latter amplifying stage be traversed by opposing currents. (It will be recalled that the points \textit{R} and \textit{S} in four-pole generators are at the same potential.) The short-circuit \textit{RS} serves to balance the currents; this is not necessary, however, in this second stage of amplification. It should nonetheless be emphasized that it is of primary importance for the first stage.

33.5.5. Other Machines. Conclusions. The Ragonot Servodyne (France) can be used as a motor, generator, or amplifier. The gain at zero frequency is between 100 and 800, according to the application made. The field time constant then lies between 0.08 and 0.28 sec. The Macfarlane Magnicon (Great Britain) is a four-pole crossed-flux machine. At present there exists a considerable number of flux machines, either crossed flux or of greater complexity (four poles, etc.). The basic idea of all these machines is that the output current from the first amplifying stage is utilized to create a second flux within the armature itself, without using an intermediary field. As we have pointed out, such devices give:

a. Much lower time constants than through the use of two Ward-Leonard-type stages
b. Very high steady-state gains

The ruggedness of such machines depends on the commutators. At present, bearing life can be considered to be indefinitely long and bearing maintenance zero (permanently lubricated bearings).

33.6. MIXED ELECTRONIC-TUBE AMPLIFIERS

33.6.1. Vacuum-tube Amplifiers. It is known that the amplification of a modulated a-c voltage is simpler than the amplification of a d-c voltage. The reader is referred to books on radio engineering for description and operating principles of these amplifiers. Figure 33-23 illustrates a system consisting of two corrective networks and a two-tube a-c amplifier (6SN7 tubes). The input voltage comes from a synchronous transducer and is therefore alternating. (The stages which follow contain a thyratron control circuit for the two-phase motor; the system is used as a remote control for a radio antenna.)

We shall not go into further detail on this type of amplifier, but we may mention that, as a result of the discontinuous cathode emission (electrons), the output current necessarily contains a certain amount of noise.

\footnote{As opposed to gas tubes (thyratrons).}
Moreover, as was pointed out in Sec. 29.1.4, an electrical resistance is a source of noise the effective magnitude of which is proportional to the magnitude of the resistance. Grid polarization resistances are generally very large (from $10^4$ to $10^6$ ohms). These two sources of noise are the ones which limit the gain of an amplifier.

33.6.2. Thyratron Tube Amplifier. The output-input relation of a thyratron corresponds to charge in coulombs ($\int idt$) expressed as a function of the grid voltage. The form of the output current generally requires the use of filtering circuits. We refer the reader to books on radio and electrical engineering for the description, operation, and theory of these tubes.

1. Principle. It will merely be recalled here that two types of thyatrons exist: the hot-cathode (mercury-vapor) tube and the cold-cathode tube (containing a mixture of several inert gases). In these triodes (Fig. 33-24) the plate is necessarily supplied from an a-c source. When the discharge has once been triggered, it can be interrupted only by establishing a negative voltage on the plate; the grid voltage has no further effect after triggering.\(^1\)

The triode is employed as follows: The (d-c) grid voltage is made very slightly positive (through impulses) at the time the discharge is to be triggered. Starting at this time, a quantity of electricity, determined by the triggering instant with respect to the a-c plate-supply voltage, will necessarily flow. The plate current will stop flowing at the instant that the plate voltage becomes negative; the time of this event is independent

\(^1\) Certain special circuits wherein the plate voltage is normally positive direct current exist; tube action is cut off voluntarily by bringing the plate to zero voltage through the use of an auxiliary device. These special setups will not be discussed here.
of the control voltage (Fig. 33-25a). It can then be seen that the resulting current is pulsed, with an average value not zero. This average value depends on the control voltage.

There are other ways of controlling a thyratron; the more usual ones are the following:

$\alpha$. The grid voltage is alternating and of low amplitude (a few volts) which is shifted out of phase with respect to the plate voltage. When this voltage reaches the operating threshold of the thyratron, the latter triggers, and the polarity and voltage of the grid play no further role until the plate potential becomes negative once more (Fig. 33-25b).

In fact, as will be seen in the example given under (b), the triggering threshold depends on the plate voltage, so that the instant of triggering is not strictly proportional to the phase lag. This amounts to considering (Fig. 33-25c) that the voltage defining the threshold for triggering is not a
straight line but is a (periodic) curve varying inversely in direction to the plate voltage.

b. The grid is subjected to positive impulses which are synchronous with the plate voltage but are out of phase with the control signal (Fig. 33-25d).

c. The grid is raised to a positive or negative d-c voltage (with respect to the triggering threshold) (Fig. 33-25e). This device is often employed to control electric motors by means of thyratrons (the currents are pulsating, the motor acting as a filter).

2. A Typical Thyatron Tube. As an example, we present the main characteristics of a thyratron 3C23, a triode having an atmosphere of mercury vapor and another inert gas. For a plate voltage of 25, 100, or 500 volts, the grid voltage above which discharge occurs is 0, −2.5, or −4.5 volts, respectively.

The ionization time is about 10 μsec, and the deionization time is about 1,000 μsec. This time limits the maximum frequency at which the thyratron can be used; for this tube, the deionization time is around 1 msec and, therefore, the maximum admissible frequency is 500 cps.

The maximum anode currents are thus limited (1) for a frequency below 25 cps, to 3 amp and (2) for a frequency above 25 cps, to 6 amp.

Average anode current (during 5 sec or more) is (1) below 210 cps, 1.5 amp; (2) between 210 and 400 cps, 1.0 amp.

Maximum peak current during 0.1 sec: 120 amp.

Figure 33-26 shows the zone in which the characteristics of such a tube lie.

3. Examples of Circuits. Frequently the push-pull circuit is used (Fig. 33-27a and b). However, a difficulty arises in preventing the simultaneous firing of the two thyratrons. If, moreover, it is desired that the threshold of the amplifier be as low as possible, the adjustment of the grid polarizations can become delicate and difficult to maintain as the result of temperature or supply-voltage variations. A dependable device consists in using a polarized relay having a very low time constant (1 msec or less) to trigger the thyratrons. With this, a single thyratron can be triggered (Fig. 33-27c).

Remark. One of the reasons for favoring the use of 400 cps for airplanes is the fact that thyratrons function well at that frequency. On some occasions, thyratrons are constructed to have a very short cutoff time in order to permit their use with 1,000-cycle alternating current.

4. Threshold of a Thyatron. In order that the ionization produced by the grid triggers the tube, the plate voltage must be above a certain minimum value. It follows that the ratio of threshold to saturation is not as high as in vacuum tubes, for example. If \( s_1 \) is the thyratron threshold and \( P \) the amplitude of the plate voltage, it may be seen that
(a) One-direction rotating controlled motor

(b) Two-direction controlled motor

(c) Security arrangement

Fig. 33-27. Motor controlled by thyratrons.
the ratio saturation/threshold is practically equal to \((2P - s_0)/s_0\). To give a specific example, with \(P = 160\) volts and \(s_0\) about 7 volts, this ratio equals about 47. It is, in fact, less than 50 for most thyratrons.

5. **Response Curve of a Thyatron.** Consider the case of a thyratron the output of which is sent into a real impedance \(R\), the tube being set off by synchronous impulses (Fig. 33-25d). If the curve of \(\int i \, dt\) is considered as a function of the phase lag \(\Delta \varphi\), a branch of a sine wave is obtained (Fig. 33-28):

\[
\int_{\Delta \varphi/\omega}^{\pi/\omega} \frac{e_0}{R} \sin \omega t \, dt = -\frac{e_0}{\omega R} (1 + \cos \Delta \varphi)
\]

where \(e_0\) is the amplitude of the plate voltage.

It can be seen that, in the neighborhood of \(\Delta \varphi = \pi/2\), the response curve is quite linear. The extremities are to be rejected because of the threshold of the thyratron. In fact, the thyratron is followed in the circuit either by a filter, introduced purposely, or by an element which plays a filtering role (for example, a motor). The properties of the filter must be considered jointly with those of the above response curve when a control by thyratrons is being designed.

### 33.7. MAGNETIC AMPLIFIER

**33.7.1. Principle.** These amplifiers, known now for several decades, are based on the principle of the **saturable reactance**. Consider the circuit shown in Fig. 33-29, in which the input voltage \(v_s\) is direct and of low power. The other winding \(B\) is traversed by the "power circuit" supplied by alternating current.

![Figure 33-29: Magnetic Amplifier Circuit](image)

The induction of the core (iron without hysteresis) is given, to a first approximation, by the curve of Fig. 33-30.

![Figure 33-30: Induction Curve](image)

Suppose that the resistance \(R_s\) be so adjusted that, for a constant voltage \(v_s(t)\), the magnetic core is saturated. The inductance is practically zero, and the value of the impedance of the winding \(B\) reduces practically to the value of its pure resistance. This resistance is generally low, and therefore the load resistance \(R_l\) is traversed by the maximum current.
When the input voltage varies, the impedance of the winding \( B \) increases as the inductance of the winding \( A \) diminishes. The current \( I_r \) flowing in the load as a function of the input ampere-turns \( NI_s \) is shown in Fig. 33-31.

![Fig. 33-31.](image)

**33.7.2. Criticism of the Preceding Circuit.** The winding \( A \) in such a circuit would be traversed by an alternating component induced by the winding \( B \). The shape of the above curve is not permissible in servo-controlled systems because there exists a *rest-position current* [if \( v_s(t) \) is zero, there exists a potential difference across the load terminals] and the system is *not polarized*. As a result, it would be desirable to seek a "response" curve like that shown in Fig. 33-32.

![Fig. 33-32.](image)

**33.7.3. Practical Circuits.** There are numerous variants which do in effect give rise to the above-indicated response curve (Fig. 33-32). We shall give only one of the possible circuits. First of all, the canceling of the a-c component in the control circuit can be accomplished through use of a reactance shown in Fig. 33-33. It will be noted that, if the windings are correctly connected, there is no a-c flux traversing the winding \( A \).

![Fig. 33-33. Saturable core.](image)

Next, a curve like that of Fig. 33-32 can be obtained as the sum of two curves like that of Fig. 33-31, as shown in Fig. 33-34.

Finally, to "polarize" the detector-amplifier, the amplifier must be made symmetric. This solution is arrived at by opposing a potential
difference corresponding to zero current (polarization) to the potential difference at the load terminals.

### 33.7.6. Examples

The schematic diagram of Fig. 33-35 represents a magnetic amplifier which is polarized, has no zero current, and is linear in an appropriate range. It will be noted that the bias is obtained from an a-c source by means of a bridge of rectifiers. The diagram also corresponds to use of the output voltage as a "d-c source."

The choice of the carrier frequency is important because the time constant is inversely proportional to the frequency, and the transformer-core dimensions necessary for a given power are considerably smaller at high frequency. The weight varies roughly as the square of the reciprocal of the frequency.

The choice of rectifiers is also very important because magnetic amplifiers generally have a pair of symmetrical elements called transductors for amplifying positive or
negative direct currents. It is absolutely necessary that the rectifiers which are used symmetrically be identical; this entails careful matching in pairs. Often the rectifiers used are of germanium. Their disadvantage is their limitation at very low or at very high temperatures; on the other hand, their volume and efficiency are remarkable (Sec. 33.8.1).

33.7.5. Conclusion. Magnetic amplifiers (static type) are very much used because they are economical when it is not absolutely necessary to have a very small time constant. These amplifiers have, in fact, a time constant of a few tens of milliseconds for low powers (from 0 to 50 watts) and a few hundred milliseconds for higher powers. They are much employed industrially for controlling d-c motor or alternator speeds.

Finally, it should be noted that, the output of a magnetic amplifier being alternating current, in certain cases the rectification of this current can be avoided—for example, where the range of power permits use of a two-phase motor (Sec. 33.8.1).

If magnetic amplifiers are compared with electronic tubes, the following may be noted in favor of the former:

a. A greater ruggedness (against shocks and vibrations) and a longer life.

b. Easier input and output circuit matching, since the impedances are easy to match and they vary over a wide range.

c. Greater mixing possibilities (addition of several input quantities).

d. Easier autoregulation (a single polarization low voltage).

e. The weight of the magnetic amplifier should be compared to that of the electronic amplifier together with its supply transformer. Comparative weights and sizes are often equivalent.

f. The gain of the magnetic amplifier is high. On the other hand, its time constant is not negligible as is that of the electron tube.

It should be noted that the utility of the magnetic amplifier increases with increase of supply frequency.

The magnetic amplifier is, compared to electronic-tube amplifiers, still rare on the market. Moreover, a magnetic amplifier for 400 cps is very different from the corresponding amplifier for 50 cps, and for the same amplification there exist as many types of amplifier as there are input and output impedance values. It may, however, be hoped that, in view of the rapidly growing use of these amplifiers, a choice among preferred types will automatically be made available on the market.

33.8. RECTIFYING, DEMODULATING, FILTERING

33.8.1. Problems and Definitions. The main cases to be considered are as follows:

1. Fixed-frequency alternating current modulated in amplitude, the variation of the positive d-c variable which contains the information being less than, or equal to, the amplitude of the alternating-cARRIER frequency (modulating rate between 0 and 100 per cent, the 100 per cent value being excluded).

This is the case of a-c amplifiers in d-c energized servo systems. A modulating system is provided before the amplifier; the output voltage of the amplifier which corresponds to a useful signal as shown in Fig. 33-36a is shown in Fig. 33-36b.
2. Fixed-frequency current modulated in amplitude, the variation of the variable containing the information being equal to that of the carrier current (Fig. 33-37a and b). In fact, everything happens from the mathematical viewpoint, as if the phase of the carrier frequency were changing simultaneously with the d-c modulating voltage.

![Fig. 33-36.](image)

![Fig. 33-37.](image)

For example, consider the case of a voltage \( V \) induced in a fixed coil \( F \) by a rotating coil \( M \), Fig. 33-37c. If \( \theta \) is the angle between the directions of the two coils, the voltage induced by \( M \) into \( F \) is proportional to \( H \cos \theta \), that is, \( V = H \cos \theta \sin \omega t \), where \( \omega \) is the angular frequency of the a-c carrier.

But, by definition, the amplitude of an alternating current is always positive. The preceding equation should then be written

\[
V = H \cos \theta \sin \omega t \quad \text{if} \quad -\frac{\pi}{2} < \theta < +\frac{\pi}{2}
\]

\[
V = |H \cos \theta| \sin (\omega t + \pi) \quad \text{if} \quad \frac{\pi}{2} < \theta < \frac{3\pi}{2}
\]

This amounts to stating that the phase undergoes a variation of \( \pi \) when the amplitude of the modulated current goes through the zero value. This condition is encountered in selsyns and selsyn-like devices such as synchro detectors, etc.
3. Alternating current of constant (or slightly variable) amplitude modulated in frequency by the useful signal (Fig. 33-38a and b). Such currents can be found at the output of detectors, where the transducer is made of an oscillating circuit whose frequency varies as a function of the physical quantity to be detected (see Sec. 29.3.5).

4. Alternating current of constant (or slightly variable) amplitude modulated in phase by the useful signal (Fig. 33-39a and b). Such currents are often found in servo systems (ensuring synchronism of rotating parts).

5. Pulsed current (current impulses of any shape, presenting or not a certain degree of periodicity). Two cases are to be considered: (a) The output impulses have an arbitrary polarity (generally they are alternately positive and negative). Such is the case for vibrating relays, and ordinary relays utilized as nonlinear amplifiers. (b) The output impulses have always the same polarity. This is the case for the output signal of thyatrons.

6. Quasi-periodic current, whose rms value is a function of the command signal. The meaning to be attached to periodicity is, in fact, an optimum correlation with the power source, the latter being in itself generally a fixed-frequency direct current. Such is the case for magnetic amplifiers (without series rectifiers in the power elements).

The devices providing for the extraction of the useful information are not the same in the various cases above. They are termed:

a. Rectifiers (or detectors) in the Cases 1, 6, and sometimes 5a. They are essentially devices with nonlinear characteristics whose "direct resistance" is negligible with respect to their "reverse resistance." Mention should be made of diodes of selenium and germanium (and more recently silicon).

b. Demodulators (or phase rectifiers) in Case 2. These devices have two inputs, one for the modulated current and one for the reference carrier. Their purpose is basically the same as that of simple rectifiers; however, their output polarity does not remain the same.
c. Discriminators, in Cases 1 and 2. These devices also have two inputs. They are very similar to the demodulators, but they incorporate circuits tuned on the reference frequency.

d. Integrating networks in Cases 5 and sometimes 6. Integration with a time constant amounts to averaging the signal over a time interval approximately equal to the time constant of the network. The polarity of the output signal depends upon the mean value of the input signal.

On the other hand, all the devices give, besides the modulation (useful signal) in a more or less deformed way (distortion), residual currents such as a d-c component, an alternative component at the carrier-current frequency, etc. These residual currents should be eliminated; this is the purpose of filtering.

Finally, it may be recalled that a motor element can itself act as a rectifier. If the alternating current from a single (constant-frequency)

![Diagram](image)

Fig. 33-40.

source is sent simultaneously into the field A of a two-phase motor (Fig. 33-40) and into the control device, the rotor of the motor will be sensitive only to the modulation introduced by the control device. We have considered the case wherein this device modulates the field current (Sec. 30.5.1); according to whether the control gain lies between 0 and 1 or between −1 and +1 (change in the direction of rotation of the motor), it may be seen that the motor, the rotation speed of which is proportional to the quantity being controlled, acts as a rectifier or a phase demodulator. Therefore, information is extracted in the motor itself without need for a rectifying element (and a filtering element).

This point should not be overlooked: If the power of the motor is not to exceed about 10 watts and if for other reasons an a-c detecting and amplifying device has been decided upon (such as selsyns, etc.), it would be a serious error to change over to direct current to drive a d-c motor. The complexity of the circuit would be considerably increased and the time constant of the whole system would be greater.

33.8.2. Rectifiers. If the modulation has an amplitude of less than half that of the amplitude of the carrier current, a (germanium) crystal

1 These are often called "detectors" in radio engineering. We shall avoid use of this term in connection with servocontrolled systems to avoid possible confusion with the term error-sensing devices.
is used for low powers; for all powers up to several kilowatts a selenium rectifier may be employed. Another choice is the diode (vacuum tube).

A bridge circuit (Fig. 33-41) makes possible the rectification of both positive and negative parts of the alternating current. [It can be seen that only two rectifiers are enough if the current to be rectified is furnished by a transformer with a middle terminal (Fig. 33-42) or by any equivalent device.]

The characteristics of "dry" rectifiers (germanium, selenium) are analogous to those shown in Fig. 33-43, which is for a silicon rectifier. There exists, in fact, a very faint reverse current of less than 1 \( \mu A \) for the silicon rectifier the characteristics of which are shown in Fig. 33-43. Diode characteristics show, on the other hand, that the reverse current is rigorously zero.

Finally, as a last point, note that all "dry" rectifiers, and, in particular, those of germanium, are very sensitive to temperature. Figure 33-44 shows the "reverse" characteristics (current in the direction opposite to the normal current) for a germanium rectifier.

33.8.3. Phase Demodulators. The modulation can vary between zero and the amplitude of the carrier current (Case 2 of Sec. 33.8.1). It
Fig. 33-45.

Fig. 33-46.

Fig. 33-47. Output is between A and B.
is, then, a question of extracting a signal proportional to the modulation, taking its sign into account (Fig. 33-37). Several devices are possible. In Fig. 33-45 the principle of one such device is shown, it being assumed there that the modulated-current source is a transformer. Depending on the phase of the modulated voltage with respect to the reference voltage, the rectifier (or diode) bridge conducts through the branch $ADEBC$ or through the branch $CDEBA$.

This device implies that the reference applied at $E$ must be effectively that which served to modulate the useful signal (without phase shift); it also implies that the carrier current must be free of harmonics.

33.8.4. Demodulators. Here it is a question of extracting the information from a frequency-modulated current. The principle is shown in Fig. 33-46; therein is found once more comparison of the modulated signal with the a-c reference voltage.

Another device which permits rectifying both the positive and negative directions of frequency-modulated alternating current requires the use of four rectifiers. Figure 33-47 shows such a device wherein appropriately polarized triodes have been used instead of diodes, in order to profit from an amplification. The output signal corresponds to the rectification of the positive and negative a-c directions; it is pulsed (Fig. 33-48) as in all the preceding devices; filters are then required.

33.8.5. Filtering. The problem consists in "smoothing" the output signals from the rectifiers or demodulators in such a way that the output signal from the smoothing element (filter) will fluctuate very little about its mean value. We find it necessary here again to specify what we mean by "fluctuate very little." It can be considered, in fact, a question of first-order average or of second-order average (quadratic mean).

The design and construction of a filter are somewhat delicate, for the fluctuations must be attenuated without disturbing the information too much. This disturbance manifests itself generally by shifting the information in time: filtering requires milliseconds.

The filtering time constant is, to a first approximation, linked to the phase-shift curve by the following relation (Secs. 7.2.5 and 8.3.2):

$$\tau = \left(\frac{d\Phi}{d\omega}\right)_{\omega=0} \quad (33-7)$$

Since, moreover, the phase $\Phi(\omega)$ and the amplitude $A(\omega)$ of a linear system having minimum phase shift are connected through Bode's equation, it can be seen that not the amplitude ratio alone (nor only the phase shift) should be taken into account. The filter to be sought should satisfy a given amplitude relation (within a certain tolerance) and have, in addition, a time constant $\tau$ which is as satisfactory as possible [see Eq. (33-7)]. By working within the limits of tolerance imposed, it is possible to obtain simultaneously an amplitude curve and a phase curve which are satisfactory.
Let us consider the following example: suppose the current to be filtered is that from a rectifying cell. The filter will be taken as that shown in Fig. 33-49 (notations on the figure). The transfer function of this filter is

\[
\frac{R}{E} = \frac{1}{LCs^2 + RCs + 1}
\]

which may also be written as

\[
\frac{S}{E} = \frac{1}{s^2/\omega_n^2 + 2s/\omega_n + 1}
\]

where \[\omega_n^2 = \frac{1}{LC}\] and \[2\tau = R \left(\frac{C}{L}\right)^{1/2}\].

The curves of amplitude vs. frequency and phase vs. frequency were shown, in non-dimensional coordinates, in Chap. 6. The filtering time constant is given approximately by

\[
\tau = \left(\frac{d\Phi}{d\omega}\right)_{\omega = 0} = \frac{2\tau}{\omega_n}
\]

(see Secs. 7.2.5 and 8.3.2). In particular, for \[\tau = 0.5\], \(\tau = 1/\omega_n\).

![Fig. 33-49. Filter.](image)

It may be remarked that, in general, better time constants are obtained with filtering circuits having self-inductances.

### 33.9. Example

#### 33.9.1. Example of an Amplifier, a Part of a Complete Servo System.

1. **Data.**

   The problem consists in automatically copying the coordinates \(D, G, S\) (distance, bearing, and site) of a pursuit radar, for the ultimate purpose of transcribing the above data in a binary code. The required precision in copy is that of the radar, i.e., approximately 1/4,000 for the coordinates \(D\) and \(G\) (\(G\) is an unlimited quantity) and 1/1,000 for the \(S\) coordinate (\(S\) varies between -10 and +90°). In fact, for upkeep and maintenance efficiency reasons, the three copying servomechanisms were built with the same design.

   For mechanical reasons, it is difficult to install a mechanical digital coder on the bearing and site shafts (the latter would require a rotating contact with as many tracks as there are digits, that is, 10 at least). For electrical reasons (multi-lead cables), the transfer of information under the form of selvyn voltages is preferable, as will be seen later when the method is outlined.

   The first step consists in obtaining the maximum rate of information that is to be transmitted. The number of levels is approximately 4,000 (or, more exactly, 4,096 = \(2^{12}\)), the maximum pursuit speed is of 0.125 rounds per second, and the maximum admissible speed (scanning) is of 0.31 rounds per second. The number of bits per second is, then, (a) 510 levels per second (9 bits/sec) in automatic pursuit, (b) 1,260 levels per second (10.3 bits/sec) in scanning rotation at constant speed. As mentioned before, the transmission of information at such a rate (between 100 and 1,000 levels per second) constitutes an interesting problem of servo system.
It should also be pointed out that the radar is a mobile piece of military equipment and thus usually provided with selsyns as means of transmission for the angles. The static precision of a selsyn is approximately 1\(^\circ\), therefore not sufficient for a direct transmission. The speed limit of a selsyn is relatively high (it depends upon the frequency of the rotor input current, but never exceeds 10 per cent of the synchronism speed). The radar is normally equipped with two selsyns for each of the bearing and site values.

For the bearing, for instance, a first selsyn \( G \) (called coarse selsyn) picks up its movement directly on the shaft (ratio 1/1). A second selsyn, called fine selsyn, picks up its movement by means of a multiplying gear train (ratio 16/1).\(^\dagger\) In these conditions, a static precision of approximately \( \frac{1}{2}^\circ \), can be expected, that is, 5,760 levels vs. 4,096 required. It should be noted that the maximum speed (scanning) of 0.31 turns per second on the bearing shaft becomes 5 turns per second on the fine selsyn, which is close to the maximum speed for a selsyn energised by a 50-cps alternating current.

In short, the use of such a double chain permits the transmission of the required information rate. This target could not have been reached with a single selsyn.

\(^\dagger\) This ratio has been selected with consideration of many factors. For the present problem (recopy of physical values) it should be noted that an odd ratio would have been preferable, as will be seen later.
2. Principle of Copying Servo Systems. The principle of servo detectors is used, as described in Chap. 29, since the selsyns are already installed on radars. The information is received through two complementary channels which must be exhausted simultaneously. The general schematic diagram is shown in Fig. 33-50 with:

a. Two selsyns, identical to the transmitting selsyns used as synchro detectors SD1 (fine) and SD2 (coarse). Their rotors are mechanically coupled through a set of gears, ratio 16/1. The amplitudes of the alternating voltages delivered on the rotors as a function of the position error $\Delta \theta$ are shown in Fig. 33-51. The direct superposition of these two voltages would lead to an indetermination for a position error $\Delta \theta = \pi$, as can be seen on Fig. 33-52 for a multiplying ratio (coarse-fine) of 2. On the other hand, the direct superposition of these two signals would give an error signal with unique value but variable polarity, therefore would lead to an intricate design. These two drawbacks were minimized through a threshold device $St$ and a peak-cutting device $Ea$.

b. Threshold device $St$ (threshold in time). Its purpose is to produce a signal as shown in Fig. 33-53 from a harmonic signal. It should be noted that the relation dealt with now is a function of time, for a given value of $\Delta \theta$.

1 Study and design by SEA (Société d'Électronique et d'Automatisme, Courbevoie, Seine, France).
Fig. 33-56. Detailed amplifier and auxiliary circuits of a servo.
c. Limiter device $E_a$ (amplitude) (transforming a harmonic signal into a signal as shown in Fig. 33-54). The superposition of these signals (point 5 of Fig. 33-50) produces a signal whose amplitude vs. time is represented by Fig. 33-55 (this figure represents the case where the multiplying ratio is 4). It can be noticed, in particular, that no indetermination remains in the vicinity of $\Delta \phi = \pi$.

d. An a-c preamplifier, whose output is directed toward two elements.

e. A demodulator provided to rectify the fraction of the signal out of the preamplifier.

f. A correcting network (integrating network).

g. An addition amplifier receiving (1) the signal from the correcting network (d-c) and (2) the signal directly from the preamplifier (a-c).

h. A magnetic amplifier.

i. A two-phase servomotor.

j. The mechanical coding device which is to be servoed.

k. A tachometric d-c generator.

l. The two selsyn (SD1) and (SD2) rotors coupled through the appropriate set of gears.

3. Realization of Copying Servo Systems. Figure 33-56 represents the industrial achievement of SEA that meets the above requirements. The input currents (polarization) of the diodes for both stages (threshold and limiter) are represented in the figure. The demodulator preceding the correcting network is of the ring type. As expressed before, the information must be extracted from a modulated signal whose amplitude is positive or negative. Therefore, a device of the phase-rectifier or demodulator type is to be used.

This example shows the fairly high complexity of actual realization of devices based on simple principles. It will be noted, and this is a general case, that several devices are incorporated between the preamplifier and the power amplifier, by which the addition of values delivered before and after is achieved. The study of the amplifier can be made only after the detector, load, and servomotor are determined.
33.9.2. Hydraulic or Pneumatic Amplifiers. It frequently happens that the power element of a hydraulic motor is composed of two stages, each of similar design. The first stage can be called a power "amplifier." We have spoken thereof in the chapter dealing with hydraulic motors. The distributor (called piston valve in steam engines) may be viewed as an amplifier, since this system element controls a flow (of oil, steam, air, etc.). It itself usually works at a very low energy level. Properly speaking, the motor (cylinder and piston) is actually linked to this distributor and cannot in fact easily be dissociated from it. In two-stage schemes (Fig. 33-57) the first stage (needle valve) is used to control the position of the piston for the distributor valve of the master cylinder.

If, to conclude, we depict the stages of electrical control, we may obtain an arrangement like that of Fig. 33-58, the complexity of which is not exaggerated.

The "control motor of the hydraulic preamplifier" may be of the "loudspeaker" type, i.e., a coil which moves linearly through an intense magnetic field (Fig. 33-59). A motor of this type has the advantage of possessing a very low time constant provided that the necessary displacements are small. Occasionally, the hydraulic preamplifier is composed of a controlled leakage.
33.10. TRANSISTORS

33.10.1. Principles. The basic materials used in the manufacture of transistors are germanium and silicon, each having a valence of 4. The incorporation into these elements of valence-5 impurities, such as arsenic, creates a composite material with free electrons called a type-n crystal (negative). In the same way, the incorporation of valence-3 impurities, such as indium, creates a composite material which can easily receive electrons. Such a crystal, called type p (positive), is said to have "holes."

1. Diodes. Consider (a) a crystal where a type-n area touches a type-p area (pn junction) and (b) a positive voltage applied to the positive area p. The electrons tend to move toward the holes, so that a low voltage creates an important current (Fig. 33-60). Inversely, if the n area has a positive polarity, it is necessary to extract electrons from the p zone, where they are "short," to deliver them to the n area, already crowded. Therefore the current is small. If not a true junction, the contact between n and p areas can be realized by a point pressed against the crystal (point-contact diodes).

Silicon diodes produce a reverse current of less magnitude than that of germanium diodes (1 instead of 100 $\mu$ at 20°C, for example), but their operating range extends into higher temperatures (150 instead of 75°C). The reverse current increases exponentially with temperature.

![Diagram of Diodes](image)

2. Transistors. The notation used is $E$ for emitter, $B$ for base, and $C$ for collector. See also Figs. 33-61 and 33-62.

Consider a set of junctions as shown in Fig. 33-63. The left, center, and right parts will be respectively called collector, base, and emitter. The emitter has a positive polarity. Experience indicates the flow of a

---

1 This section has been written in collaboration with J. Friberg of the Technical Services of the French Air Force.
very small current \( I_{\infty} \) in the collector. It is the reverse current of the \( pn \) right-hand-side junction. (Electrons cannot be extracted from the \( p \) area, where they are insufficient in number.)

If the emitter is then given a negative polarity (Fig. 33-64), however small, in respect to the base, the electrons will very easily move from the emitter toward the base (direct-current diode). They are further attracted by the positive potential of the collector, and an additional collector current is generated. The latter is not quite equal to the emitter current because of a slight loss in the base. Finally, the following equation can be written:

\[
I_e = I_{\infty} + \alpha I_e
\]

where \( I_e \) is the collector current, \( I_e \) is the emitter current, and \( I_b \) is the base current. In this expression \( \alpha \) is very close to unity; actually it is 0.95 to 0.98.

It is, therefore, conceivable that a current amplification is possible between the base current and the collector current and a voltage amplification between the base/emitter voltage and the collector/base voltage (if a resistor is introduced between the battery and the collector).

If, instead of \( npn \) type, a type-\( pnp \) transistor is used, the polarities must be exchanged (Fig. 33-65). Most of the following schematic diagrams will correspond to that type, the one most frequently used with germanium.

Symbols represented in Figs. 33-61 and 33-62 will be used, the arrow representing the base/emitter diode. Silicon transistors present the same qualities as the corresponding diodes: absence of \( I_{\infty} \) current and high-temperature performance. They exist only in \( npn \) type.

As for the diodes, point-contact transistors are also used, but not in servomechanisms, because of their poorer stability in time. The existence of multiple-electrode transistors should also be mentioned. They are not currently used, and their high-frequency performance is poor because of important internal capacitances.

33.10.2. Characteristics. 1. Curves. The condition of state of a transistor is known once four values are determined:

a. Two currents: \( I_e \) and \( I_b \) or \( I_e \). The third one is derived by use of Kirchhoff’s law, \( I_e = I_e - I_b \), following the commonly accepted sign rules.
b. Two potentials: $V_{EB}$ (emitter/base) and $V_{CB}$ or $V_{CE}$. The third one is derived by use of Ohm's law, $V_{EB} = V_{BC} - V_{CB}$. Practically, no difference can be made between $V_{CB}$ and $V_{CE}$ (about 10 volts); for $V_{EB}$ has a value of approximately 0.1 volt. Let us call $V_{E}$ the emitter potential with respect to the base and $V_{C}$ the collector potential with respect to the base (or emitter).

But the four physical values are not independent; they have two relations which can be represented by two sets of curves, customarily: (1) a set for the collector, $f(V_{e}, I_{e}, I_{b}) = 0$ (Fig. 33-66) or $g(V_{e}, I_{e}, I_{b}) = 0$ (Fig. 33-67); and (2) a set of curves for the emitter, $h(V_{E}, I_{e}, I_{b}) = 0$ (Fig. 33-68). The curves of Fig. 33-66 are to be used when the input
current of the amplifier is $I_e$, the set of Fig. 33–67 when $I_b$ is the input current. For both sets of Figs. 33–66 and 33–67, which have the same appearance as the $f(V_p, I_p, V_0)$ set of a pentode tube, the hyperbolic curve of maximum dissipation $V_cE_c = W_{\text{max}}$ can be drawn, as well as the linear load curve $V_c = E - r_eI_c$.

![Graph](image_url)

**Fig. 33–68.**

The emitter set of curves does not exist for tubes, where it is replaced by the simpler relation, grid current $I_g = 0$. Therefore, the study of transistors is more complicated.

2. **Equations.** The set of curves applying to the collector can be represented by

$$I_c = I_{ce} + \alpha I_e + \frac{V_{CB}}{r_c}$$

or, since $I_c = I_e - I_b$,

$$I_c = \frac{I_{ce}}{1 - \alpha} + \frac{\alpha}{1 - \alpha} I_b + \frac{V_{CB}}{r_c(1 - \alpha)}$$

The set of curves pertinent to the emitter can simply be represented by Ohm's law, $V_E = -(r_eI_e + r_bI_b)$. In these equations, $I_{ce}$ is the saturation collector current, as before; $\alpha$ is the current gain, as before; $r_e$, $r_b$, and $r_c$ are respectively the resistances of the emitter, base, and collector.

The above equations are generally represented by the equivalent schematic diagram, shown in Fig. 33–69, where $I_{ce}$ is neglected. The generator $G$ delivers a voltage, $-r_mI_e$, with

$$\alpha = \frac{r_m + r_b}{r_e + r_b} \approx \frac{r_m}{r_e}$$

Table 33–1 gives various values for germanium and silicon transistors.
Table 33-1.

<table>
<thead>
<tr>
<th>Temp (°C)</th>
<th>-50</th>
<th>+25</th>
<th>+100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type*</td>
<td>G</td>
<td>S</td>
<td>G</td>
</tr>
<tr>
<td>α</td>
<td>0.97</td>
<td>0.96</td>
<td>0.975</td>
</tr>
<tr>
<td>$I_{ss}$, μA</td>
<td>0.05</td>
<td>0.01</td>
<td>5</td>
</tr>
<tr>
<td>$r_s$, ohms</td>
<td>30</td>
<td>140</td>
<td>40</td>
</tr>
<tr>
<td>$r_b$, ohms</td>
<td>100</td>
<td>750</td>
<td>300</td>
</tr>
<tr>
<td>$r_e$, meghms</td>
<td>2</td>
<td>0.35</td>
<td>1</td>
</tr>
</tbody>
</table>

* G = germanium (pnp type TJN2), S = silicon (nnp type 904).

Note 1. The $I_{ss}$ current is practically zero for silicon transistors.

Note 2. It is very important to note that, because the collector resistance $r_s$ is very high in respect to $r_e$ and $r_b$, it can be assumed to have an infinite value, and the basic equations become

$$I_c = I_{ss} + \alpha I_e \quad V_e = (I_e - I_b)r_c \quad V_E = -(r_e I_e + r_b I_b)$$

3. Physical Meaning of These Equations. From the just-mentioned reduced equations, the following remarks can be made:

a. The transistor is, by nature, a current amplifier between the very small base current and the emitter or collector current, the latter two being practically equal. Numerical values with $\alpha = 0.975$ are

$$\frac{\Delta I_e}{\Delta I_b} = \frac{\alpha}{1 - \alpha} \cong 40 \quad \frac{\Delta I_e}{\Delta I_e} = 0.975 \cong 1$$

b. The transistor can be a voltage amplifier between the collector voltage $V_e$ and the emitter voltage $V_E$. This condition is due to the fact that the collector resistance is great, whereas the base and emitter resistances are small. Numerical values (Fig. 33-70) are $\alpha = 0.975$, $r_s = 40$ ohms, $r_b = 300$ ohms, $r_e = \infty$, and $r_1$ = load resistance = 10,000 ohms. From $\Delta I_b = 50$ μA, we have

$$\Delta I_c \cong \Delta I_e = \frac{1}{1 - \alpha} \Delta I_b = 40 \times 0.05 = 2 \text{ ma}$$

$$\Delta V_e = r_s \Delta I_e = 10,000 \frac{2}{1,000} = 20 \text{ volts}$$

whence,

$$\frac{\Delta V_e}{\Delta V_E} = \frac{20}{0.095} \cong 200$$

Fig. 33-70.
33.10.3. Preliminary Design of an Amplification Stage.  

1. General Remarks. Since the transistor has three connection wires, namely, one each for the emitter, base, and collector, and since moreover the input signal and the amplified output signal comprise two connections each, three combinations are possible:

a. Base in common with input and output side (Fig. 33-71). This combination is referred to as a grounded base, even though, in the case of a pnp transistor, the collector has a negative potential with respect to the base.

b. Emitter in common with input and output side (Fig. 33-72). This combination is referred to as a grounded emitter.

c. Collector in common with input and output side (Fig. 33-73). This combination is referred to as a grounded collector.

An amplification stage is designed as follows: The input current $I_i$ is obtained from the voltage $e_0$ of the input signal, the impedance of the source, and the impedance of the output network. The transistor being a current amplifier, this gives the output current $I_o$ in the load resistance $R_1$. A knowledge of the output impedance can be useful for the matching of the following stages.

The basic and important values to be known are then, current amplification, voltage amplification, power amplification, and input and output impedances. A quantity particular to transistors is the stability factor $S = dI_o/dI_{oc}$ which characterizes the variability of the output current ($I_o$, or practically, $I_e$) in relation to temperature. $S$ must be as small as possible.

The three different arrangements will now be analyzed in detail, with the derivation of two of the values involved. For these arrangements, Table 33-2 gives the approximate theoretical expression for the principal values, assuming $r_e$, $r_b$, and possibly $r_1$ small with respect to $r_e$ and the practical value for a pnp preamplification transistor, type TJN2, OC70, for which we shall take $\alpha = 0.975$, $r_e = 40$ ohms, $r_b = 300$ ohms, $r_e = 1$ megohm, $r_1 = 10,000$ ohms, and $\gamma = 1,000$ ohms.

2. Common-base Arrangement. This arrangement uses (1) a battery of voltage $E = 12$ volts to polarize the collector and (2) a battery of voltage 1.5 volts to polarize the emitter. The circuit diagram is shown in Fig. 33-74. A load resistance of 5,000 ohms and the necessary measuring instruments are inserted. A 50-ohm potentiometer is used in order to adjust the input current. It is seen that the input and output currents are practically equal, $\Delta I_e = \alpha \Delta I_i$. The output voltage is amplified (Fig. 33-79); Table 33-2 gives the effective values. It is seen that the power amplification is moderate and that the stability factor is satisfactory.

3. Common-emitter Arrangement. With the arrangement shown in Fig. 33-75, both current and voltage are amplified: $\Delta I_o/\Delta I_e = \alpha/(1 - \alpha)$ (Figs. 33-78 and 33-79).
<table>
<thead>
<tr>
<th>Type of Arrangement</th>
<th>Equation (Practical Value)</th>
<th>Current Amplifier</th>
<th>Voltage Amplifier</th>
<th>Input Impedance</th>
<th>Output Impedance</th>
<th>Stability Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common-base arrangement</td>
<td>$\alpha \frac{1}{1 + \frac{r_i}{r_e}} \cong -\alpha$</td>
<td>$1$</td>
<td>$-\alpha$</td>
<td>$200$</td>
<td>$50$</td>
<td>$800,000$</td>
</tr>
<tr>
<td>Common-emitter arrangement</td>
<td>$\alpha \frac{1}{1 - \alpha + \frac{r_i + r_e}{r_e}} \cong \alpha$</td>
<td>$40$</td>
<td>$200$</td>
<td>$50$</td>
<td>$800,000$</td>
<td>$-\alpha$</td>
</tr>
<tr>
<td>Common-collector arrangement</td>
<td>$\alpha \frac{1}{1 - \alpha + \frac{r_i + r_e}{r_e}} \cong \alpha$</td>
<td>$40$</td>
<td>$200$</td>
<td>$50$</td>
<td>$800,000$</td>
<td>$-\alpha$</td>
</tr>
</tbody>
</table>
Figure 33-78 represents the current amplification for a silicon transistor type 904. From the performance standpoint, this arrangement is the best one. There is a phase inversion between input and output \((V_e\) decreases). Table 33.2 gives the effective values. The power gain is excellent \((40\ \text{db})\); unfortunately, the stability factor is poor.

4. Common-collector Arrangement. Consider the circuit shown in Fig. 33-76. There is a current amplification between base and emitter.

\[
\frac{\Delta I_e}{\Delta I_b} = \frac{\alpha}{1 - \alpha} \approx \frac{\Delta I_e}{\Delta I_b}
\]

There is no voltage amplification, since the \(V_e\) differs from \(V_i\) only by the potential difference between emitter and base, which is very small. This arrangement is
analogous to a cathode follower. It gives small amplification but is especially useful for impedance matching. The output current reaches its maximum value when \( U_s = 12 \text{ volts} = E \), whence \( I_{\text{max}} = E/R_1 \) and \( W_{\text{max}} = E^2/R_1 \). This is true for the three arrangements. For a given supply voltage, a considerable output power is obtained, with a low load resistance. Effective values are given in Table 33-2. It is seen that the stability factor is poor and the gain rather low, so that this arrangement is suitable only for matching impedances (output stage with low impedance).

33.10.4. Alternating-current Amplification. 1. Polarizations. An arrangement with a common emitter is the best from the amplification standpoint. Unfortunately, its stability coefficient is poor. Hence the idea of amplifying the alternating current by using a common emitter

![Fig. 33-80.](image)

![Fig. 33-81.](image)

while polarizing as for the common-base arrangement, that is, giving a fixed value to the emitter current. This can be achieved in the following manner (Fig. 33-80): The base potential is maintained at its fixed value by a given divider \( R_2R_3 \). The emitter polarization is obtained by connecting the emitter to the position terminal of the battery through a rather high resistance \( R_1 \). In so doing, the stability factor approaches unity. But with alternating current, \( R_1 \) is shorted by \( C_1 \), thus giving in fact a common-emitter arrangement. There is an analogy here with the automatic polarization of vacuum tubes.

2. Association of Stages. A considerable part of the experience gained with electronic tubes can be applied to transistors. Coupling between stages can be effected through transformers if optimum impedance matching is desired, in order to minimize the number of transistors. For example, a ratio of 20/1 can be used between two stages with a common emitter. The coupling can also be done through capacitances. In this
case, it is advantageous to insert a common-collector stage between two common-emitter stages (Fig. 33-81).

The feedback can be effected by a liaison between collector and base (Fig. 33-82) or a resistor inserted in the emitter (Fig. 33-83).

**Fig. 33-82.**

**Fig. 33-83.**

**Fig. 33-84.**

Push-pull arrangements can be used as for tubes:

a. With a tapped transformer (Fig. 33-84). The voltage divider 3 ohms/220 ohms polarizes the bases. Efficiencies up to 70 per cent can be obtained.

b. With a phase-change transistor (Fig. 33-85). By insertions in series of equal resistances in the collector and the emitter, voltage variations equal in amplitude and opposite in phase can be obtained. By combining pnp- and npn-type transistors, with similar characteristics, push-pull arrangements without phase shift or transformers can be made (Fig. 33-86).

The main difficulty lies in obtaining pnp- and npn-type transistors whose characteristics are close enough to each other. It will be recalled that there is no silicon pnp-type transistor.

33.10.5. **Direct-current Amplification.** Variations in temperature or supply voltage tend to introduce a shift; that is, an output current is produced, even with no input current. On the other hand, junction-type transistors have an advantage over tubes in that they are not affected by aging.
1. Suppression of Thermal Shift. We shall first consider germanium transistors. The main problem is to minimize the influence of a variation in $I_{ce}$ which increases exponentially with temperature. One suitable method is to insert in the emitter small resistors to improve the stability factor. Series arrangements can also be made so that the different variations of $I_{ce}$ result in zero variation of the output current in the final stage. The following combinations are known to give good results: common emitter–common emitter, base-emitter, base-collector, emitter-collector, and collector-base. The influence of temperature can be compensated for by the introduction of very sensitive elements such as thermistances $r = r_0 e^{-B/T}$ or diodes, or a common-emitter arrangement with thermistance stabilization. More detailed information can be found in Shea, "Principles of Transistor Circuits," pp. 169–180. For silicon transistors, a thermistance or a diode can be inserted at a suitable point, if necessary.

2. Stabilization of Supply Voltages. Stabilization of low-voltage supply can be effected either by means of a silicon diode operating in the flash area (Fig. 33-88a and b), the latter voltage being fairly stable when tem-
perature changes, or through a nonlinear resistance (Fig. 33-89a and b). In difficult cases, transistors can be used to produce a positive voltage or current regulation. Figure 33-90 shows the regulation of the armature current in a small field-controlled d-c motor. The system is in equilibrium when the base potential is very close to the emitter potential; i.e., when the voltage drop in the \( R_2 \) resistance (30 ohms) of the emitter is equal to the 3 volts supplied by the nonlinear resistance, the latter being in turn supplied by the \( R_1 \) resistance (3,000 ohms), or

\[
I_e = \frac{3 \text{ volts}}{30 \text{ ohms}} = 100 \text{ ma}
\]

Hence, \( I_e \cong I_e = 100 \text{ ma} \).

3. Use of a Symmetrical Arrangement. The use of a symmetrical arrangement makes it possible to eliminate the shift despite considerable variations in supply voltage or temperature. Since a transformer arrangement can obviously not be used with direct current, the choice remains among (a) push-pull with phase-shifting transistors, as shown in Fig. 33-85, (b) push-pull with complementary symmetry \((pnp-npn\) as shown in Fig. 33-86), and (c) self-dephasing push-pull. This last arrangement will be examined in greater detail in a particular case (Fig. 33-91). Its advantage is that it requires only two transistors of the same type \((npn\) if possible because \(I_{ce}\) is not present). This arrangement uses two silicon transistors—type 903, 904, or 905—with a common emitter (Fig.
Typical numerical values are

\[ r = 5,000 \text{ ohms} \quad R = 180,000 \text{ ohms} \quad C = 1 \mu \text{f} \]

When no input signal is applied, the master transistor \( T' \) and the slave \( T'' \) are polarized by a current \( i = V_c/R \). Their collector currents \( I_c \) are identical, and the resistor \( R/2 \) inserted between \( E'' \) and \( C'' \) is such that its current \( I'' \) is equal to \( 2i \). Both transistors deliver equal currents \( I' = I'' = I_c - 2i \) in equal load resistances \( r \) (for instance, double-armature motor). The output signal can be characterized by \( I' - I'' \). The voltage between \( C' \) and \( C'' \) can also be utilized.

If the transistors have any characteristics, nonlinear but identical, the equality of \( I' \) and \( I'' \) will be maintained, whatever the input voltage and the temperature. When an input current \( I_e \) arrives at the base of the master transistor, \( I' \) increases, \( V_c \) decreases, the current \( i \) in the slave base decreases, consequently \( I'' \) decreases also. The behavior is then that of a push-pull. The self-dephasing arrangement is effected by the potentiometer of the master collector acting on the slave base current.

A capacitance \( C \) between \( C' \) and \( C'' \) is necessary to avoid low-frequency jamming. Self-dephasers give poor amplification of alternating currents whose frequencies exceed a few hundred cycles. The derivation of the gain formula is:

\[ I'_e = I' + 2i = \frac{\alpha}{1 - \alpha} (I_e + i) \quad I''_e \approx I'' + 2i \approx \frac{\alpha}{1 - \alpha} i \]

whence, by difference, \( I' - I'' \approx [\alpha/(1 - \alpha)]I_e \). The current gain is approximately that of a simple transistor and depends very little on the voltage, temperature, and load resistance.

The time constant due to a self-inducting load resistance \( r_e \) is not involved, since the current gain is maintained in the low-frequency range. The parallel capacitance (Fig. 33-93) of a grounded-emitter arrangement with temperature stabilization can be introduced in a continuous amplifier as a lead-factor element.

33.10.6. Conclusion. The use of transistors gives a reduction in weight and volume due not as much to the very small size of the transistors themselves as to the small size of the power-supply sources made possible by the high efficiency, low operating voltages, and absence of heating currents. Under these conditions it becomes advantageous to use very small components like tantalum capacitors, despite their high
<table>
<thead>
<tr>
<th>1. Schematic Diagram...</th>
<th>Relay</th>
<th>Vibrating relay</th>
<th>Vacuum tubes</th>
<th>Transistors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image" alt="Relay Schematic" /></td>
<td><img src="image" alt="Vibrating Relay Schematic" /></td>
<td><img src="image" alt="Vacuum Tube Schematic" /></td>
<td><img src="image" alt="Transistor Schematic" /></td>
</tr>
<tr>
<td>2. Input signal.........</td>
<td>D-c</td>
<td>D-c or modulated a-c</td>
<td>D-c or a-c</td>
<td>D-c or a-c</td>
</tr>
<tr>
<td>3. Output signal........</td>
<td>D-c</td>
<td>A-c, square, with non-zero average voltage</td>
<td>D-c or a-c according to coupling</td>
<td>Same nature</td>
</tr>
<tr>
<td>4. Principal source of energy (Q)</td>
<td>Generally d-c; low internal impedance</td>
<td>Generally d-c; low internal impedance</td>
<td>D-c high voltage (a minimum of 28 volts with special tubes)</td>
<td>D-c</td>
</tr>
<tr>
<td>5. Additional sources (E)</td>
<td>None</td>
<td>A-c at constant frequency (if desired)</td>
<td>D-c voltages for polarisation and heating</td>
<td>None grid polarisation</td>
</tr>
<tr>
<td>6. Measurement range...</td>
<td>Interrupted d-c, only two values possible</td>
<td>Limited by mechanism, 10 to 1</td>
<td>In class A: saturation — threshold = noise from 10^4 to 10^6; in other classes, distortions</td>
<td>Somewhat smaller than that of the vacuum tube</td>
</tr>
<tr>
<td>7. Output noise level...</td>
<td>Zero, except while switching</td>
<td>High; depends on the impedance of the element absorbing the output. Filtering needed</td>
<td>Very low; varies with class. Depends on input resistance</td>
<td>Very low</td>
</tr>
<tr>
<td>8. Input impedance.....</td>
<td>That of the relays: 10–1,000 ohms; fairly reactive</td>
<td>That of the deflecting coil is variable; very slightly reactive</td>
<td>Internal impedance 10^4–10^6 ohms</td>
<td>About 10^4 ohms</td>
</tr>
<tr>
<td>9. Output impedance...</td>
<td>That of the source</td>
<td>That of the source</td>
<td>Internal resistance plus that of the source</td>
<td>About 10^4 ohms</td>
</tr>
<tr>
<td>10. Steady-state gain in open loop</td>
<td>Power gain, up to 10^4–10^6</td>
<td>Max power gain is 1,000</td>
<td>Voltage gain: 10^4–10^6. Power gain: poorly defined</td>
<td>Current gain: about 100</td>
</tr>
<tr>
<td>11. Average time constant</td>
<td>Not defined; delay from a millisecond to a fraction of a second</td>
<td>At minimum, period of the vibration</td>
<td>Less than microsecond in most tubes</td>
<td>About a microsecond</td>
</tr>
</tbody>
</table>
### Amplifiers and Preamplifiers

<table>
<thead>
<tr>
<th>Thyristors</th>
<th>Magnetic</th>
<th>Ward-Leonard</th>
<th>Amplidyne</th>
<th>Hydraulie</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Thyristor Diagram" /></td>
<td><img src="image" alt="Magnetic Field Diagram" /></td>
<td><img src="image" alt="Ward-Leonard Diagram" /></td>
<td><img src="image" alt="Amplidyne Diagram" /></td>
<td><img src="image" alt="Hydraulic System Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Generally</th>
<th>D-c*</th>
<th>D-c</th>
<th>D-c</th>
<th>Position (or velocity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modulated pulsed (polarised)</td>
<td>A-c</td>
<td>D-c</td>
<td>D-c</td>
<td>Position (or velocity)—perhaps force</td>
</tr>
<tr>
<td>A-c 50-500 volts</td>
<td>A-c; low internal impedance</td>
<td>Mechanical, with quasi-constant speed</td>
<td>Mechanical—with quasi-constant speed</td>
<td>High-pressure fluid at constant pressure</td>
</tr>
<tr>
<td>Heating; grid polarisation</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Average† 10 to 1</td>
<td>Good: 50 to 1</td>
<td>Excellent: 300 to 1</td>
<td>Excellent: 500 to 1</td>
<td>Excellent: 500 to 1</td>
</tr>
<tr>
<td>See line 3. Filtering needed</td>
<td>Harmonic components; may be reduced</td>
<td>Low; high frequencies due to brushes</td>
<td>Slightly superior to the Ward-Leonard system a single stage</td>
<td>Very low if the source Q is quite stable</td>
</tr>
<tr>
<td>Grid impedance: 10⁴-10⁴ ohms—pure resistance; zero resistance when the tube conducts</td>
<td>Extremely variable, from a few ohms to 100 ohms. Reactive</td>
<td>100-1,000 ohms, slightly reactive (10%)</td>
<td>100-1,000 ohms, slightly reactive (10%)</td>
<td>Inertia of the movable part of the distributing valve</td>
</tr>
<tr>
<td>Internal impedance of the source</td>
<td>Extremely variable, great adaptive flexibility; from a few ohms to 10⁴ ohms</td>
<td>That of the armature: 10-500 ohms</td>
<td>That of the armature: 10-500 ohms</td>
<td>Depends on the head losses of the distribution</td>
</tr>
<tr>
<td>Voltage gain poorly defined. Power gain: up to 100-500</td>
<td>Poor gain up to 10⁴-10⁴</td>
<td>Dynamic amplification coefficient: 10-300</td>
<td>Dynamic amplification coefficient: 100-10,000</td>
<td>Power gain—up to 10⁴-10⁴</td>
</tr>
<tr>
<td>At the minimum period of the source Q for one thyristor; reduced if several thyristors on the same source Q</td>
<td>At the minimum, period of the source Q, plus time constant of the input circuit</td>
<td>Minimum 10 msec up to 100 watts. 10-50 msec for 1,000 watts</td>
<td>6 msec up to 100 watts; 5-50 msec up to 5,000 watts</td>
<td>May remain equal to 1 msec up to several kilowatts</td>
</tr>
</tbody>
</table>
### Table 33-3

<table>
<thead>
<tr>
<th>12. Typical transfer function</th>
<th>Relay</th>
<th>Vibrating relay</th>
<th>Vacuum tubes</th>
<th>Transistors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximation without filter:</td>
<td>$E = \frac{e^{-t}}{1 + ts}$</td>
<td>Linearisation</td>
<td>$\frac{E}{E} = k$</td>
<td>$\frac{E}{E} = k$</td>
</tr>
<tr>
<td>$T = \text{delay} = L/R$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>13. Output power range for different types</th>
<th>1-10^4 watts</th>
<th>10^-2-10^2 watts</th>
<th>10^-7-10 watts</th>
<th>Power, 3-10 watts according to class</th>
</tr>
</thead>
</table>

| 14. Recommended power | Any power if servo is not very accurate | 10-50 watts | 1 watt if followed by a 2d stage; 10 watts if direct | 2 to 20 watts according to class |

<table>
<thead>
<tr>
<th>15. Efficiency from energy standpoint</th>
<th>Very high</th>
<th>Average</th>
<th>Depends on class: 30-80%</th>
<th>Depends on class—comparable to the vacuum tube</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>16. Ruggedness</th>
<th>Good</th>
<th>Average</th>
<th>Good; excellent with special tubes</th>
<th>Excellent for junction transistor</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>17. Maintenance</th>
<th>Very Little</th>
<th>Low</th>
<th>Low; periodic checks necessary</th>
<th>Zero</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>18. Length of life</th>
<th>That of contacts: 10^7 to 10^9 complete cycles; depends on load impedance</th>
<th>That of the contacts: 10^7 to 10^9 complete cycles; depends strongly on load impedance</th>
<th>That of the tubes. Rarely under 1,000 hr</th>
<th>Over 50,000 hr</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>19. Relative cost</th>
<th>1</th>
<th>5</th>
<th>5</th>
<th>4</th>
</tr>
</thead>
</table>

| 20. Remarks | Rarely used directly; often associated with thyristors or Ward-Leonard system or magnetic amplifier | Will have growing importance (in replacing vacuum tubes) | |

* Generally d-c; however, thyristors are often driven by an a-c voltage coming from the same source as the plate supply but shifted in phase with respect to this voltage by a quantity proportional to the signal. 

† Thyatron. The response curve of a thyatron is not linear (area exists under a sine wave). In the neighborhood of an inflection point, the curve can be considered linear (Fig. 33-28).

---

The transistor is sturdy and stands up well to shock and vibration. Its life span can presently be estimated at 100,000 hr. Transistors do not represent a real revolution in the field of low-frequency a-c amplifiers. For d-c amplifiers, special care should be exercised to prevent thermal shifts. But, in so far as this condition is met, units very suitable for servo systems can be obtained, especially in the case where the supply source is a low-voltage direct current.

### 33.11. COMPARATIVE TABLE OF THE PRINCIPAL TYPES OF AMPLIFIER

Table 33-3 is intended mainly to present comparative values within the confines of a single horizontal line. The input of the amplifier has been
<table>
<thead>
<tr>
<th>Thyatrons</th>
<th>Magnetic</th>
<th>Ward-Leonard</th>
<th>Amplidyne</th>
<th>Hydraulic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximately (before filtering) $\frac{R}{E} = \frac{k}{1 + vs}$ (see line 20)</td>
<td>Approximately $\frac{R}{E} = \frac{k}{1 + s}$</td>
<td>Without compensating winding $\frac{R}{E} = \frac{k}{s(1 + rs) (1 + rs)}$</td>
<td>Without compensating winding $\frac{R}{E} = \frac{k}{p(1 + rs) (1 + rs)}$</td>
<td>Linearization about a steady flow output $\frac{R}{E} = \frac{k}{s^2 + 2\omega_0 s + \omega_0^2}$</td>
</tr>
<tr>
<td>10-100 watts</td>
<td>10-100 watts</td>
<td>10-100 watts</td>
<td>10-100 watts</td>
<td>1-100 watts</td>
</tr>
<tr>
<td>100-500 watts</td>
<td>5-10 watts</td>
<td>10 watts</td>
<td>500-5,000 watts</td>
<td>All powers</td>
</tr>
<tr>
<td>Excellent; above 90%</td>
<td>Good</td>
<td>Very high</td>
<td>Very high</td>
<td>Excellent or average according to distributor type</td>
</tr>
<tr>
<td>Poor</td>
<td>Excellent</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Frequent checks necessary</td>
<td>Zero</td>
<td>Average</td>
<td>Average</td>
<td>Average</td>
</tr>
<tr>
<td>That of the thyatrons: several hundred hours</td>
<td>Indefinitely long (except in case of overload)</td>
<td>If ball bearings are used, life limited by commutator</td>
<td>Like that of Ward-Leonard system</td>
<td>Several tens of thousands of hours</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

2 Dynamic amplification coefficient = (steady-state power gain)/(time constant).

§ It is desirable to install a real load impedance; if not, a device (correcting network) must be connected to the relay, for the purpose of suppressing the sparks due to interruption of the current.

designated by $E$, the output by $S$, the principal energy source by $Q$, and the auxiliary energy source by $q$. The indicated properties correspond mainly to the aircraft field, i.e., mainly to amplifiers of servo-controlled systems the output power of which lies between 10 and 500 watts and for which the "average time constant" lies between 5 and 50 msec.
PART FIVE

BASIC DESIGN OF A SERVO SYSTEM

CHAPTER 34

TWO EXAMPLES OF BASIC SYSTEM DESIGN

Summary

1. Some leading principles.
2. First example: servomechanism for a measuring system.

34.1. SOME MAJOR PRINCIPLES

34.1.1. Introduction. No absolutely general method is available for designing a servo system. Owing to the existence of a great variety of problems and techniques, the part played by experience is just as important when dealing with a specific problem as the use of general methods. However, provided that the engineer in charge of a servo project has already had some experience with the techniques involved, and provided that he starts by thinking over the problem that he has to solve, the methods outlined in this book provide him with a very powerful tool which can be considered to be one that is of general purpose. Rather than summarizing once more that which has been previously stated, it is considered preferable to emphasize a few points which engineers often overlook and which are the cause of frequent failures in the automatic-control field.

34.1.2. Some General Considerations and Important Pitfalls. 1. For the design of a feedback control system satisfying certain specifications, many solutions are always possible. These solutions generally are technically very different, the reason being that there are always more unknown than known parameters.

2. In order to choose among the many possible solutions, a general method consists in extrapolating, whenever possible, solutions that have been previously realized and proved. In the technical field entirely new solutions are the exception and are encountered only in advanced research problems.

3. Once the procedure that is to be used for solving the problem is finalized, the designer must translate the sponsor’s specifications into the nomenclature of servomechanisms. This step is often most difficult; for the design engineer must very frequently press the sponsor for lucid and satisfactory specifications which can be expressed in terms of quantities which define the operational characteristics of servomechanisms.

4. As far as the design itself is concerned, the principle which experience shows to be most often overlooked is that a feedback control system

1 A third example (Basic Design of a Jet Regulator) is dealt with in vol. 2 of the German edition (Oldenbourg) or the second French edition (Dunod) of this book.
constitutes an entity, and each component of it must be considered as a part of the over-all system. Experience has also shown that the majority of technical failures in the servomechanism field are due to the neglect of this principle by the design engineer. This might well be called the "golden rule" of the feedback-control-system engineer. In this respect, the most common errors are:

a. Choosing a component without making a quantitative study of the over-all influence of its characteristics

b. Using tolerances for some components without obtaining them from the over-all specifications

c. Running tests in the laboratory or in the field without a preliminary quantitative study of the performance to be tested, etc.

Most of these errors could be avoided if one would consider as a fundamental principle the fact that a complete quantitative evaluation of the over-all system should precede any decision concerning the selection of a component.

5. The validity of the results obtained as a conclusion to any step in a design procedure should always be checked. In particular, it should be remembered that the transfer functions are valid only in a given frequency band and that experiments should be carried out for higher frequencies than those which the system might encounter during normal operating conditions.

34.1.3. Remarks Concerning the Following Examples. It is not possible, within the scope of this book, to present complete design examples of a feedback control system. The choice of one from a number of possible solutions is a matter of engineering experience which cannot be condensed into a few pages. Furthermore, the final phases of a design procedure, especially the laboratory tests, depend on specialized techniques. Therefore, the two examples given below are restricted to the investigation of the basic design once the principle of the solution has been chosen.

34.2. FIRST EXAMPLE: POSITIONAL SERVO FOR A MEASURING SYSTEM

34.2.1. Data. 1. The problem. The bearing of an airplane with respect to a runway is measured by the two angles $\theta_1$ and $\theta_2$ from two symmetrically located stations (Fig. 34-1).¹ At each station the angle is measured as follows: A uhf emission from the airplane is received by antennas connected to the extremities of a coaxial line (Fig. 34-2). Standing waves are produced in the line, the position of a minimum in voltage being very nearly proportional to the bearing of the airplane. Automatic measurement of $\theta$ is obtained by slaving: for example, by using a slide which follows the minimum in voltage. When the slaving is perfect,

¹ Such a system (telémètre bistatique) was studied by the Centre National d'Études des Télécommunications, Paris. A somewhat different approach is given by G. Klein, "Étude et Réalisation d'un Asservissement de Position," Annales des Télécommunications, 6(11):313–324 (1951).
a measure of $\theta$ is given by the displacement of the slide. The purpose of the present section is to study the servo system that causes the slide to follow the minimum in voltage.

The sensing device of this servo system is made of two probes attached to the slide. If the voltages $e_1$ and $e_2$ given by the two probes are equal, then the midpoint of the probes, which is also attached to the slide, is located on a voltage minimum (Fig. 34-3).

Thus, a discriminator that measures $\varepsilon = e_1 - e_2$ can be used as the sensing device for the servo system under consideration. As a first approximation, the measured quantity $e_1 - e_2$ is proportional to the positional error of the servo system $x_n - x_c$, where $x_n$ is the position of the voltage minimum and $x_c$ that of the slide. Figure 34-4 shows a block diagram of the system used to measure $\theta$.

2. Stipulated Specifications. Control Input. It is the quantity $x_n$ which is to be measured. The most probable signal is due to an airplane on the runway moving with a velocity $V$. Such an input is continuous (Fig. 34-5), which suggests that the input should be characterized by a maximum angular velocity $(d\theta/dt)_M$ and acceleration $(d^2\theta/dt^2)_M$. For this case, the frequency spectrum of the input is of no interest. Numerically:

$$
\left(\frac{dx}{dt}\right)_M = 12 \text{ cm/sec} \quad \left(\frac{d^2x}{dt^2}\right)_M = 2.4 \text{ cm/sec}^2
$$

Noise. In the absence of quantitative data concerning noise, it is required that the resonance ratio should not exceed 1.5.
ACCUacy. The required accuracy is defined for three types of input: (a) constant input, steady error less than 0.4 mm; (b) ramp input with maximum velocity, velocity error less than 1.5 mm, and (c) harmonic input with an amplitude of 15 mm at an angular frequency of 1 rad/sec (in order to follow the movements of the airplane around the axis of the runway), error less than 1.5 mm.

OTHER SPECIFICATIONS. The mechanical portion of the system is given and is assumed to be linear. The motor drives the slide by means of a lead screw. The pitch is 4 mm and the backlash of the slide displacement is 0.1 mm. The inertia of the mechanical portion of the system, as seen from the motor output shaft, is $0.5 \times 10^4$ g-cm$^2$.

The motor is to be thyatron-controlled. It is assumed that the sensing and amplifying circuits, from the sensing device to the motor-control voltage, introduce a time lag of 0.05 sec. The motor is to be an armature-controlled d-c motor with separate field excitation. It will be assumed that the position of the output shaft does not influence the control voltage.

34.2.2. Investigation of the Basic Design. 1. General Outline. The basic design will be carried out as follows: The specifications impose certain conditions on the transfer function, especially on the corresponding amplitude and phase response. Evaluation of the tolerable lags then enables a choice to be made of the type of motor desirable, which yields a second approximation for the transfer function. A study of the stability will then show the necessity for incorporating into the system a compensating network. Designing the compensating network defines the main characteristics of the system.

2. First Approximation for the Open-loop Transfer Function. a. SENSING AND AMPLIFYING STAGES. The error signal ($E$ in volts) of the servo system and the control voltage ($\theta$ in volts) of the motor are related by the transfer function

$$\frac{E}{\theta} = \frac{K_1}{1 + 0.05s}$$

where the coefficient $K_1$ includes the adjustable gain of the amplifier.

b. MOTOR. The type of the motor, as well as the gear ratio

$$\frac{\theta_m}{\theta} = \frac{\pi}{0.2} = 15.7$$

are given. The form of the transfer function of the motor is obtained from:

1. The moment equation applied to the output shaft:

$$J \frac{d^2\theta_m}{dt^2} + C_s = C_d = K_d I$$
2. Kirchhoff's law applied to the armature circuit:

\[ E = RI + L \frac{dI}{dt} + K_v \frac{d\theta_m}{dt} \]

where \( J \) = total inertia referred to output shaft
\( C_o \) = opposing torque referred to output shaft
\( C_d \) = driving torque referred to output shaft
\( K_v \) = constant for the motor back emf
\( E \) = armature voltage
\( R, L, I \) = resistance, inductance, and intensity in the armature circuit.

These equations enable the motor to be represented by the block diagram of Fig. 34-6.

![Diagram of Fig. 34-6.]

\( \frac{E}{\text{Volts}} \rightarrow \frac{K_i}{R} \frac{1}{1 + \frac{L}{R} s} \rightarrow \text{Driving torque} \rightarrow \frac{R}{K_i K_\omega} \frac{1}{s} \frac{1 + \frac{L}{R} s}{1 + \frac{JR}{K_i K_\omega} s + \frac{JL}{K_i K_\omega} s^2} \rightarrow \theta_m \rightarrow \text{Radians} \)

![Diagram of Fig. 34-7.]

\( x \rightarrow \text{x of voltage minimum, cm} \rightarrow \text{Error} \rightarrow \frac{K_i}{1 + 0.05 s} \rightarrow E \rightarrow \text{Volts} \rightarrow \frac{0.063}{K_\omega} \frac{1}{(1 + \frac{JR}{K_i K_\omega} s + \frac{JL}{K_i K_\omega} s^2)s} \rightarrow x \rightarrow \text{x of slide, cm} \)

\( c. \) If the opposing torque \( C_o \) is neglected, the block diagram is simplified. Since

\[ \frac{x_s}{\theta_m} = \frac{0.2}{\pi} = 0.063 \]

where \( x_s \) is the abscissa of the slide in centimeters, the block diagram becomes that shown in Fig. 34-7. It is a servo system with unity feedback. A first approximation for its open-loop transfer function when the armature inductance is neglected is

\[ \frac{K_v}{s(1 + 0.05 s)(1 + T_m s)} \]

where

\[ K_v = 0.063 \frac{K_i}{K_\omega} \quad T_m \approx \frac{JR}{K_i K_\omega} \]

The next step is to find values for the gain \( K_v \) and the lag \( T_m \) that are compatible with the specifications.

\( d. \) Determining the gain. The servo system has integration and therefore (Chap. 15) has no position error. There is, however, a velocity error \( a/K_v \) for a ramp input \( x(t) = atu(t) \).
The specification that the error should not exceed 0.15 cm for \( a = 12 \text{ cm/sec} \) requires that \( K_e \geq 80 \text{ sec}^{-1} \).

(e. note.) Another minimum value for the gain results from the required accuracy specification under static conditions. In fact, when an opposing torque is present, positional error may occur even when integration is present in the system (Sec. 15.3.3, par. 3).

It can be seen from Fig. 34-6 that, neglecting armature inductance, the error-disturbance transfer function is

\[
H(s) = \frac{\varepsilon(s)}{C_s(s)} = \frac{0.063R/K_iK_\omega s(1 + T_\omega s)}{1 + K_e s(1 + 0.05s)(1 + T_\omega s)}
\]

\[
H(s) = \frac{R}{K_iK_1} \frac{1}{1 + T_\omega s} \frac{1}{1 + s(1 + 0.05s)(1 + T_\omega s)/K_e}
\]

If a step disturbance is applied \( C_e = \gamma_e(t) \) the steady-state error will be

\[
\varepsilon(\infty) = \lim_{s \to 0} sC_s(s)H(s) = \lim_{s \to 0} \gamma H(s) = \frac{R}{K_iK_1} \frac{\gamma}{\varepsilon}
\]

The steady-state error can also be obtained by equating the torque developed in the presence of a constant error, \( K_iI = K_e s/R \), and the opposing torque \( \gamma \), which gives \( \varepsilon(\infty) = (R/K_iK_1)\gamma \), where \( \varepsilon \) must be less than 0.04 cm. This results in a minimum value for the gain. It will be assumed that this condition is numerically less severe than that found in (d).

(f. determining the time lag.) Whereas the specifications pertinent to the gain were imperative, the tolerable lag can only be estimated. In fact, this depends on the form of the transfer functions that are involved and on the corresponding frequencies.

It is necessary to satisfy a requirement for static accuracy in the sinusoidal regime. As was seen in Sec. 8.3.8, the nondimensionalized error is related to the closed-loop phase shift (in radians) of the servo system. This demands that the phase shift does not exceed 1.5/15 rad = 0.1 rad at 1 rad/sec.

A first approximation for the bandwidth of the system can be obtained by considering the system to be similar to a second-order system with a damping ratio of 0.50. This is equivalent to a resonance ratio of 1.20, which is compatible with the requirements. Since the phase shift of this system is \( 5^\circ = 0.09 \text{ rad} \) at \( \omega = \omega_n/10 \), it is possible to consider 10 rad/sec to be a minimum value for the resonant frequency of the system. Hence, assuming that this resonant frequency corresponds to \( 1/T_m \), \( T_m \leq 0.1 \text{ sec} \).

3. Choice of the Motor, Second Approximation for the Transfer Function. The choice of the motor will not be discussed here, although the choice would take place at this point in the sequence of the project. The following transfer function is obtained between the input voltage \( E \) of the motor and the angular position of its output shaft

\[
\frac{\theta}{E}(s) = \frac{1.4}{s(1 + 0.05s)^2} \text{ rad/volt/sec}
\]

\( ^1 \) The problem of choosing the servomotor will be treated in the second example of the present chapter (Sec. 34.3.2, par. 3).
This gives a second approximation for the transfer function (Fig. 34-8) with an open-loop transfer function

$$KG(s) = \frac{80}{s} \frac{1}{(1 + 0.05s)^4}$$  \hspace{1cm} (34-1)

The stability and performance of this system are now to be checked.

4. Stability. Inspection of the Nyquist locus $KG(j\omega)$ (Fig. 34-9) shows that the system becomes unstable when $K$ is greater than some critical value. The system is regular, as defined in Secs. 8.3.9 and 9.3.3. The left-hand criterion applies. The amplitude and phase-vs.-frequency curves have been plotted in logarithmic coordinates (Bode plots) with $K = 1$ in Figs. 34-10 and 34-11. The gain and phase margins are as determined from these curves.

$$G_m = -13 \text{ db} \quad \Phi_m = -60^\circ$$

Both margins are negative and the system is therefore very unstable.

Stabilization might be achieved by closing a loop around the motor in order to suppress the integration and diminish the phase lag by approximately 90°. In the following, the usual compensation techniques will be applied.

5. Compensation. Introduction of tachometric feedback around the motor leads to the diagram shown in Fig. 34-12. The over-all open-loop transfer function is, setting $1.43k_3k_2 = \lambda$,

$$KG(s) = \frac{0.2}{\pi} \frac{k}{k_3} \frac{1}{1 + \frac{1}{\lambda} s} \frac{1}{1 + 0.05s} \frac{1}{1 + \frac{0.1}{1 + \lambda} s + \frac{0.0025}{1 + \lambda} s^2}$$

For the first approximation, it can be assumed that $\lambda \gg 1$ (in fact, the final value for $\lambda$ will be approximately 10). Hence,

$$KG(s) = \frac{0.2}{\pi} \frac{k_1}{k_2} \frac{1}{s} \frac{1}{1 + 0.05s} \frac{1}{1 + (0.1/\lambda)s + (0.0025/\lambda)s^2}$$

$$\approx K \frac{1}{s} \frac{1}{1 + 0.05s}$$  \hspace{1cm} (34-2)
Fig. 34-10.

Bode plot (amplitude) for

\[ G(s) = \frac{1}{s(1 + 0.05s)^3} \]

Gain margin

Fig. 34-11.

Bode plot (phase) for

\[ G(s) = \frac{1}{s(1 + 0.05s)^3} \]

Phase margin

Angular frequency in rad/sec

(logarithmic)
BODGE PLOT (AMPLITUDE) FOR
\[ G(s) = \frac{1}{s(1 + 0.05s)} \]

FIG. 34-13.

BODGE PLOT (PHASE) FOR
\[ G(s) = \frac{1}{s(1 + 0.05s)} \]

FIG. 34-14.
where the velocity gain is

$$K_v = \frac{0.2}{\pi} \frac{k_1}{k_2} = 80 \text{ sec}^{-1}$$

The Bode plots of this transfer function (Figs. 34-13 and 34-14) show that the system is stable and that it has a $30^\circ$ phase margin. This is insufficient, however, for this phase margin cannot compensate for the phase lag due to the factor $1/(1 + 0.1s/\lambda + 0.0025s^2/\lambda)$ which has been neglected. Therefore, it is necessary to increase the amount of compensation. This can be accomplished by introducing a differentiating network (Fig. 34-15), placed in cascade with the tachometric generator. The transfer function of this network is $RCS/(1 + RCS)$. The block diagram of the inner loop is shown in Fig. 34-16.

The closed-loop transfer function is therefore

$$\frac{\Theta(s)}{E(s)} = \frac{1.43k_1}{s^2} \frac{1 + Ts}{1 + (\lambda + T + 0.1)s + 0.1(T + 0.0025)s^2 + 0.0025Ts^2}$$

(34-3)

where $T < 0.1 \ll \lambda$. The over-all open-loop transfer function is then

$$KG(s) = 0.091 \frac{k_1k_2}{s} \frac{1}{1 + 0.05s} \frac{1 + Ts}{1 + \lambda s}$$

(34-4)

It is seen that a differentiating network in the feedback path is equivalent to an integrating network in the forward path. The magnitude of $KG(j\omega)$ approaches $80T/\lambda\omega^2$ at high frequencies.

Let a phase margin of $40^\circ$ be prescribed in the neighborhood of the resonant frequency, which was found to be approximately $10$ rad/sec.

Two approximations are now to be carried out; they should not change the phase by more than $10^\circ$ at $\omega = 10$ rad/sec if they are to be valid. The first consists in replacing Eq. (34-3) by Eq. (34-4) in the expression for the transfer function. The second consists in replacing $(1 + Ts)/(1 + \lambda s)$ by $T/\lambda$ in the neighborhood of resonance. Hence, the simplified
expression for the transfer function that is valid near resonance is

\[ K_1G_1(s) = \frac{T}{\lambda} \frac{80}{s} \frac{1}{1 + 0.05s} \frac{1 + 0.6s}{1 + 9s} \frac{1}{1 + 0.5s/s0 + s^2/s0^2} \]

The corresponding magnitude and phase curves have already been drawn (Figs. 34-13 and 34-14). Demanding for \( K_1G_1(s) \) a phase margin of \( 40 + 2 \times 10 = 60^\circ \) leads to a choice of \( T/\lambda \approx 14 \). In spite of this change, the resonant frequency still remains approximately 10 rad/sec.

The approximations made have now to be checked. The phase lag due to \((1 + Ts)/(1 + \lambda s)\) must be less than \(10^\circ\) at 10 rad/sec, whence \( T \geq 0.6 \) sec. If \( T \) is taken as 0.6 sec, it is possible to choose \( \lambda = 14 \) and \( T = 8.4 \) and to check that the phase difference between Eqs. (34-3) and (34-4) is less than \(10^\circ\) at \( s = 10/j \) rad/sec.
The setting of the amplifiers is given by

\[ k_1 k_2 \approx \frac{80}{0.091} = 880 \]

The open-loop transfer function of the system is

\[ \frac{80}{s \left( 1 + 0.05s \right) \left( 1 + 9s \right) \left( 1 + 0.5s/80 + s^2/80^2 \right)} \left( 1 + 0.6s \right) \]

and can be approximated by

\[ KG(s) = \frac{80}{s \left( 1 + 0.05s \right) \left( 1 + 9s \right) \left( 1 + 0.5s/80 + s^2/80^2 \right)} \]
which represents the open-loop transfer function of a system that meets the specifications.

The Nichols chart (Fig. 34-17) enables the transfer function of the closed-loop system to be represented by the corresponding frequency-response curves (Fig. 34-18). From this it is verified that the phase shift at 1 rad/sec is less than 5°. It will be noted that the magnitude drops below 1 near \( \omega = 20 \text{ rad/sec} \) (gain crossover frequency) but that the phase becomes \(-90^\circ\) at \( \omega = 10 \text{ rad/sec} \).

The frequency range over which the system operates is approximately 0 to 1 rad/sec. However, it has been necessary to consider frequencies that are 10 to 20 times higher; for if the gain were increased (the gain margin is 5 db), hunting would occur at \( \omega = 20 \text{ rad/sec} \).

34.3. SECOND EXAMPLE: AUTOPilot FOR A GUIDED MISSILE

34.3.1. Data. 1. Purpose of the Problem. It is desired to stabilize a missile along its vertical trajectory. The missile is roll-stabilized and has a cruciform wing configuration. As a first approximation, it is assumed that the problem of stabilization may be resolved into two planes. For the time being, only stabilization in a vertical plane will be considered, this stability being obtained by an autopilot which operates the control surfaces.

2. Aerodynamic Data. The application of the law of dynamics to the missile involves aerodynamic, gravitational, and propulsive forces. The transfer function relating the control-surface deflection to the pitch angle of the missile relative to the vertical is

\[
\frac{\Theta}{\Delta} = \frac{335s + 237}{20.7s^2 + 39s^2 + 257s - 9.5}
\]

It is assumed that an aerodynamic torque proportional to \( \delta \) and equal to \( 4 \times 10^4 \) (cgs units) per radian is exerted on the control surface. The deflection of the control surface is limited by a stop to \( \frac{1}{10} \) radian. The moment of inertia of the control surface is \( 10^4 \) in cgs units.

Note. The transfer function of a missile varies during flight, and consequently the operation given above is valid only for one speed and altitude. In the following analysis only flight conditions numerically corresponding to the transfer function given above will be considered.

The general form of the block diagram of the stabilized missile is given in Fig. 34-19, the transfer-function of which has been given above. The elements still to be determined in the synthesis of the autopilot are:

a. The vertical reference sensing device
b. The servomotor
c. The amplifiers and the compensating networks

3. Specifications. It is desired to guide the missile by controlling the velocity of the control surface (see Sec. 15.3). This requires that
a constant-error signal into the servomotor should theoretically result in an angular velocity of the control surface. In other words, the transfer function of the servomotor in the block diagram of Fig. 34-19 must be, as a first approximation, of the form \( M(s) = K/s(1 + rs) \). This type of operation helps to reduce the effect of constant disturbances (Chap. 15). A 27-volt d-c motor with separate excitation will be used.

![Fig. 34-19.](image)

The sensing device must detect the deviations of the trajectory from the vertical. A rate gyro which measures \( d\theta/dt \) (see Secs. 29.1.2 and 29.2.5) is used for that purpose. It can be adjusted as follows: (1) natural frequency between 5 and 15 cps, (2) damping range between 0.3 and 2.0, operating range between 6 and 60°/sec, and (3) saturation-to-threshold ratio 100.

34.3.2. Preliminary Design. 1. General. The preliminary design consists in choosing the detector, the servomotor, the compensating network, and the amplifier. The input to be considered is the disturbing torque, the effect of which must be minimized. This requires a system with an open-loop gain that is as high as possible and, if also possible, a high natural frequency.

The function of the amplifier will be to provide the necessary gain and to allow adjustment of it. The compensating network will enable a satisfactory margin of stability to be maintained. The gain adjustment must not be too critical. This requires a satisfactory gain margin (see note in Sec. 34.3.1, par. 2).

2. General Stability, Approximate Performance. If as a first approximation it is assumed that

\[
R(s) = \text{const} \quad M(s) = K/s \quad D(s) = D_s
\]

then the open-loop transfer locus of the system is similar to that of the missile, within a gain factor.

The Nyquist locus has the form shown in Fig. 34-20. The open loop has two unstable modes; the first is one that is slowly divergent, the
second is one that is underdamped and oscillatory. These modes are defined by the roots of

$$20.7s^4 + 30s^2 + 257s - 9.5 = 0$$

Application of the Nyquist criterion shows that the curve must enclose the $-1$ point. This is an example of a servo system with conditional stability. If the gain is increased, the system becomes stable and further increase of the gain results in hunting.

For each value of the resonance ratio, two values for the gain which are compatible with stability may be found. Compromise, however, between a high gain and satisfactory stability leads to the choice of the greater of the two values.

It can be seen in the above equation that 9.5 is negligible as compared to the other terms when $\omega > 2$ rad/sec. Moreover, it is verified from the Nichols locus that it is permissible to neglect the term for the nonoscillatory unstable mode in the region of utilization. The transfer function becomes, therefore,

$$\frac{\Theta}{\Delta} \approx \frac{1}{s} \frac{1 + 1.4s}{1 + 0.15s + 0.08s^2}$$

It can be seen from the locus (Fig. 34-22) that the open-loop and closed-loop resonance frequencies are in the neighborhood of 3 and 10 rad/sec, respectively.

The gain that provides a resonance ratio of 1.3 is $-5$ db. It can be increased by incorporating phase lead into the transfer function. Figure 34-21 shows the gain corresponding to $Q = 1.3$ as a function of the additional phase lead. It shows the necessity for the phase-lead compensating network to have a high value of the coefficient $\alpha$. The value $\alpha = 10$ will be chosen. Because of the important decrease in gain due to high values of $\alpha$, it will not be possible to take advantage of the maximum phase lead at the resonant frequency of the servomechanism. For $\alpha = 10$ it is reasonable to expect a phase lead of $35^\circ$ at that frequency (10 rad/sec). The lags which have been neglected in the above brief analyses must now be distributed between the components in such a manner that the phase lead be as effective as possible.
To decrease the phase lag due to the sensing rate gyro, it is necessary to increase the gyro's natural frequency $\omega_n$ and to decrease its damping $\zeta$. The extent to which this may be carried out is limited, however, by noise. A damping ratio of 0.5 and a natural frequency approximately 5 times that of the servo system are, however, acceptable values. For this condition the rate gyro causes a $10^\circ$ phase lag at 10 rad/sec.

![Diagram](image)

**Fig. 34-22.**

As far as the amplifier is concerned, the use of electronic amplifiers enables a time constant of the order of 10 msec to be obtained. This gives a phase lag of $5^\circ$ at 10 rad/sec.

Allowing an additional $10^\circ$ phase lag to account for the undesirable characteristics of the system, such as backlash and finite stiffness, it can be seen that the effect of the above phase lead will be nullified if the motor exhibits a phase lag of $10^\circ$. This occurs if the time constant
is approximately 20 msec. An attempt will be made to meet this specification.

3. Choosing the Motor. A d-c armature-controlled electric motor will be used to actuate the control surface through a reduction gear. The motor characteristics \( (\Omega, C) \) and the gear ratio required will now be determined in the following analysis by the modulus method described in Sec. 30.5.3. The motor transfer function \( M(s) \) will be calculated first.

a. Positioning the Motor Characteristic. In order to compute the maximum torque and angular velocity that will have to be developed, it is necessary to evaluate the acceleration and maximum speed of the control surface. It will be assumed (Sec. 30.5.3, Note) that they are attained through a sinusoidal movement of the control surface the amplitude of which equals the maximum deflection and the frequency of which is equal to twice the resonant frequency of the servo system, that is, about \( \frac{1}{\sqrt{10}} \) rad at an angular frequency of 20 rad/sec. Hence,

\[
A_c = 40 \text{ rad/sec}^2 \quad \Omega_c = 2 \text{ rad/sec}
\]

The aerodynamic load torque is equal to \( C_e = 0.4 \times 10^8 \) cgs. The inertia of the control surface is \( J_e = 10^4 \) cgs. If one then writes

\[
\Gamma_c = J_e A_c + C_e = 0.4 \times 10^8 \text{ cgs}
\]

the specifications for the modulus method become

\[
\frac{C_m^2}{J_m} \geq 6.4 \times 10^9 \quad \Omega_m \geq 2\alpha
\]

\[
\alpha_0 < \alpha < \alpha_1 \quad \text{with} \quad \alpha_0 \text{ and } \alpha_1 = \frac{C_M \pm (C_m^2 - 6.4 \times 10^9 J_m)^{1/2}}{80 J_m}
\]

A condition pertaining to the speed of response will also have to be verified for the transfer function that is going to be evaluated.

b. Transfer Function. The restoring torque on the control surface is equal to \( K_c \delta \). This results in the following expression for the transfer function \( \Delta/C \)

\[
\frac{\Delta}{C}(s) = \frac{K_c/\alpha}{RK_c/\alpha^3 + (K_c K_\omega + L K_c/\alpha^3)s + R/s + L J s^3}
\]

where \( C \) is the armature control voltage.

The corresponding static gain is \( K_c \alpha/RK_c \). Thus, at low frequencies one has position and not velocity control for the control surface. This is due to the restoring torque \( K_c \delta \). This expression is valid only for a small angle, the theoretical position of equilibrium never being attained in practice.

This system performs, however, with one integration in the path as soon as \( RK_c/K_c K_\omega \alpha^2 \) becomes negligible with respect to \( \omega \). It will be possible to verify that this condition is satisfied in the neighborhood of \( \omega = 10 \text{ rad/sec} \). Moreover, neglecting \( L K_c/K_c K_\omega \alpha^2 \) with respect to unity
yields
\[
\frac{\Delta}{C}(s) = \frac{1}{\alpha K_u} \frac{1}{s + (2\pi/\omega_n)s + \omega_n^2}
\]
with
\[
\omega_n^2 = \frac{K_i K_u}{LJ} \quad \frac{2\pi}{\omega_n} = \frac{RJ}{K_i K_u} = J \frac{\Omega_M}{C_M}
\]

The time constant of the motor for open-loop operation is, then,
\[
\tau \approx \left( J_m + \frac{J_c}{\alpha^2} \right) \frac{\Omega_M}{C_M}
\]

A gear ratio such that \( J_c/\alpha^2 \) is negligible with respect to \( J_m \) will be chosen. One then has, if \( \tau < 20 \) msec,
\[
\frac{C_M}{J_m} > 50\Omega_M
\]

c. **FIRST APPROXIMATION.** Starting with the value \( C_M = 10^8 \) cgs as a first approximation, this value having been obtained by analogy with similar problems, the values \( J_m = 50 \) cgs, \( \alpha = 100 \), and \( \Omega_M = 200 \) rad/sec are found to satisfy the specifications.

The maximum power at the middle point of the characteristic, in mks units, is
\[
P = \frac{C_M \Omega_M}{4} = \frac{E^2}{4R}
\]

where \( E \) is the maximum input voltage. This leads to choosing a motor of about 5 watts. The resistance of the armature circuit for a maximum voltage of 27 volts is about 35 ohms. A more complete computation will be given in the appendix to this chapter (Sec. 34.3.2, par. 5).

4. **Summary of the Transfer Functions and Compensation.** According to the preceding discussion, the four main elements have the following transfer functions:

**Rate gyro:**

Input: pitch angle in radians
Output: a voltage proportional to the signal detected by the rate gyro
\[
D(s) = K_d \frac{s}{1 + \frac{2 \times 0.6}{40} s + \frac{s^2}{40^2}}
\]

Assuming that the output signal is 27 volts for an angular velocity of 1 rad/sec, it is found that \( K_d = 27 \).

**Compensating network** (electric phase-lead network):

Input: voltage
Output: voltage
\[
R(s) = \frac{1}{10} \frac{1 + 10\tau s}{1 + \tau s}
\]
Amplifier:

Input: voltage
Output: voltage

\[ A(s) = \frac{K_a}{1 + 0.01s} \]

Servomotor:

Input: voltage
Output: control-surface deflection
Neglecting armature inductance, one obtains for this motor

\[ M(s) = \frac{1}{107} \frac{1}{1 + 0.13s + 0.001s^2} \]

Missile:

Input: deflection of the control surface in radians
Output: pitch angle in radians

\[ F(s) = \frac{1}{s} \frac{1 + 1.4s}{1 + 0.15s + 0.08s^2} \]

where \( F(s) \) is valid at approximately 2 rad/sec and over but is not valid at low frequencies.

The Nichols chart enables the adjustment of the gains and the lead-time constant \( \tau \). The static gain is \( K = K_a/400 \). The over-all amplitude-phase plot for unit open-loop gain without a compensating network is shown in Fig. 34-22, curve II. It corresponds to the transfer function

\[ \frac{1 + 1.4s}{1 + 0.15s + 0.08s^2} \frac{1}{1 + 1.2s/40 + s^2/40^2} \frac{1}{1 + 0.13s + 0.001s^2} \frac{1}{1 + 0.01s} \]

which is valid beyond 2 rad/sec.

At the present stage a second, more detailed approximation should be performed and a study of the motor transfer function should be undertaken. In order not to overcomplicate this example, the transfer function represented by curve II will be retained, and the compensation will be performed by means of phase lead.

By introducing the function \((1 + 10\tau_s s)/(1 + \tau_s s)\) and by adapting the methods given in Chap. 18, it is found that \( \tau_s = 0.01 \) sec (curve III of Fig. 34-22) produces instability. The over-all open-loop gain of the servomechanism is then approximately one\(\dagger\) when the resonance ratio is adjusted to 1.3:

\[ K \approx 1 \text{ rad/rad/sec} \]

The gain margin is 8 db, which enable the system to operate satisfactorily over a certain range of speed and altitude. The phase margin is 45°, which provides a certain margin against unwanted lags due to backlash, stiffness, aerodynamic effects, etc.

\(\dagger\) This is a usual order of magnitude for autopilots, contrary to most servo systems, where the open-loop gain is much higher in general.
The natural frequency is of the order of 13 rad/sec. Closing the loop thus results in an increase of the natural frequency, which was 3 rad/sec in the open-loop case.

5. Appendix. Suggestions for the Design of the Motor (Armature Design). A first approximation for the motor will be the reduction of the motor that is shown in Fig. 34-23, whose length-to-diameter ratio is \( \frac{L}{d} = b = 3 \). The field coil consists of a pair of poles covering approximately seven-tenths of the circumference of the armature. The armature has six slots, and its design can be performed as indicated in Sec. 31.1.4.

![Diagram of motor](image)

The notations used henceforth are those of Sec. 31.1.4. It is assumed that (1) the loss of flux is equal to 20 per cent, (2) the pole pitch is 0.7, and (3) the ratio \( \beta = 4N\pi/\pi d^2 \) is 0.3. The maximum driving torque is, according to the expression for the torque given in Sec. 31.1.4,

\[
C = 0.0175N \frac{dL B v}{R} \tag{34-5}
\]

The maximum power is

\[
P = \frac{C\Omega}{4} = \frac{v^2}{4R} \times 10^7 \tag{34-6}
\]

The resistance \( r \) of the winding of the armature is a fraction \( 1/k \) of the total resistance of the circuit. The resistance of one winding is

\[
R_1 = \rho \frac{2(L + d)}{\sigma} = 1.6 \times 10^{-4} \times \frac{2(L + d)}{\sigma}
\]

Since two circuits with \( N/4 \) winding in each are placed in parallel across the brushes,

\[
r = \frac{N}{4} \frac{R_1}{2} = 0.4 \times 10^{-4} \frac{N}{\sigma} (L + d) \tag{34-7}
\]

The quantities,

\[
C = 10^4 \text{ cgs} \quad P = 5 \times 10^7 \text{ cgs} \quad v = 27 \text{ volts}
\]

are data. The following:

\[
\frac{L}{d} = b = 3 \quad \frac{N}{d^2} = \frac{\beta \pi}{4} \cong \frac{1}{4}
\]

are estimates.

From Eq. (34-6) one obtains

\[
R \cong 35 \text{ ohms}
\]
It is now necessary to state that the current density must be less than $\delta$. Since the maximum armature current $v/R$ is only intermittent, the values of $\delta$ will be taken from 500 to 1,000 amp/cm$^2$. Since the current $v/R$ is divided into two on the brushes of the collector, one arrives at the condition

$$\frac{v}{2Re} \leq \delta$$  \hspace{1cm} (34-8)

The application of that condition to Eqs. (34-3) to (34-7) leads to the conditions

$$Bd \leq 114k \left(1 + \frac{1}{b}\right) \delta \frac{C}{P}$$  \hspace{1cm} (34-9)

$$d^2 \geq \frac{2.8}{b} \frac{C}{\delta^2 \pi / 4} Bd$$  \hspace{1cm} (34-10)

which can be written as

$$d^2 \geq \frac{P}{k} \frac{1}{40(b + 1) \delta^2 \pi r / 4}$$  \hspace{1cm} (34-11)

If the equality in condition (34-9) is chosen, then Eqs. (34-10) and (34-11) can be written as

$$Bd < 1,500k \quad d > \left(\frac{2.5}{k}\right)^{1/4}$$

The following set of values is acceptable:

- $B = 5,000$ gauss
- $d = 2$ cm
- $L = bd = 6$ cm
- $k = 7 = \frac{R}{r}
- r = 5$ ohms

The armature input impedance (see Chap. 10) must be equal to $R - r = 30$ ohms with a maximum voltage rating of 27 volts. Equation (34-5) then gives

$$N \approx 1,200$$

There are, therefore, 200 conductors per notch. Condition (34-8) then gives for $\delta = 500$ amp/cm$^2$

$$\sigma = 8 \times 10^4 \text{ cm}^2$$

If the armature is assumed to be homogeneous, and to be a complete circular cylinder with a specific density of 7.5, its inertia is

$$J_i = \frac{\pi d^2 L}{32} \times 7.5 = 75 \text{ cgs}$$

It is seen that the effect of the first gear inertia necessitates that a redesign of the motor be carried out, with the inertia $J_i$ being greater than 50 cgs. Since the results obtained are not too different from those estimated, a second approximation would be sufficient for the design of the armature.
PROBLEMS

1. Regulation of the Water Level in a Reservoir. A reservoir $C$ is used to supply water the flow of which is controlled by the tap $T$. The reservoir itself is fed by a rotating pump the flow of which is proportional to its angular velocity. It is desired to maintain the water level $H$ in the reservoir constant. For this the pump is driven by a separately excited d-c motor $M$ the field of which is produced by a current $I$. This current is controlled by the position of the water gauge by means of the potentiometer $R$ the arm of which is connected to the water gauge $F$. It will be assumed that a field-current variation will produce a variation in the speed of the pump motor.

1. Draw a block diagram of the open-loop system, defining the blocks of the diagram, the chosen main inputs and outputs, and the secondary inputs and outputs.

2. With the loop closed, draw a detailed diagram of the system in the form of a servo system. Define the blocks of the diagram and show the control input and the output. Indicate how the disturbance, the presence of which makes the regulation necessary, is introduced into the system.

3. Draw the block diagram of this system considered as a regulator; that is, consider the disturbance as the main input.

2. Block Diagram of a Beam-riding Missile. One method of guiding a missile consists in controlling it by means of a radar beam directed from the ground. It is assumed that the beam is locked onto the target and that a receiver within the missile transmits a signal $k_1x$ that is proportional to the distance $x$ between the center of gravity of the missile and the beam. To this signal is added one that is obtained from a rate gyro that is mounted on the missile. This rate gyro is used to stabilize the movement of the missile about its center of gravity; it accomplishes this

---

1 The problems grouped here arise from the authors’ teaching experience at the École Nationale Supérieure de l’Aéronautique, Paris, and at the Faculté des Sciences de Laval University, Québec. Some of them have been taken advantage of in order to discuss subjects which have been only treated briefly, owing to the space limitation of the book. Most of them have been taken from our “Problèmes d’asservissements avec solutions” (Dunod, Paris, 1957), containing 130 problems on automatic control.
by sensing the pitching rate $d\theta/dt$ of the missile. The autopilot forces the control-surface deflection $\delta$ to be proportional to the signal, which is the sum of $k_1 x$ and the rate-gyro signal $k_2 d\theta/dt$. The autopilot consists of a potentiometer pick-off which is used to detect the control-surface deflection; an amplifier with a gain $K$; and a motor $M$ the input to which is the motor control voltage, the output being the control-surface deflection. The effects of hinge moments will be neglected, and only the case of the vertical beam will be considered. In addition, the problem will be resolved into two planes, the analysis being limited to only one. The black box corresponding to the missile (input, control-surface deflection $\delta$; output, distance $x$ between missile and beam) can be separated into two black boxes in cascade, the first one (input, $\delta$; output, $\theta$) corresponding to its movement about its center of gravity and the second (input, $\theta$; output, $x$) corresponding to the movement of the center of gravity.\footnote{This is permissible mathematically, since $x/\delta = (x/\theta)(\theta/\delta)$, and physically, since the effect of the control surface is essentially that of a torque which makes the missile rotate about its center of gravity, the subsequent movement of the center of gravity being produced by a force due to the resulting angle of attack.} An aerodynamic disturbing torque around the pitch axis, the effect of which is equivalent to a control-surface deflection $\delta_0$, will be considered. The problem consists in drawing, with as much detail as the above description permits, the block diagram of the servo system that slaves the missile to the beam.

3. Equations of an Electric System. Consider the electric network shown in Fig. P3. $E$ is a constant-voltage source. The values for the passive elements are indicated in the figure. The switch $K$ is closed and the system has reached its steady-state condition. Suddenly, at time $t = 0$, the switch $K$ is opened. It is desired to study what happens after the opening of the switch.

1. Find the number of independent meshes and independent node pairs of the circuit.

2. Write the mesh equations and the essential initial conditions for the problem: i.e., the independent initial conditions with which the equations will completely define the behavior of the system for $t > 0$. Express the values of these initial conditions as a function of the known parameters.

3. Answer the same questions for the node equations.
4. Electrical Analog of a Mechanical System. A tape-feeding machine, as used in film or telegraph tape recorders, uses a mechanical filter described below to compensate for the rotational irregularities of the driving sprocket. It is desired to make the speed of the tape on the drum as constant as possible.

![Diagram of mechanical system](image)

**Fig. P4.**

Starting from an equilibrium position of the system where \( v_1 = v_2 = v_0 \), the incremental speed of the tape at the input, i.e., on the sprocket, will be designated by \( \Delta v_1 \) and the resulting irregularity of speed on the drum by \( \Delta v_2 \). In the same manner, the variations of longitudinal stress in the tape due to variations in speed will be denoted by

\[
T_1 = T_0 + \Delta T_1 \quad T_2 = T_0 + \Delta T_2
\]

The pulleys \( P \) and \( P' \) have fixed axes; the axis of the moving pulley of mass \( M \) moves in the vertical plane. The film stretch and the moment of inertia of pulleys \( P \) and \( P' \) will be neglected. The inertia \( J \) of the moving pulley about its axis will be taken into account. The angular velocity of the mobile pulley will be called \( \omega \). It will also be assumed that the film does not slide on the drum or on the moving pulley, so that

\[
r \omega = v_1 - v = v_2 - v
\]

where \( v \) is the translational speed of the pulley in the upward direction. The oil dashpot or shock absorber and the magnetic brake on the drum create a resisting force equal to \( f v \) and a tangential force equal to \( f_2 v_2 \), respectively.

1. Write the mechanical equations for the system: moment equations for the drums and the mobile pulley, force equations for the moving pulley.
2. Write the differential equation relating \( \Delta v_1 \) and \( \Delta v_2 \).
3. Derive and comment, from a filtering point of view, the transfer function \( \Delta v_2/\Delta v_1 \).
4. Determine the \( V \sim v \) electric analog of the system for the case in which the moment of inertia of the moving pulley is neglected. Indicate the values of the constants in the electric circuit.
5. Repeat (4) for the \( I \sim v \) analog.
6. Find the electric analog when the moment of inertia of the pulley is not neglected. Use a diagram including a center-tapped ideal trans-
former, or a lattice quadripole terminated by an impedance representing the drum.

5. Electric Analog of a Pickup. Figure P5 represents a phonograph pickup with vertical coil displacement. It is assumed that the stylus follows exactly the groove of the record, which is assumed infinitely rigid.

![Diagram](image)

**Fig. P5.**

The coil has $N$ turns of mean radius $r$; its mass is $M_2$. The stylus, with a mass $M_1$, is connected to the coil by means of a spring of stiffness $k_1$ and to the magnet by a spring of stiffness $k_2$ and by a mechanical resistance $f_2$. The mass $M_3$ of the magnet is connected to the frame by a stiffness $k_3$ and a mechanical resistance $f_3$. The flux density is $\beta$. The resistance of the coil is $R_L$ and its self-inductance is $L_1$. The coil is connected to an external circuit equivalent to a resistance $R_2$.

The following notation is used:

- $v_1(t)$ = vertical velocity of stylus
- $v_2(t)$ = vertical velocity of coil
- $v_3(t)$ = vertical velocity of magnet
- $i(t)$ = current in coil

where the upward direction is considered positive.

1. Write the differential equations of the system. For this, apply the basic law of dynamics projected on the positive vertical axis to the coil and to the magnet and apply Kirchhoff's law to the one-mesh network formed by the coil circuit. Also apply the law of dynamics to the stylus and explain how the equation obtained is related to the solution of the problem.

2. Write the equations of the all-electric analog system for this electromechanical system, with voltage analogous to velocity. Draw the circuit, indicating the variables and the values of the components.

3. Find, by using the topological method of Sec. 2.3.6, the all-electric analog of the system, with current analogous to velocity.

6. Transients. A sinusoidal voltage $v(t) = V_m \sin (\omega t + \psi) u(t)$ is suddenly applied to the terminals of each of the following circuits: (a) $R$ and $L$ in series, with zero initial current in $L$; (b) $R$ and $C$ in series, with zero initial voltage across $C$; (c) $R$, $L$, and $C$ in series, with zero initial current in $L$ and zero initial voltage across $C$; (d) $R_1$ and $L$ in series, shunted by $R_2$.

Is it possible to choose the phase angle $\psi$ in each of the above four cases so that there will be no transient?
7. Laplace Transforms. Give the Laplace transform of the functions of time shown in Figs. P7a and b. (Do not use the mathematical definition of the Laplace transform, but start from the Laplace transforms of simple functions and apply the lag theorem.)

\[ f(t) = \sin \omega t \quad \text{for} \quad 0 < t < 2\pi/\omega \]
\[ f(t) = 0 \quad \text{for} \quad t < 0 \text{ and } t > 2\pi/\omega \]

8. Laplace Transforms. Give the Laplace transform of the following function:

\[ f(t) = \sin \omega t \quad \text{for} \quad 0 < t < 2\pi/\omega \]
\[ f(t) = 0 \quad \text{for} \quad t < 0 \text{ and } t > 2\pi/\omega \]

9. Laplace Transforms. Give the Laplace transforms of the following functions of time. In each case specify the definition threshold and the location of the poles in the complex plane:

a. \( (e^{-at} - e^{-bt})/(a - b) \)

b. \( e^{-at} \sin (bt + \varphi) \)

c. \( te^{-at} \cos bt \)

d. \( \cos^2 bt \)

e. \( a(1 + b \sin \beta t) \sin \omega t \)

f. \( [u(t) - u(t - \pi/\omega)] \sin \omega t \)

10. Laplace Transforms. Find the necessary and sufficient conditions to be satisfied by the Laplace transform of a function of time in order that the function be periodic. Consider only Laplace transforms having the form of rational fractions.

11. Transient Response of an Electric Filter. Figure P11 shows the diagram of a filter mounted at the output of industrial voltage rectifiers;

![Fig. P11.](image)

the purpose of the filter is to smooth out the rectified voltage. \( R \) is the resistance of the inductance coil \( L \); \( R_L \) is the load impedance, assumed purely resistive. It is desired to study the response \( r(t) \) for a step function \( e(t) = e_0 u(t) \). Take \( R = 275 \) ohms, \( L = 4 \) henrys, \( C = 100 \mu F \), \( e_0 = 5 \) volts, \( R' = 2,000 \) ohms, \( R_L = 1,000 \) ohms, and \( C' = 32 \mu F \).

1. Write the equations of the system and the essential initial conditions.

2. Find directly the initial values of \( dr/dt \) and \( d^2r/dt^2 \) at the instant \( t = 0 \), and find the final value of \( r(t) \).
3. Evaluate the literal and numerical expressions for the Laplace transform $R(s)$ of $r(t)$ and find again the initial and final values obtained in part 2 by applying the initial- and final-value theorems to $R(s)$.

4. Represent the position of the zeros of the denominator of $R(s)$ in the complex plane. Perform the partial-fraction expansion for $R(s)$. (The zeros are $s = 0$, $-45.7$, and $-37.5 \pm j37.4$.)

5. Derive and plot $r(t)$.

6. Considering the filter as a system with the voltage delivered by the source as input $e(t)$ and the voltage $r(t)$ as the output delivered to the terminals of the load resistance $R_L$, what is its transfer function, i.e., the ratio $R(s)/E(s)$?

12. Transient of an Electronic Circuit. Figure P12a shows schematically a circuit belonging to a radar set. This circuit is intended for producing transient oscillations in the $LCR$ circuit when a negative pulse with a duration $T$ is applied to the grid of the tube. Figure P12b shows the equivalent circuit. $K$ is a switch that is open for the duration of the negative pulse. Compute and plot $v_2(t)$ when a negative pulse is applied at $t = 0$ and lasts until $T = 150$ $\mu$sec. Take $L = 15$ mH, $C = 250$ $\mu$F, $R = 1,500$ ohms, $R_p = 7,300$ ohms.

13. Transient Response of a Ward-Leonard System. The Ward-Leonard system, studied in Chap. 33, is based on the principle that the same machine can be used as generator or motor simply by reversing the process of energy conversion. A mechanical torque applied to the shaft transforms mechanical energy into electrical energy (generator action), whereas forcing an electric current into the machine converts it to motor action. In the Ward-Leonard system, both armature windings are in series. The generator acts as a current amplifier, deriving its energy from a mechanical source, while the motor, with separate constant excitation, transforms the electrical energy back into mechanical energy and assumes a speed proportional to its armature current. The control input to the system is the generator field current, while the output is the motor speed. The load torque on the motor constitutes a secondary input (or disturbance) to the system.
It is desired to investigate the response of the system for the two types of input:

1. Response to a control input (starting up the motor from rest)
2. Response due to a disturbance such as load-torque variation

**System Equations.** The symbols used are as follows:

- $I_f$: Generator field current
- $I$: Armature current (common to both armature windings)
- $R, L$: Resistance and inductance of both armature windings in series
- $C$: Motor load torque
- $C_2$: Developed motor torque on output shaft
- $J$: Load inertia on output shaft
- $\theta_1(t), \omega_1(t)$: Angular position and speed of generator
- $\theta_2(t), \omega_2(t)$: Angular position and speed of motor

The equations as developed in Chap. 33 are

**Motor torque:**

$$C_2 = K_2 I$$

Kirchhoff's law applied to armature winding circuits:

$$K_1 I_f = RI + L \frac{dI}{dt} + K_2 \frac{d\theta_2}{dt}$$

**Torque equation:**

$$C_2 = J \frac{d\omega_2}{dt} + C$$

Typical numerical values (cgs system of units) are:

- $K_1 = 1,500$
- $L = 0.60$ henry
- $K_2 = 3.4 \times 10^4$
- $K_3 = K_2 \times 10^{-7}$
- $R = 90$ ohms
- $J = 1.77 \times 10^{-4}$

1. Find the response to a control ramp input. Assuming that the system is initially at rest with zero load torque, calculate the response of the system to a "motor start" signal by considering the input $I_f = atu(t)$.

2. Find the transient response to a load-torque variation. The generator armature current is adjusted to give a steady-state motor speed of 1,200 rpm under a load torque of 3.5 kg-cm ($3.43 \times 10^4$ cgs units). Calculate the transient response $\omega_2(t)$ for the case in which the load torque is suddenly changed from 3.5 to 5 kg-cm.

**14. Transient Response of Hydraulic Transmission.** **Principle of Operation.** A hydraulic system is described in Chap. 32. It consists of a rotating hydraulic motor, with small pistons of fixed stroke, fed by a pump similar to the motor with the difference that the piston stroke is variable and acts as the control input to the system. The output of the system is the angular speed $\omega_2(t)$ of the motor shaft. The load torque $C$ is a secondary input to the system.
There are two important aspects in the study of the transient response of this system: (1) response to a command, or control input, such as the starting and stopping of the motor and (2) response to a disturbance such as the effect of load-torque variation on the system.

**System Equations.** The symbols used are as follows:

\[
\begin{align*}
P & \quad \text{Oil pressure} \\
V & \quad \text{Volume of oil under pressure} \\
\theta_1(t), \omega_1(t) & \quad \text{Angular position and speed of pump shaft} \\
\theta_2(t), \omega_2(t) & \quad \text{Angular position and speed of motor shaft} \\
C_2 & \quad \text{Torque developed on motor shaft} \\
C & \quad \text{Load torque applied on motor shaft} \\
\delta_1 & \quad \text{Oil displacement per radian rotation of pump shaft when } x \text{ is set at its maximum value} \\
\delta_2 & \quad \text{Motor volumetric displacement} \\
J & \quad \text{Inertia of load on motor shaft} \\
F & \quad \text{Leakage factor for the oil (leakage flow: } FP) \\
B & \quad \text{Coefficient of compressibility of oil (} \Delta P = B \Delta V/V) \\
\end{align*}
\]

The following equations, developed in Chap. 32, permit a study of the system with sufficient accuracy:

Torque produced in motor:

\[ C_2 = P \delta_2 \]

Equation of continuity:

\[ \delta_1 x \omega_1 = \delta_2 \omega_2 + FP + \frac{V}{B} \frac{dP}{dt} \]

Torque equation:

\[ C_2 = J \frac{d\omega_2}{dt} + C \]

Typical numerical values for a unit (cgs system of units) are:

\[
\begin{align*}
\delta_1 = \delta_2 & = 1 \\
\omega_1 & = 3,500 \times 2\pi/60 \text{ rad/sec} \\
J & = 1.77 \times 10^4 \\
F & = 2.47 \times 10^{-7} \\
V & = 65 \\
B & = 1.95 \times 10^{10} \\
\end{align*}
\]

1. Find the response of the system for a ramp control input. Assuming that the system starts from rest with zero load torque \((C = 0)\), find the motor response for an input function \(x = atu(t)\).

2. Find the transient response to a change in load. Assuming that \(x\) is so adjusted that the motor has a constant (steady-state) speed of 500 rpm under a steady load torque of 67.5 kg-cm \((66.2 \times 10^6 \text{ cgs})\), find the motor transient response \(\omega_1(t)\) for a sudden change in the load torque from 67.5 to 108 kg-cm.

15. **Comparison between Hydraulic and Electric Systems.** Compare the hydraulic and electric systems studied in Probs. 13 and 14 from the viewpoint of their equations and their transient performance for a sudden change in command and load torque. In particular, can the two systems be considered analogous?
16. Step and Ramp Responses of a First-order System.\textsuperscript{1} A mercury thermometer, initially at 20°C, is suddenly placed in an air current of 250 m/min at sea level and 10°C. The following temperatures are read:

<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20.0</td>
<td>10</td>
<td>16.1</td>
<td>20</td>
<td>13.7</td>
</tr>
<tr>
<td>5</td>
<td>17.8</td>
<td>15</td>
<td>14.7</td>
<td>25</td>
<td>12.9</td>
</tr>
</tbody>
</table>

The same thermometer is then placed in a balloon which ascends at 250 m/min. The following temperatures are read. What is the temperature at an altitude of 600 m?

<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
<th>Time (sec)</th>
<th>Indicated temp (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15.0</td>
<td>50</td>
<td>13.4</td>
<td>100</td>
<td>11.0</td>
</tr>
<tr>
<td>10</td>
<td>14.9</td>
<td>60</td>
<td>13.0</td>
<td>150</td>
<td>8.5</td>
</tr>
<tr>
<td>20</td>
<td>14.6</td>
<td>70</td>
<td>12.5</td>
<td>170</td>
<td>7.5</td>
</tr>
<tr>
<td>30</td>
<td>14.3</td>
<td>80</td>
<td>12.0</td>
<td>180</td>
<td>7.0</td>
</tr>
<tr>
<td>40</td>
<td>13.9</td>
<td>90</td>
<td>11.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

17. First-order System. Introduction to Filtering. An unwanted sinusoidal voltage with an amplitude of 1 volt and a frequency of 50 cps is superimposed on a periodic saw-tooth voltage $e_1$. The voltage $e_1$ is to be measured by means of a voltmeter. The inertia of the voltmeter movement will be assumed negligible, so that it can be considered as a first-order system with a time constant $T$.

![Fig. P17.](image)

1. What value can be chosen for $T$ if it is desired that (a) the indicated amplitude of the unwanted oscillations not exceed 0.2 volt and (b) the error in the reading of $e_1$ not exceed 1 volt?

2. Show that the accuracy of measurement obtained cannot exceed a certain value. Calculate this value. What is the corresponding value of $T$?

18. Determining the Transfer Functions of First- and Second-order Systems.\textsuperscript{1} It is desired to obtain the transfer function of a certain number of systems (a to g) which are known to be first- or second-order systems with transfer functions of the form $K/(1 + Ts)$ or $K/(1 + as + bs^2)$. The inputs of these systems are forces, the outputs are displacements. A static test has shown that a force of 300 g results in a displacement of

\textsuperscript{1} After C. S. Draper.
1 cm. Systems $a$ to $d$ are then submitted to a step input $a[1 - u(t)]$ the response to which is recorded in Fig. P18a to $d$. Systems $e$ to $g$ are subjected to harmonic tests the result of which is given in the form of amplitude response for $e$ and $f$ and in the form of frequency response locus for $g$. Give the expression for the transfer function of each system. Sketch the Nyquist and Nichols loci.

19. **Transfer Function of a Simple Mechanical System.** A mass $M$ is suspended by a spring the stiffness constant of which is $k$. When the mass is subjected to a vertical force $F$ in addition to its own weight $Mg$, a
certain displacement $x$ results. Find the transfer function $X(s)/F(s)$. Draw the transfer locus. Explain.

20. Study of a Phase-lead Mechanical Network. The purpose of this problem is to study a phase-lead mechanical device. The cylinder represented in Fig. P20 is filled with oil, assumed incompressible, and consists of two parts with different cross sections $A$ and $A'$. The ends of the cylinder are connected by frictionless tubing, so that the instantaneous pressure $p_1$ is the same at both ends. The two pistons, with respective cross sections $A$ and $A'$, are connected by a spring of stiffness constant $k$. The linear displacements of the pistons are respectively $x$ and $y$. The volume $V$ of the fluid between the two pistons is under a pressure $p_2$. A leak, represented by a small hole in piston $A'$, causes a fluid flow between regions $p_1$ and $p_2$ determined by Poiseuille's law:

$$\text{Volumetric flow} = \frac{1}{B} (p_2 - p_1)$$

where $B$ is Poiseuille's constant.

1. Neglecting the inertia of the small piston $A'$ as well as the external forces acting upon this piston, find the equation relating $y$ to $x$. Derive the transfer function $H(s) = Y(s)/X(s)$. Let $a = A/A'$ and $\tau = A'^2B/V$.

2. Draw, for $a = 7$, the Nyquist locus $H(j\omega)$, $\omega$ varying from zero to infinity. What can be said about the phase of such a system? What is the maximum value of this phase? What is the value of $\omega$ for maximum phase?

3. Supposing a disturbance $x(t) = u(t)$, what is the shape of $y(t)$? Calculus can be used, but physical reasoning should give a qualitative solution.

4. Compare $H(s)$ with the transfer function $V_r(s)/V_s(s)$ of the electric system of Fig. P20b when the output voltage $V_r$ is assumed to be applied to an infinite impedance.
5. Derive the transfer function \( H(s) \) for the case in which the load on the small piston is (a) an inertia \( J \), (b) an inertia \( J \) plus a constant load force, (c) an inertia \( J \) plus a load force proportional and opposite to the displacement \( y \). Comment on the results obtained. What conclusions can be drawn concerning the derivation of transfer functions?

21. Bode Plots. Sketch the Nichols loci for the following transfer functions

\[
KG(s) = \frac{32}{(5s + 1)(1.25s + 1)(0.1s + 1)}
\]

\[
KG(s) = \frac{40}{s(3.3s + 1)(s^2 + 0.6s + 1)}
\]

\[
KG(s) = -\frac{1 + 0.85s}{5.8s(1 + 0.27s)(1 - 0.48s)}
\]

22. Nyquist Loci. Compare the Nyquist loci for the two following transfer functions: \( \exp(-sT) \) and \( (1 - sT)/(1 + sT) \).

23. Principle of the \( j\omega \)-breakdown Representation for the Study of Stability.\(^1\) Consider an algebraic equation in \( s \) depending on two parameters \( \lambda \) and \( \mu \), \( F(s,\lambda,\mu) = 0 \). In the \( (\lambda,\mu) \) plane, stable regions correspond to values of \( \lambda \) and \( \mu \) for which the roots of \( F = 0 \) have negative real parts; such regions are separated from unstable regions. Boundaries which are the loci of the \( (\lambda,\mu) \) points at which \( F \) has pure imaginary roots can be obtained by writing that \( F(j\omega,\lambda,\mu) \) is zero, whatever the real quantity \( \omega \); that is, they are defined by the two equations \( \text{Re} F(j\omega,\lambda,\mu) = 0 \) and \( \text{Im} F(j\omega,\lambda,\mu) = 0 \). Once the boundary curves are obtained, the stability or instability of the different regions they limit in the \( (\lambda,\mu) \) plane is determined by testing the stability in one of these regions (taking simple values of \( \lambda \) and \( \mu \)) and noting that, each time a boundary curve is crossed, one (or many) root changes the sign of its real part.

Apply this method to Example 2 of Sec. 9.2.3:

\[
F(s,\lambda,\mu) = s^2 + (\lambda + 1)s^2 + (\lambda + \mu - 1)s + \mu - 1
\]

and to the following equations:

\[
F(s,\lambda,\mu) = 10\mu p^4 + (11\mu + 10)p^2 + (11 + \mu)p + \lambda = 0
\]

\[
F(s,\lambda,\mu) = p^2 + 10p^4 + \mu p^3 + \lambda p + 5p + 100 = 0
\]

24. Electric Quadripoles. The study of a positional servo system has shown that it had to be compensated by the product of a phase-lead

factor and an undercompensated integral control factor:

\[
\frac{1 + \alpha e^{-s}}{1 + \tau e^{-s}} \times \frac{1 + \tau_i}{1 + \alpha \tau_i}
\]

where \(\alpha = 5\), \(\alpha_i = 10\), \(\tau_e = 2.5\) msec, and \(\tau_i = 60\) msec. These factors are separately realized by means of two standard resistance-capacitance networks, located after the potentiometric sensing device and before the electronic amplifier (the grid resistance of the first tube being larger than 100,000 ohms). The natural frequency of the servo is equal to 10 cps.

1. Approximate adaptation of the two compensation networks. (a) Determine the quadripoles associated with each of the two networks (inverse transfer matrix). Compute the open-circuit input and output impedances. (b) Give a rule enabling a first approximation of the matching of these networks so that the properties of the system formed by the two cascaded networks are equivalent to the properties separately computed for each. (c) Determine which network should follow the other. Discuss this choice, give numerical values to the resistances and capacitances of the networks. Comment on the proposed adaptation.

2. Accurate matching of the two networks by means of quadripole synthesis. In this part, one will try to determine an electrical quadripole such that its characteristic impedance is a given real resistance \(R_e\) and when it is closed on the characteristic resistance \(R_c\), its transfer function is equal to the desired transfer function of the network. (a) Show that, if such quadripoles can be physically realized, the problem of the network adaptation, as stated in the first part, can be accurately solved in a way that does not depend on the frequency. (b) \(Hint\). Use a quadripole the shape of which is a symmetrical lattice (see Fig. P24a). Write the relations which correspond to the above conditions. Show that the transfer locus which corresponds to \(Z_A\) or \(Z_B\) is also a circular one. Show that the \(Z_A\) and \(Z_B\) networks are reciprocal. (c) Assuming that the lead network follows the lag network, numerically determine the corresponding quadripole. \(Hint\). The network shown in Fig. P24b may be used for either \(Z_A\) or \(Z_B\). (d) Compute the phase-lag network.

3. Adaptation of the sensing device. Assuming that the whole compensating system (consisting of the two networks) is directly coupled to a potentiometric sensing device, what can be said about the adaptation of the sensing device to the compensating devices?

25. Mechanical Quadripoles. A volumetric fuel pump is driven by the motor of a hydraulic rotative transmission which controls the pump rotational speed. The hydraulic transmission motor is supplied by a hydraulic pump which is driven at constant speed by an electrical motor. The discharge of this pump is controlled by the tilt of the pump plate,
which causes an incremental increase of the quantity of oil displaced per revolution. This tilt is the main input to the system. The motor is equivalent to a quadripole, the input variables being flow $Q_1$ and oil pressure $P_1$ and the output variables being motor rotational speed $\Omega$ and torque $T$ applied by the motor to the fuel-pump shaft. The fuel pump is also equivalent to a quadripole, the output variables being fuel pressure $P_2$ and fuel flow $Q_2$ closed on a dipole which is the load circuit (fuel circuit).

![Fig. P25.](image)

1. Write the equations of the motor and of the fuel pump. The power-conservation equation derived from the kinetic-moments equation and the volumetric-flow conservation equation will be written for each system. For this purpose, the inertias, oil leaks (proportional to the pressure), compressibility of the oil, and expansibility of the tubing will be taken into account and grouped as a unit. Write the equations in the quadripole standard form and display the inverse transfer matrices.

2. From now on, the fuel-pump factors of inertia, compressibility, and leak are assumed to be negligible; besides, the equation relating the flow $Q_2$ to the pressure $P_2$ in the load circuit may be linearized about the normal operation values (the increments of the flow $Q_2$ are proportional to those of the pressure $P_2$). What can be said about the load impedance of the motor and about the impedance matching of the various power stages? Write the condition for which the hydraulic motor is adapted to its characteristic impedance, and comment on it.

26. Autocorrelation and Spectrum Function. A random function $f(t)$, physically represented by a voltage, is equal to the sum of a message $m(t) = 2 \sin 10\pi (t + \frac{1}{2})$ and a stationary noise $n(t)$ the frequency spectrum of which is $\Phi_{nn}(\omega) = K/(\omega^2 + 100)$. The mean-square value of the noise is equal to 4 volt$^2$. Draw the autocorrelation function and the spectrum function corresponding to the random function $f(t)$.

Note. One may have to use the following integrals:

\[ \int_{-\infty}^{\infty} \frac{d\omega}{\omega^2 + a^2} = \frac{\pi}{a} \]
\[ \int_{-\infty}^{\infty} \frac{e^{i\omega t}}{\omega^2 + a^2} d\omega = \frac{\pi}{a} e^{-at} \]
\[ \int_{-\infty}^{\infty} \frac{d\omega}{(\omega^2 + a^2)(\omega^2 + b^2)} = \frac{\pi}{a + b} \left( \frac{1}{ab} \right) \]
\[ \int_{-\infty}^{\infty} \frac{\omega^2 + c^2}{(\omega^2 + a^2)(\omega^2 + b^2)} d\omega = \frac{ab + c^2}{ab(a + b)} \pi \]
27. Extrapolation. An automatic fire-control system predicts the future position of a moving target. This future position is the estimated position of the target \( T \) sec later, the estimation being sometimes called prediction or extrapolation. The simplest prediction consists in measuring the speed \( V_1 \) of the real target \( B_1 \) and defining the future position of the target \( B_2 \) by means of the equation \( B_2 = B_1 + V_1 T \). In the following analysis, only the unidimensional case will be considered and it will be assumed that the measure of the speed is obtained by differentiating the measure of the position, with an additional time constant \( \tau \).

1. If the real target has a constant acceleration \( \gamma \), what is the steady-state error in the position of the extrapolated target? What condition must be satisfied by \( \tau \) in order that this error be not larger than 10 per cent of the extrapolation distance, when the acceleration is equal to 10 m/sec\(^2\) and the speed equal to about 200 m/sec?

2. The target's speed is assumed to be constant, but a stationary noise is added to the measure of the position. The frequency spectrum of the noise is \( \Phi_{nn}(\omega) = k/(\omega^2 + a^2) \) and its mean-square value is equivalent to \( n^2 \) m\(^2\). In the expression of the error in the position of the extrapolated target, display the error in the position of the real target, and that caused by the extrapolation. With \( V = 200 \) m/sec, \( T = 20 \) sec, \( a = 10 \) sec\(^{-1}\), \( n = 20 \) m, is it possible to choose \( \tau \) so that the condition of part 1 is verified and that the rms error caused by the extrapolation is less than 5 per cent?

28. Minimization of Mean-square Error. The feedback control system shown in Fig. P28 is used to filter a noise \( n(t) \) which contaminates the message \( m(t) \). The functions \( m(t) \) and \( n(t) \) are uncorrelated and are stationary random. Their frequency spectra are

\[
\Phi_{mm}(\omega) = \frac{a^2 A^2}{\pi(a^2 + \omega^2)}
\]

and \( \Phi_{nn}(\omega) = B^2/\pi \). Is there a value of \( K \) which minimizes the mean-square error \( \bar{e}^2 = [r(t) - m(t)]^2 \)?

29. Closed-loop Transfer Function. Find the \( R/E \) transfer function for the system shown in Fig. P29.
30. **Gain Adjustment of a Servo System.** In the servo system shown in Fig. P30 the gain $k$ of the amplifier can be adjusted and

$$F(s) = \frac{0.1}{s(1 + 0.6s)(1 + 0.3s)}$$

Adjust $k$ in order that $Q = 1.3$ (a) when the transfer function $H(s)$ of the sensing device is unity, (b) when $H(s) = 2$, and (c) when

$$H(s) = \frac{2}{1 + 0.2s}$$

Compare (b) and (c).

31. **Servo Stabilization by Use of a Friction Damper.** Figure P31 shows a positional servomechanism with unity feedback. The output shaft has a moment of inertia $J_1$ and the error torque is $Ce$. A shaft with a moment of inertia $J_2$ is linked to $J_1$ by a viscous friction $f$ and to a fixed reference by a torsional stiffness $k$.

1. Write the differential equation that relates the angular positions of the input and output shafts. (Neglect initial conditions.)
2. Comment on this method for stabilization.¹

32. **Bode Plots.** Assuming the transfer functions sketched in Prob. 21 are the open-loop transfer functions of three servo systems with unity feedback, find for each system (a) the open-loop and closed-loop gains, (b) the velocity constant, and (c) the resonant frequency.

33. **Transient Response of a Servomechanism.** Consider a unity-feedback servomechanism with an input $e$ (angular position), an output $r$ (angular position), and a deviation $\varepsilon(t) = e(t) - r(t)$. It is assumed that this servomechanism is represented by the following differential equation, which relates the error $\varepsilon$ to the output $r$:

$$J \frac{d^2r}{dt^2} + f \frac{dr}{dt} = K_1 \varepsilon + K_2 \frac{d\varepsilon}{dt} + K_3 \int \varepsilon \, dt \hspace{1cm} \text{(P33-1)}$$

At time $t = 0$, when the servo has a constant angular speed $\omega$ and the transient state has been completely damped out, the input is suddenly set to zero.

1. Compute the Laplace transform $R(s)$ of the output (defined for $t > 0$).

2. For \( t = 0 \), what is the value of the integral term \( K \int \varepsilon \, dt \) in Eq. (P33-1)? Comment upon its effect.

3. Derive the open-loop and closed-loop transfer functions of the servo, assuming that the initial conditions are equal to zero.

34. Left-hand Criterion and Nyquist's Criterion. For each of the systems

\[
KG(s) = \frac{K}{s} \quad \frac{1 + s}{1 - s} \quad KG(s) = \frac{K}{s^2 - 2s + 1}
\]

sketch the open-loop transfer locus \( G(s) \) with unit open-loop gain. Give the range of values of \( K \) for which the system is stable. Does the left-hand criterion apply?

35. Routh’s and Nyquist’s Criteria. The open-loop transfer function of a servo with unity feedback is

\[
KG(s) = \frac{K(s^2 - 2s + 1)}{s(s^4 - 6s^3 + 37s^2 + 96s + 52)}
\]

and the Nyquist locus \( G(j\omega) \) is sketched in Fig. P35. Study the stability of the system without finding the zeros of the denominator of \( KG(s) \). Does the left-hand criterion apply to this system?

36. Nyquist’s Criterion. The open-loop transfer functions of three servo systems are

\[
\frac{s(s - 1)}{(s + 2)(s + 3)} \quad \frac{s(s + 1)}{(s + 1)(s + 2)(s + 3)} \quad \frac{1}{s^2(s + 1)}
\]

Study their stability if the transfer function of the feedback path is (1) unity and (2) \( s/(1 + Ts) \).

37. Nyquist Loci and Root Loci. The Nyquist locus \( KG(j\omega) \) for an open-loop transfer function \( KG(s) \) is the representation in the \( KG \) plane of the imaginary axis \( s = j\omega \) (more exactly, of its upper half \( \omega > 0 \)) of the \( s \) plane. It is possible to consider the representation in the \( KG \) plane of the parallels to the imaginary axis of the \( s \) plane \( (s = \alpha + j\omega \) with \( \alpha \) constant and \( \omega > 0 \)). The present problem attempts to show how the consideration of such generalized transfer loci can create a connecting link between the study of stability by means of root loci and Nyquist loci.

1. Consider the functions

\[
G_1(s) = \frac{1}{s(s + 1)} \quad G_2(s) = \frac{1}{s(s + 1)(s + 2)}
\]

Calling \( O \) the origin \( s = 0 \), \( A \) the point \( s = -1 \), and \( B \) the point \( s = -2 \), the value of \( G(s) \) at the point \( M \) \( (s = \alpha + j\omega) \) is

\[
G_1(\alpha + j\omega) = \frac{1}{{\text{OM}} \cdot \text{AM}} \quad G_2(\alpha + j\omega) = \frac{1}{{\text{OM}} \cdot \text{AM} \cdot \text{BM}}
\]

1 Adapted from W. G. Johnston.
Using these relations, sketch the $G_1(\alpha + j\omega)$ and $G_2(\alpha + j\omega)$ loci as $\omega$ varies from zero to infinity for different values of $\alpha$. In particular, note that the $\alpha = 0$ loci are the Nyquist loci, that the loci $\alpha = -0.5$ for $G_1$ (and $\alpha = -1$ for $G_2$) are particularly simple to draw, and finally, that values of $\alpha$ symmetric with respect to $-0.5$ for $G_1$ (and $-1$ for $G_2$) give rise to symmetric loci.

2. Use the sketches just drawn to study the equation $G_1(s)$ or $G_2(s) = -1/k$ with $k$ real and positive. In particular, discuss the sign of the real parts $\alpha$ of the roots of this equation as $k$ is varied from zero to infinity. Compare with the root loci, Figs. 14-2 and 14-12.

3. Show that the discussion of (2) leads to considering the relative position of the Nyquist locus and the $-1/k$ point, that is, that this discussion constitutes a proof of Nyquist’s criterion.

4. Repeat (1), (2), and (3) for the function $G_2(s) = 1/(s-1)(s+3)(s+7)$, that is, for a case in which the open-loop transfer function has one unstable pole.

38. Pressure Oscillations inCombustion Chambers. It is the purpose of the present problem to study the possibility of “chugging,” i.e., low-frequency oscillations that occur in combustion chambers, the pressure fluctuation being the same throughout the chamber at a given instant. An attempt has been made to explain the occurrence of such oscillations by the presence of a delay $\tau$ between the instants of injection and combustion. The present problem concerns only the case in which the injection is constant. It can then be shown that the relative pressure fluctuation $\varphi(t)$ in the chamber is governed by the equation

$$\frac{d\varphi}{dt} + (1 - n)\varphi + n\varphi(t - \delta) = 0$$

where $n$ is a number between zero and one that depends on the propellant and is mainly controlled by the pressure drop at the injectors, and $\delta = \tau/\theta$ is the delay $\tau$ nondimensionalized with respect to the residence time $\theta$, characteristic of the chamber.

1. Show that the above equation can be interpreted as that of a regulator (servo system with zero command) whose open-loop transfer function is the product of a delay factor and a first-order factor. Write this open-loop transfer function in the form $Ke^{-ts}/(1 + Ts)$ and express $K$ and $T$ as functions of $n$. Sketch $KG(j\omega)$ in either Nichols or Nyquist coordinates.

2. Show by means of the Nyquist or the left-hand criterion\(^1\) that no instability can occur if $K < 1$ (unconditional stability). What is the corresponding condition imposed on $n$?

3. If the condition just stated is not fulfilled, show that instability occurs if

\(^1\) Rocket specialists use a less general method and study the phenomenon of chugging by means of the Satche diagram. The technique of Satche diagrams and the study of some more complicated cases will be found in F. Marble and D. Cox, “Servo-stabilization of Low-frequency Oscillations in a Liquid Bipropellant Rocket Motor,” J. Am. Rocket Soc., 32(2): 63–81 (1963).
δ is greater than a critical value δc. Find the frequency of the resulting oscillation in terms of n and δ. Take τ = 0.001 sec and θ = 0.0045 sec.

39. Response to a Gust and Stabilization of a Missile. The subject of this problem is the study of a missile's response to a gust and stabilization in vertical flight. The missile is assumed to be in a vertical plane, and hypotheses are made to simplify the problem.

The missile is launched vertically. At the moment of the take-off (t = 0), there is a gust which lasts for t1 sec; its action is localized and is equivalent to a transversal force f3, applied during the interval (0, t1) at a point located at a distance l2 beneath the center of gravity G, its effect being a torque f3l2 about G (see Fig. P39a for the positive direction).

Inside the missile, a feedback control system maintains at zero the angle θ between the axis and the vertical. For this, it deflects a control surface through an angle δ related to θ (measured by means of a gyroscope) by

\[ \frac{d\delta}{dt} + k_1\delta = k_2 \frac{d^2\theta}{dt^2} + k_3 \frac{d\theta}{dt} + k_4\theta \]

where \( k_1 \) is the time constant of the servomechanism and \( k_2, k_3, \) and \( k_4 \) are the setting parameters of the automatic pilot.

This deflection creates a transversal force \( f_1 \) which is applied at a distance \( l_1 \) beneath the center of gravity \( f_1 = k_4\delta \). Its effect is a torque \( -k_4l_1\delta \) about \( G \). The quantity \( J \) is the transversal inertia momentum of the missile (rotation). The small movements of the missile about a steady vertical flight are to be studied, and the only considered aerodynamic torques are those caused by the control surface and by the gust. (The neutral point is assumed to be very close to \( G \), and the aerodynamic damping is negligible.) The initial condition for the deflection is \( \delta(0) = 0 \).

**Numerical Values.** (mts system, sn denotes sthenes. One sthen is 1,000/9.81 = 102 kg-force, i.e., about 225 lb.)

\[
\begin{align*}
J &= 105 \text{ ton-m}^2 \\
f_1 &= 1.4 \text{ sn} \\
&\quad \text{(for a 50-km/hr gust)} \\
k_2 &= 1.0 \text{ sec} \\
k_3 &= 36 \text{ sn/rad} \\
k_4 &= 10 \text{ sec}^{-1} \\
l_1 &= 6 \text{ m} \\
l_2 &= 1.2 \text{ m} \\
k_1 &= 1.3 \text{ sec}^{-1} \\
t_1 &= 5 \text{ sec}
\end{align*}
\]

1. Write the equations of the system and evaluate system stability by means of Routh's criterion.

2. Compute and draw \( \theta(t) \) for \( t > 0 \). Is Lin's method convenient for the computation of the modes? Discuss the realized stabilization and the magnitude of the oscillations for \( t > 5 \text{ sec} \).
**Problems**

*Hint.* First compute the function \( \theta_1(t) \) defined for \( 0 \leq t < \infty \), which represents \( \theta(t) \) for \( 0 \leq t < t_1 \); then, for \( t > t_1 \), \( \theta(t) \) is given by

\[
\theta(t) = \theta_2(t) = \theta_1(t) - \theta_1(t - t_1)
\]

3. Draw the block diagram of the stabilized missile, as shown in Chap. 1. State what each block represents and give the respective input and output variables. Give the algebraic and numerical values of the following transfer functions (servo viewpoint):

- Transfer function of each block
- Transfer function of the open-loop system
- Transfer function of the closed-loop system
  (input, desired pitch \( \theta_2 \); output, pitch \( \theta \))

4. Give the algebraic and numerical values of the following transfer functions (regulator viewpoint):

- Open-loop transfer function
- Closed-loop system function
  (input, disturbance \( f_2 \); output, pitch \( \theta \))

5. Study the stability by means of the Nyquist criterion. Determine the value of the resonance ratio \( Q \) and the phase and gain margins. Give the damping ratio and the overshoot of the second-order systems which have the same value of \( Q \).

6. (a) What can be done to improve the stability margin? Consider in particular the effect of increments in the system's setting parameters; by means of the Nyquist diagram, evaluate their influence on the open-loop static gain and on either the phase lead or the phase lag of the forward path. (b) What would happen if the unwanted time lags, which have been neglected when writing the equation of the automatic pilot, were taken into account? (c) Discuss the rule \( Q = 1.3 \) as applied to this case. Is the control system a regular one (i.e., similar to a second-order or second-order-plus-lag feedback control system in the vicinity of the resonant frequency)?

For all of this question, only qualitative considerations and statements that are as precise as possible need be made regarding the effects of increments of \( k_3 \), \( k_2 \), and \( k_1 \)—omitting the case where \( k_2 = 0 \), which will be considered in the following question.

7. In order to simplify the study, it is now assumed that \( k_2 = 0 \). How does that affect the problem? (Use the Nyquist diagram.) (a) Determine the optimum values of the parameters \( k_3 \) and \( k_4 \) with the following criteria: stability margin defined by \( Q = 1.3 \), high open-loop gain. (b) Draw an approximate sketch of the response of the missile to the above gust, with the chosen setting values. Discuss the application to this case of the high-resonance-frequency criterion. Give a simple way of suppressing the error caused by a constant disturbance, by adding an element to the forward path.
**PROBLEMS**

*Hint.* The open-loop transfer function being written as

\[ KG(s) = \frac{K(s + aTs)}{s^2(s + Ts)} \]

use the Nichols locus, setting \( K \) for different values of \( a \). A lower bound will be chosen for \( a \), the upper bound being \( a = 15 \).

8. Discuss the stabilization by means of the root-locus method.

![Nyquist plot](image)

**Appendix.** The Nyquist plot of

\[ KG(s) = \frac{15.8(0.1s^2 + 0.35s + 1)}{s^2(0.77s + 1)} \]

is shown in Fig. P39b.

40. **Phase-lead Compensation of a Regulator.** The forward path of a speed regulator consists of an amplifier with a variable gain \( k \) which supplies the field current to the generator of a Ward-Leonard system (Fig. P40). The characteristics of the latter are:

**Generator (constant angular speed)**
- Inductor resistance: 50 ohms
- Inductor inductance: 0.5 henry
- Generator open-circuit voltage per unit field current: 500 volts/amp.

**Motor (constant field current)**
- Motor torque constant: 1 newton-m/amp
- Motor back emf: 1 volt/rad/sec
- Total inertia: 0.01 kg \( - \) m\(^2\)
The resistance of the armature circuit is 10 ohms and its inductance is 10 mh. The mechanical frictions will be neglected. The voltage gain of the amplifier will be assumed to be independent of the frequency. The feedback loop incorporates a tachometer generator producing an output of 110 volts at a speed of 3,000 rpm. The nominal torque of the motor is 3 newton-m and this corresponds to 3 amp nominal intensity.

1. Write (after Laplace transformation) the relation that exists between the input voltage to the Ward-Leonard system, the rotation speed \( \Omega_m \) of the motor, and the load torque \( T_m \) on the motor shaft.

2. The feedback loop being open, draw the steady-state characteristic speed \( \Omega \)-load torque \( T \) for a zero load speed of 3,000 rpm. With respect to the zero load speed, what is the percentage of the speed loss that is caused by a load torque on the shaft of the motor equal to the nominal torque of the motor?

3. Set the gain \( k \) so that the closed-loop peak value of magnification is \( Q = 1.3 \). What is the value of the corresponding natural frequency?

4. Repeat (2) for the closed-loop system, the setting being that found in (3). For this case, what is the meaning of the rule \( Q = 1.3 \)?

5. A phase-lead network is inserted in the feed forward path before the amplifier. Its transfer function is \( (1 + 10\tau s)/10(1 + \tau s) \). What must be the value of \( \tau \)? First, choose the value \( \tau_1 \) for which the maximum phase lead is obtained for uncompensated resonance frequency (this value is too large). The resonance frequency thus obtained gives a second approximation \( \tau_2 \). Then draw the Nichols loci which correspond to values of \( \tau \) in the neighborhood of \( \tau_2 \) (20 per cent difference) and observe the values of \( k \) and \( \omega_R \) which give the peak value of magnification \( Q = 1.3 \). To choose \( \tau \), one will state precisely the meaning of the chosen criteria. What is the setting value of \( k \)?

6. Same questions as in (2) for the above settings. Discuss the compensating network and its setting when the system is used as a variable-speed control device rather than a speed regulator.

7. The network of (5) is shown in Fig. P20b. Given \( C = 32 \text{muf} \), what must be the values of \( R_1 \) and \( R_2 \) to obtain the settings of (5)? Give an approximate value of the amplifier’s minimum driving-point impedance for which the foregoing computation is valid.

41. Integral Control. Networks of the type \( (1 + r s)/(1 + b r s) \) \((b > 1)\) are rather often used in series in the low-power stage of the forward path of a feedback control system the open-loop gain of which has to be increased. Explain why the network \( 1/(1 + r s) \), which is also a phase-lag network, is not used instead.

42. Introduction to Compensation. A positional servomechanism is actuated by a field-controlled d-e motor. The load is assumed to be properly represented by an inertia \( J \) and a viscous friction \( f \). The driving torque \( C \) is proportional to the field current \( i \). (The inductance of the field coil is neglected.) This current, produced by a vacuum-tube amplifier, is proportional to the angular error \( \varepsilon = e - r \), where \( e \) is the input angular position and \( r \) is the response indicated by a potentiometer. Thus
\[ C = K_1i = K_1K_2e = J\frac{d^2r}{dt^2} + J\frac{dr}{dt} \]

(It will be assumed that the gain of the potentiometer is included in \( K_2 \).)

1. Write the open-loop transfer function of the system. What is the velocity constant \( K_r \)? Assume for the rest of the problem that the time constant is 0.05 sec.

2. Adjust \( K_2 \) for \( Q = 1.3 \), using Nyquist or Nichols loci. What are \( K_r \) and \( \omega_R \)?

3. A phase-lead network with a transfer function

\[
\frac{1}{\frac{1 + a\tau_\alpha s}{1 + \tau_\alpha s}}
\]

is inserted before the amplifier \( K_2 \). Take \( a = 10 \) and \( \tau_\alpha = 0.005 \) sec. What is the new adjustment of \( K_2 \) for \( Q = 1.3 \)? (Use Nichols chart.) What are \( K_r \) and \( \omega_R \)? How are the results modified if \( \tau_\alpha \) is made twice as large, or twice as small?

4. Answer the questions of (3) for the case in which the network inserted is an undercompensated integral network with a transfer function

\[
\frac{1 + r_i s}{1 + b r_i s} \quad \text{with} \quad b = 10, \ r_i = 0.4 \ \text{sec}
\]

5. The circuits of (3) and (4) are removed but a tachometric feedback \( K_sG_s(s) = K_s \) is introduced, so that the field current is now

\[ i' = i - K_3\frac{dr}{dt} \]

Suppose \( K_3 = 9f/K_1 \) (units: radians, seconds). How should \( K_3 \) be chosen for \( Q = 1.3 \)? Compare with (2). What are \( K_r \) and \( \omega_R \)?

6. Compare the methods used in (3), (4), and (5).

43. Compensation of a Hydraulic Servo. The forward path of a positional servo with a hydraulic servomotor comprises (a) a sensing device for the error \( \varepsilon = e - r \); (b) a passive compensating network whose output is a voltage \( E \); (c) an electronic amplifier with a gain \( K_1 \), the output of which is a current \( I \); (d) a constant-torque motor which slaves the position \( X \) of the valve (which is restrained by a spring) to the current \( I \) [assume \( X/E = K_1 \lambda/(1 + \tau s) \), where the time constant \( \tau \) accounts for the lags in the control of the valve]; and (e) the hydraulic power stage, which is assumed linear and which controls the angular rate of the output shaft according to the equation

\[
\frac{R}{X}(s) = \frac{K_3}{s^2 + 2\omega_n s + \omega_n^2}
\]

where \( \omega_n \) is due to oil compressibility and \( \varepsilon \) to leakage.
PROBLEMS

Take for the open-loop transfer function

\[
\frac{R}{\varepsilon} = \frac{K_1 K_3 \lambda}{s(s^2 + 2\omega_n s + \omega_n^2)(\tau_n + 1)}
\]

the following numerical values:

\[
\omega_n = 130 \text{ rad/sec} \quad \tau = 0.3 \quad \tau_n = 0.02 \text{ sec}
\]

It is desired to boost the velocity constant and to increase the bandwidth of this servo system. For this, an RC circuit is used with a transfer function

\[
\frac{1 + \tau_i s}{1 + \alpha_s \tau_i s} \frac{1 + 1 + \alpha_s \tau_i s}{1 + \alpha_s \tau_i s}
\]

where the time constants \(\tau_i\) and \(\tau_a\) should not be too high in order to avoid too large values for the capacitances, and the time-constant ratios \(\alpha_i\) and \(\alpha_a\) should not exceed 10. Is it possible to multiply the velocity constant by 15 and the bandwidth by 2? Discuss the influence of unwanted time lags at the amplifier stage.

44. Compensating Network in the Feedback Path. An angular-position servo system has in its forward path two amplifiers which have adjustable gains \(k_1\) and \(k_2\). The motor and its load, which includes a speed-reducing gear, the ratio of which is \(\alpha = \theta_m / \theta_r = 50\), may be represented by a first-order transfer function relating the control voltage \(V_n\) to the angular speed \(s\theta_m\), the gain being \(k_m = 10 \text{ rad/sec/volt}\), and the time constant \(\tau_m = 0.25 \text{ sec}\). Furthermore, it is desired that an output rotation \(\theta_r = 1 \text{ rad}\) correspond to an input voltage \(V_n = 2 \text{ volts}\), the origin of \(\theta_r\) corresponding to \(V_n = 0\).

![Fig. P44.](image)

1. Determine the gains \(k_1\) and \(k_2\) of the amplifier and the potentiometer gain \(k_3\) (volts per radian) which correspond to a value of the resonance ratio \(Q\) equal to 1.3. What is the value of the angular deviation \(\varepsilon_0\) which corresponds to a linear input \(dV_n/dt = 1 \text{ volt/sec}\)?

2. A tachometer generator is attached to the motor shaft \(\theta_m\) and its output signal is fed back to the amplifier \(k_2\) as shown by the dotted line in Fig. P44. The gain of the tachometer is \(g = 0.025 \text{ volt/rad/sec}\). Is it possible to set \(k_1\) and \(k_3\) so that the angular-position deviation \(\varepsilon_0\) is
reduced by a factor of 4 and that the peak value of magnification remains equal to $Q = 1.3$? What is the value of the resonant frequency?

3. Before being fed back into the amplifier $k_2$, the output voltage of the tachometer is fed through a high-pass filter made of a resistance $R$ and a capacitance $C$. The gain $k_2$ is assumed to have the same value as in (2). What deviation $\varepsilon_0$ can be obtained if $k_1$ is so set that the peak value of magnification is equal to 1.3? Show that the deviation $\varepsilon_0$ of (2) can be obtained with a smaller value of the gain $k_1$. What is the new resonant frequency? In which case is such a compensation network similar to a phase-lag network placed in the forward path?

4. The tachometer of (2) is replaced by an angular accelerometer the output of which is a voltage proportional to the angular acceleration of the motor shaft. How must the resistance $R$ and the capacitance $C$ be placed in order to obtain a result theoretically similar to that of (3)? Discuss the use of the angular accelerometer alone, without compensation network.

45. Comparison of Different Techniques for Compensation. Given a conventional angular-position feedback control system the load of which is a pure inertia, the following devices may be used to compensate the system: (a) a viscous damping acting directly on the output shaft, (b) a passive phase-lead network in the forward path, (c) a passive undercompensated integral control network (phase lag) in the forward path, and (d) an auxiliary tachometer feedback. Briefly compare these four kinds of compensation from a qualitative point of view.

46. Autopilot for a Missile. It is required to stabilize a missile in vertical flight. The missile is piloted by means of control surfaces located at point $B$, behind the center of gravity $G$. The only case which will be considered is that of motion of the missile in the plane of Fig. P46, the control surfaces being perpendicular to the figure. The roll, i.e., the rotation about the longitudinal axis, is assumed to be stabilized by other means. Study the stabilization for a given constant speed $V$ of the missile.

The aerodynamic forces which are to be considered are resolved as follows: (a) Along the missile's speed vector, there are a constant aerodynamic drag $D$, applied at the center of thrust $C$, and a negligible control-surface drag. (b) Along a perpendicular to the direction of the speed vector there are a lift force $K_1\ddot{i}$, applied at the center of thrust $C$, and a force $K_2 d$, where $K_i$ is a constant, $i$ is incidence of the missile (i.e., the angle between the speed vector and the axis of the missile), and $d$ is the deflection of the control surface. The engine produces a constant force $T$ along the axis of the missile. The angles $\theta$, $\gamma$, $i$, and $d$ are assumed to
be small. The inertia of the missile about the axis perpendicular to the
figure at point G is J. Numerical values in mts units (sn = sthenes) are:

\[ a = 0.44 \text{ m} \quad M^* = 2.5 \text{ tons} \]
\[ l = 4.80 \text{ m} \quad J = 20.2 \text{ ton-m}^2 \]
\[ K_s = 950 \text{ sn/rad} \quad T = 61 \text{ sn} \]
\[ K_d = 68 \text{ sn/rad} \quad D = 36 \text{ sn} \]
\[ V = 610 \text{ m/sec} \quad g = 10 \text{ m/sec}^2 \]

1. Write the equations of the system (but first give serious considera-
tion to the problem). In what manner is the assumption of constant
speed used when the equations are written?

Compute algebraically and numerically the following relations (Laplace
transforms): \( \gamma/\theta(s) \), \( i/\theta(s) \), \( \theta/d(s) \), \( \gamma/d(s) \) and state their physical
meaning.

2. The computations of (1) give in an approximate way the following
relations (Laplace transforms):

\[
\frac{\theta}{d}(s) = \frac{1 + 1.9s}{(60s - 1)[1 + (2 \times 0.08/4.6)s + (s/4.6)^2]}
\]
\[
\frac{\gamma}{d}(s) = \frac{1 - (s/15)^2}{(60s - 1)[1 + (2 \times 0.08/4.6)s + (s/4.6)^2]}
\]

Show that stabilization is needed for vertical flight.

The autopilot may include the following devices: The sensing device,
which is either a gyroscope (detecting \( \theta \)) or a rate gyro (detecting \( d\theta/dt \)).
For the power stage, one may choose either position or velocity piloting
(Sec. 14.3).

Discuss the stability of the system for each of these four cases, assuming
that the unwanted time lags in the forward path (sensing device, ampli-
fier, motor, etc.) are negligible and that the feedback systems are not
compensated. For this study, the Nyquist loci should be roughly
sketched by use of the asymptotic Bode diagrams.

3. For one of the combinations studied in (2), the setting of the gain
with a proportional control does not stabilize the system. State whether
it is possible to stabilize it by means of a conventional passive compensa-
tion network (either phase lead or phase lag). If the answer is yes,
give the values of the parameters of the network.

4. For the stable cases of (2), give the values of the open-loop gain,
with the correct units, and of the resonant frequency of the feedback
control system, assuming that all the unwanted time lags are equal to
zero. What is the value of the over-all gain? Compare the different
cases, considering the system as a regulator and assuming that the dis-
turbing aerodynamic torques have the same effect as an equivalent deflec-
tion of the control surface.

5. It is now assumed that the actual behavior of the elements other
than the sensing device may be correctly represented by a 0.05-sec time

\^[1\] For definition of a sthene, see Prob. 39.
constant and that the rate gyro has a natural frequency equal to 30 rad/sec and a damping ratio equal to 0.5. How are the performances modified in the different cases? Would the real elements, which have the above characteristics, be satisfactory?

6. For the different cases, study qualitatively—giving only an approximation of the value of the network parameters—whether it is advisable to insert lead and lag networks in the forward path in order to increase the open-loop gain of the feedback control system.

47. Optimum Linear Filters (Wiener). It is desired to transmit a message which is contaminated by noise. The message and noise are noncorrelated stationary functions with frequency spectra

\[ \Phi_m(\omega) = \frac{1}{\omega^2 + 1} \quad \Phi_n(\omega) = d^2 \]

1. Find the transfer function of the optimum filter in the sense of the rms error criterion.

2. Find the optimum "predicting" filter \((\alpha \text{ seconds in advance})\), assuming \((a) d = 0 \text{ and } (b) d \neq 0\).

48. Steady State of On-Off Servos. The forward path of a positional servo (command \(e\); output \(r\)) with unity feedback consists of: (1) the sensing device (input, error \(\varepsilon\); output, \(\eta = K_\varepsilon \varepsilon\); (2) a relay with dead zone \(\Delta\) and hysteresis \(h\) (input \(\eta\); output \(w\); output magnitude \(M\); negligible lag); (3) a motor that develops a torque \(K_2w\) on a load with predominant viscous friction \((dr/dt = K_4C\), where \(C\) is the sum of \(K_2w\) and of a disturbing torque \(C'\)).

Draw the block diagram of the system. Briefly analyze the behavior of \(\varepsilon(t)\) and analyze more specifically the steady-state error in each of the following cases: (1) \(e = au(t), C' = 0\); (2) \(e = atu(t), C' = 0\); (3) \(e = 0, C' = au(t)\). Show the influence of the gains \(K_1, K_2, \text{ and } K_3\) and of the quantities \(\Delta, h, \text{ and } M\) on the steady-state error. Compare these results with the case in which the relay is replaced by a linear amplifier \(w = K_\eta\).

49. Performance and Compensation of a Contactor Servomechanism using the First-harmonic Approximation. Assume the relay is excited by a current \(k_1\varepsilon\), with \(\varepsilon\) in radians, and the steady-state angular velocity of the motor is \(k_m\) times the voltage at the input of the relay \((k_m\) takes the variable gear ratio of the motor into account). Assume the inactive zone \(\Delta = 2 \text{ ma}\) and consider successively \(h = 0\) and \(h = 1 \text{ ma}\).

1. Plot the critical locus \(-1/N(\varepsilon)\) of the relay (use Nichols coordinate).

2. What is the open-loop gain \(K\)? For what value \(K_0\) of \(K\) is there a limit cycle in the absence of an input signal? Find its amplitude \((\theta_0\) in radians) and frequency. For what value \(K_1\) of \(K\) is the amplitude 2 \(\theta_0\)? What is the corresponding frequency?

3. A phase-lead network with a transfer function \((1 + a\tau s)/(1 + b\tau s)\) is inserted after the potentiometer. The gains of the components are assumed to be so adjusted that \(K = K_1\). Choose \(a\) and \(\tau\) in order to eliminate the limit cycle.

4. Discuss the effect of a phase-lag network \((1 + \tau s)/(1 + b\tau s)\) for
compensating the system. (Specify the influence of the gain and that of
the amplitude response curve.)

50. On-off Stabilization of a Missile. A missile is stabilized by
spoilers (Sec. 22.3.3) which apply to it a constant rolling torque ±C.
The spoilers are actuated by electromagnets into which is fed the current
from a roll gyroscope, so that the aerodynamic torque due to the spoilers
is −C sign φ, where φ is the roll angle.
1. Neglect all lags and thresholds. Assume the missile is a pure
inertia J (Sec. 14.3.6). Study the stability of the system. What relation
exists between the amplitude and the period of the oscillation?
2. Assume the electromagnet has a constant lag θ. Show that the
commutation points in the phase plane are located on two straight lines.
3. If phase-lead compensation (1 + ars)/(1 + rs) is introduced, find
the frequency, the amplitude, and the stability of the limit cycle by using
the first-harmonic approximation. How are the frequency and amplitude
affected by C and J? Adjust r when a = 3 and a = 10.
Numerical values: J = 10^6 cgs, C = 5 × 10^7 cgs, θ = 0.05 sec.

51. On-off Servomotor with Tachometer Stabilization. An angular
position feedback control system controlled by relays and stabilized by a
tachometer feedback is shown in Fig. P51. The output of the relay

![Fig. P51.](image)

is a current i given by i = i_M sign c. The motor produces a torque
Γ = k_i, applied to a load which is a pure inertia J. The problem should
first be studied algebraically and then numerically:

\[ J = 10^6 \text{ cgs} \quad \Gamma_M = k_i \lambda = 10^7 \text{ cgs} \quad \lambda = 0.2 \text{ sec} \]

1. Write the differential equation defining the output r for a positive
control variable c of the relay and also for a negative one, assuming that
the input e is zero (regulator viewpoint). Write \( \Gamma_M/J = k_iM/J = a \).
2. Assuming that the initial conditions \( r_0 \) and \( (dr/dt)_0 = r'_0 \)
are given, give as functions of time the solutions of the two foregoing
differential equations. Draw in the phase plane the corresponding trajectories.
Their parametric representation as functions of time is \( x = r(t) \) and
\( y = dr/dt \). Show a commutation line. How can these trajectories be
graduated as a function of time?
3. Assume now that the relay has a constant time delay r which does
not depend on the switching sense (numerical application: \( r = 0.025 \text{ s} \).
Choosing as the time origin the moment when the control variable c
switches from \( 0^+ \) to \( 0^- \) (at point \( x_1, y_1 \)), compute the coordinates \( (X_1, Y_1) \) of the point where the switching actually takes place (change of sign of \( t \)). Show that the locus of points \( (X_1, Y_1) \), which are called negative switching points, is a line \( \Delta^- \). Write the equation of this line, as well as that of the locus \( \Delta^+ \) of the positive switching points.

4. Assuming \( \lambda > 0 \) and \( \tau < \lambda \), sketch in the phase plane the curve representing the system's evolution from the initial conditions \( (x_0, y_0) \). Display in a simple way a periodical oscillation, and compute its period \( T \) and its magnitude \( A \). Show that \( T \) and \( A \) can be written as

\[
A = a \frac{\tau^2}{2} f^2 \left( \frac{\lambda}{\tau} \right)
\]

Draw the curve \( f(\lambda/\tau) \) for values of \( \tau \) smaller than \( \lambda \). Comment upon the influence of the quantities \( k, J, \lambda, \tau \). Draw \( r(t) \) for the given numerical values and \( x_0 = 1^o, y_0 = 0 \).

5. Write the describing functions of the system. Show that there is a limit oscillation. Compute the period \( T'' \) of this oscillation and draw the curve giving in this case the ratio \( T''/4\pi \) as a function of \( \lambda/\tau \). Determine the magnitude \( M' \) of this limit oscillation. Compare these results to those obtained by means of the phase-plane method.

6. Consider the case where the relay has a symmetrical dead zone.
(a) Assuming that the relay has no time lag, show that, for \( \lambda > 0 \), the trajectory cannot be produced beyond the stopping point (nonzero speed). It is not possible to obtain a limit oscillation with a nonzero magnitude.
(b) Assuming that the relay has a time lag \( \tau \), show that in the phase plane the switchings will take place on either two or four lines, and state precisely the different cases. Geometrically determine the limit oscillation.

52. Servo System with Coulomb Friction and Saturation. A position-type servomechanism develops a torque \( k\varepsilon \) proportional to the angular error \( \varepsilon = e - r \). The load consists of an inertia \( J \) with static friction \( f_0 \) (i.e., a torque of constant magnitude \( f_0 \) with a sign opposite to the angular velocity \( dr/dt \)). Take \( J = 10^4, f_0 = 10^6, k = 10^7 \) (all figures are in cgs).

1. Assume zero command and an initial condition \( \varepsilon_0 = 0.1 \) rad. Study the phase trajectory in the plane \( [x = \varepsilon(k/J)^{1/2}, y = d\varepsilon/dt] \). Prove that the peaks occur at periodic instants and are located on straight lines. Compare with the case of viscous friction. Sketch \( \varepsilon(t) \). How is its shape affected by \( k \) and \( f_0 \)?

2. The motor is now assumed to have angular-velocity saturation at \( dr/dt = y_M = 0.6 \) rad/sec. Show that the presence of saturation results in an increase of damping and speed of response. What is the maximum allowable value for \( y_M \)? (Consider both the regulator and the servo viewpoints.)

3. Now assume torque saturation at \( C = C_M = 3 \times 10^4 \) instead of angular-velocity saturation. Study the phase-plane trajectory and the time response \( \varepsilon(t) \). What is the influence of \( C_M \) on the system's damping and speed of response?
53. Oscillation of On-off Feedback System. The Nichols locus of an SE-1010 airplane in longitudinal motion for horizontal flight at an altitude of 8,000 m and with a forward velocity of 120 m/sec is shown in Fig. P53 (scaling is in radians per second). Assume this airplane is stabilized by an on-off autopilot without dead zone and hysteresis with a transfer function \( k(1 + 0.5s)/(1 + 1.5s) \) for the linear part. Find the frequency and amplitude of the possible oscillations and discuss their stability by using (1) the first-harmonic approximation and (2) the Tsypkin approach.

54. Duration Modulation. A duration-modulation sensing device modulates the durations of the positive and negative values of a voltage switching periodically between the values +E and -E. The mean value of the rectangular voltage thus obtained is proportional to the input variable of the sensing device. This voltage is the input to a filter the transfer function of which is approximated by a time constant as \( 1/(1 + Ts) \).

In this case, a residual oscillation is added to the mean value of the output. The magnitude of this oscillation must be smaller than 0.05 E. Moreover, the phase shift caused by the sensing device and the filter must be smaller than 20° for a 3-cps input sine wave. Give a satisfactory approximate value of \( T \) and of the switching frequency of the sensing device.

55. Linearization of a Relay. In a regulator, the sensing device gives the pressure and its derivative; it is made as follows: A capsule (sensitive to the pressure \( P \)) drives the support rod of an elastic blade which is the mobile armature of an electromagnetic relay \( R \). The vibrating relay is in the armature circuit of a motor \( M \), the field voltage of which is constant. This motor acts on the base of the capsule by means of a gear train. The detected outputs of the system are the motor rotation \( \theta \) and the voltage \( v \) at the terminals of the armature.

1. Explain qualitatively the general working of the system. Show that it is a feedback control system, and sketch its block diagram.
2. Describe the shape of the signals along the feedback loop. Neglect

---

1 For example, altitude stabilization of an aircraft.
the forces caused by the blade elasticity with respect to the increments of the forces exerted by the internal pressure on the capsule.

3. Briefly analyze how the relay vibration (frequency $F$) enables one to linearize the characteristics of the relay. What is the interest of this device?

4. What is the transfer function which relates the actual pressure $P$ to the measured pressure?

**Numerical Values.** As soon as $x_1$ is equal to or larger than 1 mm, the relay sticks completely on one side; the armature voltage is then a maximum and equal to 25 volts. A 100-millibar ($10^4$ cgs) pressure increment causes a 2-mm elongation of the capsule.

![Diagram](image)

**Fig. P55.**

The transfer function relating the motor armature voltage and the rotation of the loaded motor shaft is equal to

$$\frac{\theta}{\varepsilon} = \frac{2}{s(1 + 0.04s)} \text{ rad/volt}$$

The gear ratio is

$$\alpha = \frac{x_2}{\theta} = \frac{1}{50} \text{ mm/rad}$$

Give an approximate value of the residual relay vibrations which appear on $\theta$ and on its derivative (measured by their pressure equivalent), when the vibration frequency is equal to 50 cps.

**56. Equivalent Block Diagrams for Servo Systems Incorporating One Nonlinear Component.** For each of the four on-off servo systems briefly described below derive (1) the usual block diagram (call $x$ and $y$ the input and output of the nonlinear component) and (2) the equivalent block diagram of the type shown in Fig. 22-30 (single loop, unity feedback, error signal at the input of the nonlinear element). Give the expression for $L(s)$ in each case.

a. **Follow-up system.** Figure P56a shows an elementary follow-up system in which $\theta_s$ is slaved to $\theta$. The error signal is applied to the relay $N$ which actuates in one or another direction the servomotor $M$ that

---

1 Adapted from J. Tsypkin,
drives the output shaft through the gear \( R \). (The auxiliary transformer \( T \) is merely intended for the purpose of introducing dynamic lubrication by means of dither effect.) Assume the sensing of the error is ideal and the control of the servomotor by the relay is represented by

\[
T_1 T_2 \frac{d^2 \theta_0}{dt^2} + T_2 \frac{d^2 \theta_0}{dt^2} + \frac{d\theta_0}{dt} = k_M y(t)
\]

Assume the relay has a delay of \( T_R \) sec.

b. Temperature regulator (Fig. P56b). The temperature \( \theta \) of a furnace \( F \) is measured by means of a resistance \( R \) inserted in a Wheatstone bridge. Any unbalance in the bridge causes the relay \( N \) to actuate, in one or another direction, the servomotor \( M \) which controls the admission valve \( V \) into the furnace. A tachometric feedback consists of the mechanism (dashpot and spring) shown in the figure, which changes the null point of the bridge. Assume that the gas flow \( v \) is controlled by the servomotor according to \( T_M \frac{dv}{dt} = y \) and that the tachometric feedback gives

\[
T_E \frac{d\eta}{dt} + \eta(t) = T_E \frac{dv}{dt}
\]

where the quantity \( \eta \) is added to the measured temperature at the input of the relay. Assume the thermocouple has a time constant \( T_R \): \( T_E (d\theta_m/dt) + \theta_m(t) = k_R \theta(t) \) and that the temperature \( \theta \) in the furnace builds up according to a first-order equation: \( T_F (d\theta/dt) + \theta(t) = k_F v(t) \).
c. Voltage regulator (Fig. P56c). The voltage \( V(t) \) across the generator is applied to the solenoid \( S \). When \( V(t) \) exceeds a certain value, the relay \( N \) operates and opens the contact \( C \). This decreases the conductance of the field circuit (not shown in the figure) of the generator and thus produces a drop in the field current and in \( V(t) \). When \( V(t) \) has dropped enough, the spring \( r \) causes the relay to close the contact \( C \), and the process starts again. Assume that, for small changes in the variables, \( V(t) \) is related to the conductance \( y \) of the field circuit by \( T_o \, dV/dt + V(t) = k_o \, y(t) \), the current \( i(t) \) in the solenoid is established according to \( T_r \, di/dt + i(t) = k_i \, V(t) \), and the motion of the arm of the relay is represented by a second-order equation

\[
\frac{d^2 z}{dt^2} + 2\xi \omega_n \frac{dz}{dt} + \omega_n^2 z(t) = k_f i(t)
\]

The nonlinear relation between \( z \) and \( y \) is shown in Fig. P56d, where \( y_0 = \frac{1}{2}(y_{\max} + y_{\min}) \).

d. Autopilot. An airplane is roll-stabilized as follows: The roll angle \( \varphi \) is detected by a gyroscopic sensing device and sent into a compensating network with a transfer function \( Z/\Phi = (1 + a_1 s + a_2 s^2)/(1 + b_1 s + b_2 s^2) \). The corrected signal \( z(t) \) operates a relay the output \( y(t) \) of which controls the aileron deflection \( \delta \) according to \( T_M \, d^2 \delta/dt^2 + d\delta/dt = k_i y(t) \). An auxiliary signal \( z_r = k_r \delta \) is fed back to the input of the relay, where it adds to the corrected error signal \( z \). Assume that the dynamic behavior of the airplane is represented (Sec. 14.3.6) by

\[
J \frac{d^2 \varphi}{dt^2} + f \frac{d\varphi}{dt} = A \delta
\]

57. Forced Oscillations of a Particular Nonlinear Servo.\(^1\) Jump Phenomenon. Consider the servo shown in Fig. P57a. Take \( K_1 = 1 \), \( K_2 = 3 \), \( T_1 = T_2 = 1 \) sec and assume the characteristic of the nonlinear component is that shown in Fig. P57b.

1. Find the describing function \( B(x_1) \) for the nonlinear component. Derive the expression for \( L(s) \) of the equivalent block diagram of the type shown in Fig. 22-30.

2. Show the existence of a limit cycle. Is it stable or unstable?

\(^1\) After C. H. Wilts.
3. When the system is subjected to an input \( e = e_1 \sin \omega t \), derive the expression for \( X/E \) as a function of \( e_1 \), \( \omega \), and \( B(x_1) \). Plot \( e_1 \) vs. \( x_1 \) for values of \( \omega \) ranging from 0.1 to 2.0. Interpret the results obtained in terms of the physical behavior of the system when subjected to harmonic inputs of given frequency and varying amplitude.

4. What value \( e_0 \) of a step input \( e(t) = e_0 u(t) \) will throw the system into instability?

![Diagram](image)

(b) 

Fig. P57.

58. Analysis of a Particular Type of Nonlinear Servo. 1. Consider an ideal on-off servo system (input \( u \), output \( v \)) in which the output rate is assumed to be exactly \( M \) times \( \text{sign} (u - v) \). By making use of the first-harmonic approximation, derive the generalized transfer function \( V/U \) as a function of the amplitude \( u_1 \) and angular frequency \( \omega \). Show that this function can be represented by one locus \( \Gamma \) scaled in the product \( u_1 \omega \). Plot that locus \( \Gamma \).

2. The servo just studied is assumed to be the servomotor stage of the longitudinal control system of an airplane. In the forward path of the whole system it is preceded by the compensating network and amplifier stage, with a transfer function \( F(s) \), and followed by a black box representing the dynamics of the airplane, with a transfer function \( G(s) \). Show that the limit cycle(s) of the system can be obtained as intersection(s) of the \( F(j\omega)G(j\omega) \) locus with a locus \( C \) which can be easily derived from the locus \( \Gamma \) obtained in the first part. Plot the locus \( C \) in Nichols coordinates.

3. Assume there exists one limit cycle and that it is unstable. What conclusions can be drawn concerning the stability of the over-all system from the Nichols plot of \( F(j\omega)G(j\omega) \)?

4. Draw for the control system described in (2) an equivalent block diagram of the type of Fig. 22-30. Give the expression for \( L(s) \) in terms of \( F(s) \) and \( G(s) \). How can the possible limit cycle(s) of the system be derived from the Nyquist plots of \( F(j\omega) \times G(j\omega) \) by making use of the first-harmonic approximation? Explain why, when there is one limit
cycle, it is in general unstable. Can the over-all system be stable in the presence of any disturbance?

59. Basic Design of an Electric Servomotor. It is desired to design the control-surface servomotor for an autopilot in a fast aircraft.

a. Characteristics of the autopilot. The natural frequency of the stabilized aircraft is 0.8 cps. Velocity piloting (see Sec. 14.3) is implied without an internal loop around the motor. The motor specifications must hold at a frequency \( f_o \) equal to 5 times the natural frequency of the aircraft. It is desired that a command in angular velocity be followed by the control surface with a phase lag which should not exceed 10° at the frequency \( f_o \). A separately excited d-c motor with armature control is to be used. The power source in the airplane is 27 volts d-c.

b. Characteristics of the control surface. Inertia, 20 kg-cm²; aero-dynamic torque, 20 m-kg/rad; and maximum deflection, 6°.

c. General observation. The data should be considered in such a way as to determine which requirements are necessary, sufficient, or even compatible. The basic design of the motor can be done only by successive approximations. First a certain number of parameters necessary for the preliminary design are estimated. The equations should be solved with these estimated parameters in order to check their validity. In the present problem only one iteration should be made, but if another iteration appears to be necessary, the parameters to be reestimated should be indicated as well as the parameter modification itself. In each case, a check should be made in order to see if all the required specifications are met and the assumptions satisfied.

1. Determination of the motor. Comment on the choice of an armature-control motor. Derive the literal expression for the transfer function of the motor (neglect inductance in first approximation). Using this transfer function, comment on the specifications relative to the control-surface deflection rate. Draw, in the torque/speed plane, the characteristic curve of the motor. For this, estimate a value of the inertia of the motor and apply the harmonic method (Sec. 30.5.2; use the graphical approach) and the modulus method (Sec. 30.5.3) successively. Explain how mechanical resonance of the control surface is introduced into the transfer function and into the determination of the characteristic curve of the motor. Compare the harmonic and the modulus methods. Give values for the motor power, the gear ratio, and a numerical expression of the transfer function that meets the required specifications.

2. Determination of the size and electrical characteristics of the motor. Determine the specifications of the motor armature; this should include length, diameter, number and cross section of conductors, number of slots, brush dimensions, etc.

Hints. a. A bipolar motor can be chosen with a ratio

\[
\frac{\text{pole arc}}{\text{pole pitch}} \approx 0.7
\]

This ratio indicates the quantitative relation between active conductors and total conductors on the armature.
b. Correction for leakage flux can be made by assuming the armature flux to be 80 per cent of the air-gap flux. The latter generally corresponds to a field of 3,000 to 6,000 gauss. Take 5,000 gauss as a first approximation. The flux density should not exceed approximately 15,000 gauss in the poles of the stator, 12,000 gauss in the stator yoke, and 2,500 to 8,000 gauss in the armature itself.

c. Use the electromagnetic notations and relations of Sec. 25.1.4 and justify these relations. For the characteristic curve take as reference point the one corresponding to maximum power. As a first approximation, assume a 50 per cent efficiency. Limit the current density to about 7 to 10 amp/mm². Keep the number of ampere-conductors (product of the number of armature conductors by the conductor current per centimeter of armature circumference) roughly between 100 and 500. As a first approximation, the following numbers of slots could be taken.

<table>
<thead>
<tr>
<th>Slots</th>
<th>Armature diam (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-8</td>
<td>2-3</td>
</tr>
<tr>
<td>9-12</td>
<td>3-4</td>
</tr>
<tr>
<td>12-24</td>
<td>4-6</td>
</tr>
</tbody>
</table>

The resistivity of copper is $1.6 \times 10^{-6}$ ohm-cm.

d. For the commutator, a current density of 10 to 25 amp/cm² should be assumed. The brush pressure should be of the order of 150 g/cm²; the brush voltage drop should be assumed equal to about 5 per cent of the applied voltage.

e. Losses by eddy current and hysteresis. Typical values in industry for these losses are 2.6, 2.2, 1.6, and 0.9 watts/kg at 10,000 gauss and 50 cps.

f. Mechanical losses. Brush friction involves a loss equal to

$$P_f = f_b P_b V_o S \times 10^{-4} \quad \text{watts}$$

where $f_b = \text{friction coefficient between brushes and commutator (take } f_b = 0.1)$

$S = \text{brush cross section, cm}^2$

$P_b = \text{brush pressure, g/cm}^2$

$V_o = \text{tangential velocity of commutator surface, cm/sec}$

Rolling friction loss is

$$P_f = F V_o f_a \times 10^{-4} \quad \text{watts}$$

where $F = \text{weight of armature on each bearing}$

$f_a = \text{friction coefficient of bearing (0.0015 for ball bearings, 0.08 for sleeve bearings)}$

$V_a = \text{tangential velocity of shaft surface, cm/sec}$

Windage losses should be neglected in the first approximation. All these losses can be added in order to determine the efficiency of the armature at the operating point chosen, and this efficiency can then be compared with the value originally assumed.
g. Dimensions of field circuit. Determine the geometric and magnetic characteristics of the field circuit. A check should be made to see if the flux density in different parts of the stator is as originally assumed.

60. Determination of a Hydraulic Motor. An airplane control surface is to be controlled under the following conditions: (a) maximum deflection of the control surface, 0.1 rad; (b) maximum disturbing torque on the control surface, 4 m-csn; (c) maximum control-surface deflection rate, 1 rad/sec; (d) the system should be a position servo; (e) under sinusoidal excitation, the maximum phase lag should not exceed 0.1 rad at 5 cps; (f) the maximum deflection-rate variation between full-load and no-load conditions should not exceed 10 per cent; (g) coefficient of compression of oil, $5 \times 10^{-5}$ cm/csn (1 csn is $\frac{1}{1000}$ sthene, see Prob. 39); (h) inertia of the control surface, 0.01 kg-m²; and (i) all dimensions of the whole servo system should be less than 10 cm.

1. Let the valve be of the form shown in Fig. P60. The liquid goes from the high pressure (100 hectopieze—1 hectopieze, or kg/cm², is approximately 14.5 psi) to the low pressure through the contractions 1 and 2 on one side and 3 and 4 on the other side of the spool. The hydraulic pressure is carried to the hydraulic motor through holes A and B. The load impedance connected between A and B is represented in an electrical analogy by a resistance. This impedance will be assumed small compared to the bypass impedances 2, 3, and 4. The variations of admittances of 1, 2, 3, and 4 are proportional to the spool displacements in the neighborhood of the position $y = 0$ (the resistances 1, 2, 3, and 4 are equal). Show that between A and B there is a flow proportional to the displacement $y$ of the core when no load is applied to the motor.

2. Make a free-hand sketch of the adopted system indicating the dimensions (use a linear piston-type motor).

3. Assume, as in (1), that the valve output flow is $Q = ky$, where $y$ is the spool displacement. Calculate the transfer function of the system. The inertias of the piston and rods, roughly estimated, should be taken into account. Calculate the value of the gains involved. Are any conditions imposed on the value of the natural frequency of the valve?
FIVE-LANGUAGE GLOSSARY OF IMPORTANT AUTOMATIC-CONTROL TERMS

The following five-language glossary (English, German, Spanish, French, and Russian) is intended to help non-English-speaking readers understand the terms of the present book, and to aid English-speaking readers of foreign publications on automatic control. In order to facilitate the manipulation of this glossary, the words have been grouped by topics, an alphabetical index making it possible for anyone to find the four equivalent words for any term desired.

The terminology in the automatic-control field is not unified, and the terms used often vary considerably from one author to another within the same language. Therefore, the present authors have restricted themselves to the terms most commonly used. Equivalent terms used by different writers are separated by a comma; a semicolon instead of a comma between the two terms indicates that they are not exactly equivalent. Terms enclosed in brackets are added for the sake of clarity, e.g., "transfer function [of a linear system]." Terms enclosed in parentheses are those not most commonly used, e.g., "transfer function (performance operator)."

Three complementary notes have been placed at the end of the glossary; they are devoted to certain questions of German and Russian terminology that require special explanations.

In order to write Russian words with Latin letters the use of transliteration is necessary. The transliteration used here is a compromise between that recommended by philologists and the popular transliteration used in England and in the United States. Most letters are pronounced approximately as in English:

a as in park      ch as in church       y as i in mill
e as in bed      o as in lot             sh as in shut
i as in pit      u as in put

However:

a. The letter j corresponds to the English y in yes, yolk, yard; zh should be pronounced like s in treasure, pleasure; kh should be pronounced like ch in German dach, loch.

b. The apostrophe indicates a "softening" of the preceding consonant, i.e., the simultaneous sounding of an i with the consonant.

c. The letters e and i actually represent je and ji, respectively, and automatically imply softening of the preceding consonant, except if the latter is zh or sh. Thus net stands for njet and is pronounced nyet, etc.

d. The vowel that bears the tonic accent (udarenie) is written in boldface.

1 The present glossary is restricted to terms used in the automatic-control field. For other technical terms the reader should refer to a technical dictionary. In this connection we especially recommend P. Naailin, "Three-language Technical Vocabulary," Revue d'Optique, Paris, 1951, for French, German, and English terms.

2 For German terms and special features of German concepts and terminology, the authors acknowledge the important help of B. Wolf.

3 The Spanish vocabulary is by Dr. G. Fucha, of the Universidad Nacional de Cordoba, Argentina. We are grateful to G. Klopfeinstein for his adaptation of the Spanish vocabulary to the rest of the glossary.

4 Except in some foreign words, e.g., rels, element, etc.
### Automatic control (general terms)

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 control system</td>
<td>sistema de mando</td>
<td>le système de commande</td>
</tr>
<tr>
<td>2 automatic control</td>
<td>control automático,</td>
<td>l'automatisme, les</td>
</tr>
<tr>
<td>3 servo system, servomechanism</td>
<td>automatismo</td>
<td>asservissements</td>
</tr>
<tr>
<td>4 regulator</td>
<td>regulador</td>
<td>l'asservissement, le système</td>
</tr>
<tr>
<td>5 self-adjusting system</td>
<td>regulación automática, sistema</td>
<td>le système d'auto-régulation</td>
</tr>
<tr>
<td>6 block diagram</td>
<td>diagrama funcional</td>
<td>le diagramme fonctionnel</td>
</tr>
<tr>
<td>7 without [or with] power</td>
<td>sin [o con] amplificación</td>
<td>sans [ou avec] amplification de</td>
</tr>
<tr>
<td>8 cascade system</td>
<td>sistema en cadena</td>
<td>puissance</td>
</tr>
<tr>
<td>9 feedback</td>
<td>el retorno</td>
<td>le système en chaine</td>
</tr>
<tr>
<td>10 loop</td>
<td>el lastro, el bucle</td>
<td>le boucle</td>
</tr>
<tr>
<td>11 feedback system, closed-loop</td>
<td>sistema con retorno, en bucle</td>
<td>le système à retour, le système</td>
</tr>
<tr>
<td>12 input</td>
<td>cerrado</td>
<td>bouclé</td>
</tr>
<tr>
<td>13 output</td>
<td>la entrada</td>
<td>l'entrée</td>
</tr>
<tr>
<td>14 control input, command</td>
<td>la salida</td>
<td>l'entrée</td>
</tr>
<tr>
<td>15 disturbance</td>
<td>perturbación</td>
<td>la perturbation</td>
</tr>
<tr>
<td>16 error (deviation)</td>
<td>error, desviación</td>
<td>l'erreur, l'écart</td>
</tr>
<tr>
<td>17 system to be controlled</td>
<td>el sistema a controlar;</td>
<td>le système à asservir; le système</td>
</tr>
<tr>
<td>18 follow-up system</td>
<td>(sistema seguidor)</td>
<td>À régular</td>
</tr>
<tr>
<td>19 level regulator</td>
<td>regulador de nivel</td>
<td>le système suivre</td>
</tr>
<tr>
<td>20 autopilot</td>
<td>piloto automático</td>
<td>le régulateur de niveau</td>
</tr>
<tr>
<td>21 guided missile, controlled</td>
<td>proyectil dirigido</td>
<td>le pilote automatique</td>
</tr>
<tr>
<td>22 design</td>
<td>proyecto</td>
<td>l'engin guidé</td>
</tr>
<tr>
<td>23 specifications</td>
<td>especificaciones</td>
<td>le projet</td>
</tr>
<tr>
<td>24 component</td>
<td>elemento, componente</td>
<td>les clauses techniques</td>
</tr>
<tr>
<td>25 accuracy</td>
<td>precisión</td>
<td>l'organe</td>
</tr>
<tr>
<td>26 velocity</td>
<td>velocidad</td>
<td>la vitesse</td>
</tr>
<tr>
<td>27 angular velocity</td>
<td>velocidad angular</td>
<td>la vitesse angulaire</td>
</tr>
<tr>
<td>28 acceleration</td>
<td>aceleración</td>
<td>l'accélération</td>
</tr>
<tr>
<td>29 force</td>
<td>fuerza</td>
<td>la force</td>
</tr>
<tr>
<td>30 torque, couple</td>
<td>par</td>
<td>le couple</td>
</tr>
<tr>
<td>31 moment of inertia</td>
<td>momento de inercia</td>
<td>le moment d'inertie</td>
</tr>
<tr>
<td>32 mass</td>
<td>la masse</td>
<td>la masse</td>
</tr>
<tr>
<td>33 spring</td>
<td>resorte</td>
<td>le ressort</td>
</tr>
<tr>
<td>34 stiffness</td>
<td>rigides</td>
<td>la raideur</td>
</tr>
<tr>
<td>35 viscous friction</td>
<td>rodamiento viscoso</td>
<td>le frottement visqueux</td>
</tr>
<tr>
<td>36 forced state</td>
<td>regimen forzado</td>
<td>le régime forcé</td>
</tr>
<tr>
<td>37 harmonic function of time</td>
<td>función armónica del tiempo</td>
<td>la fonction harmonique du temps</td>
</tr>
<tr>
<td>38 frequency</td>
<td>frecuencia</td>
<td>la fréquence</td>
</tr>
<tr>
<td>39 angular frequency</td>
<td>frecuencia angular</td>
<td>la pulsación, (la fréquence</td>
</tr>
<tr>
<td>40 phase angle</td>
<td>angulo de fase</td>
<td>angulaire)</td>
</tr>
<tr>
<td>41 transient</td>
<td>régimen transitorio</td>
<td>le régime transitoire</td>
</tr>
<tr>
<td>42 unit-step response</td>
<td>respuesta a la función unitaria</td>
<td>la réponse unitaire</td>
</tr>
<tr>
<td>43 dimensionless [reduced]</td>
<td>variable adimensional, [reducida]</td>
<td>la grandeur réduite [a</td>
</tr>
<tr>
<td>44 direct current</td>
<td>corriente continua</td>
<td>dimension]</td>
</tr>
<tr>
<td>45 alternating current</td>
<td>corriente alterna</td>
<td>le courant continu</td>
</tr>
<tr>
<td>46 voltage</td>
<td>tensión</td>
<td>le courant alternatif</td>
</tr>
<tr>
<td>47 current</td>
<td>corriente (intensidad)</td>
<td>la tension</td>
</tr>
<tr>
<td>48 circuit; network</td>
<td>circuito; red</td>
<td>le courant</td>
</tr>
<tr>
<td>49 resistance</td>
<td>resistencia</td>
<td>le circuit; le réseau</td>
</tr>
<tr>
<td>50 inductance</td>
<td>inductancia</td>
<td>la résistance</td>
</tr>
<tr>
<td>51 capacitance</td>
<td>capacidad (capacitancia es 1/Cu)</td>
<td>l'inductance</td>
</tr>
<tr>
<td>52 quality factor</td>
<td>factor de calidad</td>
<td>la capacité</td>
</tr>
</tbody>
</table>

### Fundamentals of mechanics

### Fundamentals of electricity
1. Automatic control (general terms)

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>das Steuersystem</td>
<td>sistema upravlenija</td>
</tr>
<tr>
<td>die selbsttätige Regelung</td>
<td>avtomaticheskoe upravljenie; avtomaticheskoe regulirovanie</td>
</tr>
<tr>
<td>die Folgeregulierung, das Regelsystem</td>
<td>sistema avtomaticheskovo upravlenija, upravljajushchee ustroistvo</td>
</tr>
<tr>
<td>der Regler</td>
<td>sistema avtomaticheskovo regulirovanija, reguljator</td>
</tr>
<tr>
<td>das selbstregelnde System</td>
<td>samonстраивающееся sistema</td>
</tr>
<tr>
<td>das Blockschaltbild [1]</td>
<td>blok-skhema; strukturajace skhema [1.3]</td>
</tr>
<tr>
<td>ohne (oder mit) Hilfenergie</td>
<td>[ne] prinovno dejstvja</td>
</tr>
<tr>
<td>die Kettenschaltung</td>
<td>kaskadnaja sistema</td>
</tr>
<tr>
<td>die Rückkopplung; (for internal loops only) die Rückführung [2]</td>
<td>obratnaja svjas'</td>
</tr>
<tr>
<td>der Regelkreis</td>
<td>petlja</td>
</tr>
<tr>
<td>das rückgekoppelte System</td>
<td>sistema s obratnoj svjas'ju, svjazannaja sistema</td>
</tr>
<tr>
<td>die Eingangsgröße</td>
<td>vkhodnaja velichina</td>
</tr>
<tr>
<td>die Ausgangsgröße</td>
<td>vykhodnaja velichina</td>
</tr>
<tr>
<td>die Führungsgroesse</td>
<td>upravljajushchee vosdejstvie, zadajushchee vosdejstvie</td>
</tr>
<tr>
<td>die Störgröße</td>
<td>vormushajushchee vosdejstvie</td>
</tr>
<tr>
<td>die Regelabweichung [1.2]</td>
<td>osibka (otklesenie)</td>
</tr>
<tr>
<td>die Regelstrecke</td>
<td>upravljajusjemyj objekt; reguliruemuyj objekt</td>
</tr>
<tr>
<td>der Nachlaufregler</td>
<td>sledjajuschaja sistema</td>
</tr>
<tr>
<td>der Flächensteuerregler</td>
<td>reguljator urovnya</td>
</tr>
<tr>
<td>der Flugregler</td>
<td>avtopilot</td>
</tr>
<tr>
<td>die Lenkwaage, der gelenkte Flugkörpere</td>
<td>upravljajusjemyj snarjad</td>
</tr>
<tr>
<td>der Entwurf</td>
<td>proektirovanie</td>
</tr>
<tr>
<td>die technischen Bedingungen</td>
<td>tekhnicheskie trebovanija</td>
</tr>
<tr>
<td>das Regelkreisglied; der Bauteil</td>
<td>element</td>
</tr>
<tr>
<td>die Genauigkeit</td>
<td>tochnost'</td>
</tr>
</tbody>
</table>

2. Fundamentals of mechanics

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>die Geschwindigkeit</td>
<td>skorost'</td>
</tr>
<tr>
<td>die Winkelgeschwindigkeit</td>
<td>uglovaja skorost'</td>
</tr>
<tr>
<td>die Beschleunigung</td>
<td>uskorenie</td>
</tr>
<tr>
<td>die Kraft</td>
<td>sila</td>
</tr>
<tr>
<td>das Moment</td>
<td>moment</td>
</tr>
<tr>
<td>das Trägheitsmoment</td>
<td>moment inertial</td>
</tr>
<tr>
<td>die Masse</td>
<td>massa</td>
</tr>
<tr>
<td>die Sprungfeder</td>
<td>pruzhina</td>
</tr>
<tr>
<td>die Federkonstante</td>
<td>shokost'</td>
</tr>
<tr>
<td>die fliessige Raibung</td>
<td>vjasko renie</td>
</tr>
<tr>
<td>der eingegewogene Zustand</td>
<td>vynushshennyj reshim</td>
</tr>
<tr>
<td>von der Zeit abhängige Sinusfunktion</td>
<td>garmonicheskaja funktsija vremen</td>
</tr>
<tr>
<td>die Frequenz</td>
<td>chastota</td>
</tr>
<tr>
<td>die Kreisfrequenz</td>
<td>uglovaja chastota</td>
</tr>
<tr>
<td>der Phasenwinkel</td>
<td>edvig fazy</td>
</tr>
<tr>
<td>der Einhand viewModel</td>
<td>perekhodnyj reshim</td>
</tr>
<tr>
<td>die Sprung-antwoord,</td>
<td>perekhodnya funktsija</td>
</tr>
<tr>
<td>die Übergangsfunction</td>
<td></td>
</tr>
<tr>
<td>die dimensionless [besogen] Größe</td>
<td>besrasmernaja velichina</td>
</tr>
</tbody>
</table>

3. Fundamentals of electricity

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>der Gleichstrom</td>
<td>postolannyj tok</td>
</tr>
<tr>
<td>der Wechselstrom</td>
<td>peremennyj tok</td>
</tr>
<tr>
<td>die Spannung</td>
<td>naprjashenie</td>
</tr>
<tr>
<td>der Strom</td>
<td>potok</td>
</tr>
<tr>
<td>der Stromkreis; das Netwerk</td>
<td>set'</td>
</tr>
<tr>
<td>der Widerstand</td>
<td>soprotivlenie</td>
</tr>
<tr>
<td>die Induktivität</td>
<td>induktivnost'</td>
</tr>
<tr>
<td>die Kapazität</td>
<td>lomkost'</td>
</tr>
<tr>
<td>der Gültfaktor</td>
<td>dobrotnost'</td>
</tr>
</tbody>
</table>

1 Numbers in brackets refer to the Complimentary Notes.
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>charge</td>
<td>carga</td>
<td>la charge</td>
</tr>
<tr>
<td>impedance</td>
<td>impedancia</td>
<td>l'impédance</td>
</tr>
<tr>
<td>magnetic flux</td>
<td>flujo magnético</td>
<td>le flux magnétique</td>
</tr>
<tr>
<td>electromotive force (emf)</td>
<td>fuerza electromotriz</td>
<td>la force électromotrice (f.a.m.)</td>
</tr>
<tr>
<td>direct-current generator</td>
<td>generador de corriente continua</td>
<td>le générateur de courant continu</td>
</tr>
<tr>
<td>power</td>
<td>potencia</td>
<td>la puissance</td>
</tr>
<tr>
<td>efficiency</td>
<td>rendimiento</td>
<td>le rendement</td>
</tr>
</tbody>
</table>

4. Fundamentals of mathematics

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>variable; function</td>
<td>variable; función</td>
<td>la variable; la función</td>
</tr>
<tr>
<td>derivative with respect to time</td>
<td>función derivada con relación al tiempo</td>
<td>la dérivée par rapport au temps</td>
</tr>
<tr>
<td>differential equation</td>
<td>ecuación diferencial</td>
<td>l'équation différentielle</td>
</tr>
<tr>
<td>(non) linear</td>
<td>(no) lineal</td>
<td>(non) linéaire</td>
</tr>
<tr>
<td>with constant coefficients</td>
<td>con coeficientes constantes</td>
<td>à coefficients constants</td>
</tr>
<tr>
<td>with variable coefficients</td>
<td>con coeficientes variables</td>
<td>à coefficients variables</td>
</tr>
<tr>
<td>initial conditions</td>
<td>condiciones iniciales</td>
<td>les conditions initiales</td>
</tr>
<tr>
<td>discontinuity</td>
<td>discontinuidad</td>
<td>la discontinuité</td>
</tr>
<tr>
<td>limit</td>
<td>límite</td>
<td>la limite</td>
</tr>
<tr>
<td>tend toward infinity</td>
<td>tirar hacia el infinito</td>
<td>tendre vers l'infini</td>
</tr>
<tr>
<td>approximate method</td>
<td>método aproximado</td>
<td>la méthode approchée</td>
</tr>
<tr>
<td>simplified</td>
<td>simplificado</td>
<td>simplifié</td>
</tr>
<tr>
<td>rational fraction</td>
<td>fracción algebraica</td>
<td>la fraction rationnelle</td>
</tr>
<tr>
<td>numerator [of a fraction]</td>
<td>numerador</td>
<td>numérateur [d'une fraction]</td>
</tr>
<tr>
<td>denominator</td>
<td>denominador</td>
<td>dénominateur</td>
</tr>
<tr>
<td>partial-fraction expansion</td>
<td>descomposición en elementos simples</td>
<td>la décomposition en éléments simples</td>
</tr>
<tr>
<td>periodic solution</td>
<td>solución periódica</td>
<td>la solución periódica</td>
</tr>
<tr>
<td>unit step</td>
<td>salto unitario</td>
<td>l'échelon unitaire</td>
</tr>
<tr>
<td>unit impulse</td>
<td>impulso unitario</td>
<td>l'impulsión unitario, la función de Dirac</td>
</tr>
</tbody>
</table>

5. Laplace transforms

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace transformation</td>
<td>transformación de Laplace</td>
<td>la transformation de Laplace</td>
</tr>
<tr>
<td>Laplace transform</td>
<td>Laplaciana (transformada de Laplace)</td>
<td>la transformée de Laplace (l'image)</td>
</tr>
<tr>
<td>inverse transform</td>
<td>transformada inversa</td>
<td>la transformée inverse</td>
</tr>
<tr>
<td>Fourier expansion</td>
<td>desarrollo en serie de Fourier</td>
<td>le développement en série de Fourier</td>
</tr>
<tr>
<td>lag theorem</td>
<td>teorema del retardo</td>
<td>le théorème du retard</td>
</tr>
<tr>
<td>initial- and final-value theorems</td>
<td>teorema del valor inicial, final</td>
<td>les théorèmes de la valeur initiale et finale</td>
</tr>
</tbody>
</table>

6. Stability

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>equilibrium</td>
<td>equilibrio</td>
<td>l'équilibre</td>
</tr>
<tr>
<td>characteristic equation</td>
<td>ecuación característica</td>
<td>l'équation caractéristique</td>
</tr>
<tr>
<td>damping</td>
<td>amortiguamiento</td>
<td>l'amortissement</td>
</tr>
<tr>
<td>stable; stability</td>
<td>estable; estabilidad</td>
<td>stable; la stabilité</td>
</tr>
<tr>
<td>unstable; instability</td>
<td>inestable; inestabilidad</td>
<td>instable; l'instabilité</td>
</tr>
<tr>
<td>open-loop stable</td>
<td>estable en bucle abierto</td>
<td>estable en boucle ouverte</td>
</tr>
<tr>
<td>left half plane</td>
<td>semi plano izquierdo</td>
<td>le demi-plan de gauche</td>
</tr>
<tr>
<td>absolute stability margin</td>
<td>margen absoluto de estabilidad</td>
<td>la marge absolue de stabilité</td>
</tr>
<tr>
<td>Hurwitz's criterion</td>
<td>criterio de Hurwitz</td>
<td>le critère de Hurwitz</td>
</tr>
<tr>
<td>Routh's criterion</td>
<td>criterio de Routh</td>
<td>le critère de Routh</td>
</tr>
<tr>
<td>absolute stability margin</td>
<td>criterio de Nyquist-Cauchy</td>
<td>le critère de Nyquist-Cauchy</td>
</tr>
<tr>
<td>Nyquist's criterion</td>
<td>criterio de Nyquist-Cauchy</td>
<td>le critère du revers</td>
</tr>
<tr>
<td>left-hand criterion</td>
<td>criterio de mano izquierda</td>
<td>le domaine de stabilité</td>
</tr>
<tr>
<td>stability domain</td>
<td>región de estabilidad</td>
<td>le pompage</td>
</tr>
</tbody>
</table>
### 4. Fundamentals of mathematics

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>die Veränderliche; die Funktion</td>
<td>peremennaja; funktsija</td>
<td>60</td>
</tr>
<tr>
<td>die Ableitung nach der Zeit</td>
<td>proizvodnaja po vremeni</td>
<td>61</td>
</tr>
<tr>
<td>die Differenzialgleichung</td>
<td>differentsial’noe uravnenie</td>
<td>62</td>
</tr>
<tr>
<td>[nicht] linear</td>
<td>[ne] linejnyj</td>
<td>63</td>
</tr>
<tr>
<td>mit konstanten Koeffizienten</td>
<td>s postojannymi koeffisientami</td>
<td>64</td>
</tr>
<tr>
<td>mit veränderlichen Koeffizienten</td>
<td>s peremennymi koeffisientami</td>
<td>65</td>
</tr>
<tr>
<td>die Anfangsbedingungen</td>
<td>nachal’nye uslovija</td>
<td>66</td>
</tr>
<tr>
<td>die Unstetigkeit</td>
<td>rasryv</td>
<td>67</td>
</tr>
<tr>
<td>der Limes</td>
<td>predel</td>
<td>68</td>
</tr>
<tr>
<td>gegen Unendlich gehen</td>
<td>stremitsa k beskonechnosti</td>
<td>69</td>
</tr>
<tr>
<td>die Annäherungsmethode</td>
<td>prishljannyj sposob</td>
<td>70</td>
</tr>
<tr>
<td>vereinfacht</td>
<td>uproshchonnyi</td>
<td>71</td>
</tr>
<tr>
<td>der rationale Bruch</td>
<td>drobno-racional’naja funktsija</td>
<td>72</td>
</tr>
<tr>
<td>der Zähler [eines Bruches]</td>
<td>chislitel’ [drobi]</td>
<td>73</td>
</tr>
<tr>
<td>der Nenner</td>
<td>znamenitel’</td>
<td>74</td>
</tr>
<tr>
<td>die Partialbruchserlegung</td>
<td>rasloshenie na prostye drobi</td>
<td>75</td>
</tr>
<tr>
<td>die periodische Lösung</td>
<td>periodicheskoe reshenie</td>
<td>76</td>
</tr>
<tr>
<td>der Einheitsprung</td>
<td>edinichnyj skachok</td>
<td>77</td>
</tr>
<tr>
<td>die Einheitsimpulse-funktion, die Nadelfunktion</td>
<td>78 funktsija, edinichnyj impuls’a, impul’saja funktsija Diraka</td>
<td></td>
</tr>
</tbody>
</table>

### 5. Laplace transforms

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>die Laplace-transformation</td>
<td>preobrazovanie Laplassa</td>
<td>79</td>
</tr>
<tr>
<td>die Bildfunktion, die Unterfunktion</td>
<td>preobrashenie Laplassa</td>
<td>80</td>
</tr>
<tr>
<td>die Rücktransformation</td>
<td>obratnoe preobrashenie</td>
<td>81</td>
</tr>
<tr>
<td>die Fourier serlegung</td>
<td>rasloshenie v rjad Furje</td>
<td>82</td>
</tr>
<tr>
<td>der Verschiebungsatz</td>
<td>teorema zapasyvaniya</td>
<td>83</td>
</tr>
<tr>
<td>der Anfangswertzaits, der Endwertzaits</td>
<td>teoremy o predel’nykh znacheniyakh</td>
<td>84</td>
</tr>
<tr>
<td>der Faltungsaits</td>
<td>teorema svjortyvaniya</td>
<td>85</td>
</tr>
<tr>
<td>der Originalbereich, der Oberbereich</td>
<td>p-lozkoسط’</td>
<td>86</td>
</tr>
<tr>
<td>die p-Ebene, der Unterbereich</td>
<td>dejstviel’naja chast’</td>
<td>87</td>
</tr>
<tr>
<td>der Realteil</td>
<td>mnimaja chast’</td>
<td>88</td>
</tr>
<tr>
<td>der imaginäre Teil</td>
<td>radius-vektor; argument</td>
<td>89</td>
</tr>
<tr>
<td>der Radiusvektor; das Argument</td>
<td>radius-vektor; argument</td>
<td>90</td>
</tr>
<tr>
<td>die imaginäre Achse</td>
<td>mnimaja os’</td>
<td>91</td>
</tr>
<tr>
<td>die geschlossene Kurve</td>
<td>samknyutjy kontur</td>
<td>92</td>
</tr>
<tr>
<td>der Residuenaits</td>
<td>teorija vychotov</td>
<td>93</td>
</tr>
<tr>
<td>konjugierte Wurseln</td>
<td>sopriaboanye korin</td>
<td>94</td>
</tr>
<tr>
<td>die mehrfache Wursel</td>
<td>kratnyj koren</td>
<td>95</td>
</tr>
<tr>
<td>die Vielfachigkeit</td>
<td>kratnost’</td>
<td></td>
</tr>
</tbody>
</table>

### 6. Stability

<table>
<thead>
<tr>
<th>German</th>
<th>Russian</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>das Gleichgewicht</td>
<td>ravnovesie</td>
<td>97</td>
</tr>
<tr>
<td>die charakteristische Gleichung</td>
<td>kharakteristicheskoe uravnenie</td>
<td>98</td>
</tr>
<tr>
<td>die Dämpfung</td>
<td>satschanie</td>
<td>99</td>
</tr>
<tr>
<td>stabil; die Stabilität</td>
<td>ustojobivyi; ustojobivost’ [3]</td>
<td>100</td>
</tr>
<tr>
<td>stabil, instabil; die Labilität</td>
<td>neustojobivyi; neustojobivost’</td>
<td>101</td>
</tr>
<tr>
<td>stabil bei offenen Regelkreis</td>
<td>ustojobivyi v rasomkatom sostojanii</td>
<td>102</td>
</tr>
<tr>
<td>die line Halbebene</td>
<td>levaja poluploskoسط’</td>
<td>103</td>
</tr>
<tr>
<td>der absolute Stabilitätstand der Wurseln</td>
<td>stepen’ ustojobivost’</td>
<td>104</td>
</tr>
<tr>
<td>das Hurwit’s-Kriterium</td>
<td>kriterii Rauta-Gurvitsa</td>
<td>105</td>
</tr>
<tr>
<td>das Routh-Kriterium</td>
<td>kriterii Mikhallova</td>
<td>106</td>
</tr>
<tr>
<td>die Nyquist-Bedingung</td>
<td>pravilo akhirivkov</td>
<td>108</td>
</tr>
<tr>
<td>(das line-Hand Kriterium)</td>
<td>oblast’ ustojobivost’</td>
<td>109</td>
</tr>
<tr>
<td>der Stabilitätsbereich</td>
<td></td>
<td>110</td>
</tr>
<tr>
<td>das Pendeln</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 7. Elementary linear systems

<table>
<thead>
<tr>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>111. first-order system</td>
<td>sistema de primer orden</td>
</tr>
<tr>
<td>112. first-order factor ([1/(1 + Ts)])</td>
<td>factor de primer orden</td>
</tr>
<tr>
<td>113. time constant</td>
<td>constante de tiempo</td>
</tr>
<tr>
<td>114. responses time to within 5%</td>
<td>tiempo de respuesta a 5%</td>
</tr>
<tr>
<td>115. lag; delay</td>
<td>retraso, retraso</td>
</tr>
<tr>
<td>116. second-order system</td>
<td>sistema de segundo orden</td>
</tr>
<tr>
<td>117. quadratic factor</td>
<td>factor cuadrático</td>
</tr>
<tr>
<td>118. natural frequency</td>
<td>frecuencia propia</td>
</tr>
<tr>
<td>119. resonant frequency</td>
<td>frecuencia de resonancia</td>
</tr>
<tr>
<td>120. damped oscillation</td>
<td>oscilación amortiguada</td>
</tr>
<tr>
<td>121. decrement</td>
<td>decremento (razón)</td>
</tr>
<tr>
<td>122. damping ratio ([\zeta, \text{DR}])</td>
<td>factor de amortiguamiento</td>
</tr>
<tr>
<td>123. overshoot</td>
<td>exceso de elongación</td>
</tr>
<tr>
<td>124. one-degree-of-freedom system</td>
<td>sistema a un grado de libertad</td>
</tr>
<tr>
<td>125. many-degree-of-freedom system</td>
<td>sistema a varios grados de libertad</td>
</tr>
</tbody>
</table>

## 8. Linear systems

<table>
<thead>
<tr>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>126. principle of superposition</td>
<td>principio de superposición</td>
</tr>
<tr>
<td>127. lumped-constant [system]</td>
<td>circuitos a constantes localizadas</td>
</tr>
<tr>
<td>128. with distributed parameters</td>
<td>circuitos a constantes repartidas</td>
</tr>
<tr>
<td>129. behavior for any input</td>
<td>respuesta a una entrada cualquiera</td>
</tr>
<tr>
<td>130. transfer function (performance operator)</td>
<td>función de transferencia</td>
</tr>
<tr>
<td>131. gain</td>
<td>ganancia</td>
</tr>
<tr>
<td>132. pole-zero configuration</td>
<td>configuración de los polos y de los ceros</td>
</tr>
<tr>
<td>133. mode</td>
<td>modo</td>
</tr>
<tr>
<td>134. transfer locus, Nyquist plot</td>
<td>lugar de transferencia</td>
</tr>
<tr>
<td>135. amplitude response</td>
<td>respuesta en amplitud</td>
</tr>
<tr>
<td>136. phase response</td>
<td>respuesta en fase</td>
</tr>
<tr>
<td>137. Nichols locus</td>
<td>la curva de Black</td>
</tr>
<tr>
<td>138. resonance ratio ([Q]), peak value of amplification ([M_p, \text{or } M_\infty])</td>
<td>factor de resonancia</td>
</tr>
<tr>
<td>139. passband</td>
<td>pasabanda</td>
</tr>
<tr>
<td>140. slope of 6 dB per octave</td>
<td>pendiente de (-6) db por octava</td>
</tr>
<tr>
<td>141. minimum-phase-shift system</td>
<td>sistemas de desfasaje mínimo</td>
</tr>
<tr>
<td>142. all-pass element ([(1 – ae) / (1 + ae)])</td>
<td>integración (1/z)</td>
</tr>
<tr>
<td>143. integration ([1/s])</td>
<td>integración (1/s)</td>
</tr>
<tr>
<td>144. phase lead ([1 + Ts])</td>
<td>avance de fase</td>
</tr>
<tr>
<td>145. quadrupole</td>
<td>cuadripolo</td>
</tr>
</tbody>
</table>

## 9. Linear servo systems

<table>
<thead>
<tr>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>146. closed-loop transfer function (system function)</td>
<td>función de transferencia en bucle cerrado</td>
</tr>
<tr>
<td>147. open-loop transfer function</td>
<td>función de transferencia en bucle abierto</td>
</tr>
<tr>
<td>148. Nichols chart</td>
<td>abasto de Black</td>
</tr>
<tr>
<td>149. position error</td>
<td>error de posición</td>
</tr>
<tr>
<td>150. (system with zero position error)</td>
<td>sistema sin error de posición</td>
</tr>
<tr>
<td>151. system with one integration</td>
<td>sistema con una integración</td>
</tr>
<tr>
<td>152. compensation by integral control</td>
<td>compensación por control integral</td>
</tr>
<tr>
<td>153. phase-lead compensation</td>
<td>compensación por avance de fase</td>
</tr>
<tr>
<td>154. root locus</td>
<td>lugar de Evans, de raíces</td>
</tr>
<tr>
<td>155. phase margin; gain margin</td>
<td>margen de fase, margen de ganancia</td>
</tr>
<tr>
<td>156. performance criterion</td>
<td>criterio de performance</td>
</tr>
</tbody>
</table>
7. Elementary linear systems

**German**

- das System erster Ordnung
- die Vers"ugerung erster Ordnung
- die Zeitkonstante
- die 95%-Zeit
- die Vers"ugerung, die Totzeit
- das System zweiter Ordnung
- die Vers"ugerung zweiter Ordnung
- die Eigenfrequenz
- die Resonanzfrequenz
- die abklingende Schwingung, die ged"ampfte Schwingung
- die Abklingkonstante
- der D"ampfungsgrad [D]
- der "berschwingungsfaktor; die gr"osste Abweichung
- das System mit einem Freiheitsgrad
- das System mit mehreren Freiheitsgraden

**Russian**

- sistema pervovo porjadka
- aperiodicheskoe sveno
- postojannaja vremeni
- zapasdyvanie
- sistema vtorovo dejstvija
- kolobaten'noe sveno
- sobstvennaja chastota
- chastota resonansa
- satuhaigubohee kolebanije
- dekrement satukhanija
- (nailbol'shee mgnoevennoe znachenije oshibki)
- sistema s odnoi stepeni svobody
- sistema so mnogimi stepenjami svobody

8. Linear systems

- das "ubertragungsgesetz, das superfpositionprincip mit konsentrierten Elementen
- mit verteilten Elementen
- das Verhalten f"ur einen beliebigen Eingang
- der Frequenzgang, die Ubertragungsfunktion
- der Ver"uckungsgrad
- die Lage der Pole und Nullstellen
- der Eigenwert
- die Ortskurve (des Frequenzganges)
- der Amplitudenverlauf
- der Phasenverlauf
- (die Nichols-Ortskurve)
- die Resonanz"uberh"ohung der Amplitude
- das Passband, der Durchlassbereich
- das "ubertragungssystem
- das Phasenminimumsystem
- der J-Faktor, das Integrerglied
- der Horizont-Faktor, der D-Faktor
- der Vierpol

9. Linear servo systems

- der "ugungsfrequenzgang
- der Frequenzgang des aufgeschnittenen (offenen) Regelkreises (die Nichols-Tafel)
- die P-Abweichung (die bleibende Abweichung)
- das System ohne P-Abweichung (das System mit I-Verhalten
- der Wursort
- der Phasenrandwinkel; der Amplitudensrand
- das G"utekriterium

**Russian**

- printaip naloshenija
- sistema s sozreotochennymi parametrami
- s raspredeljonnymi parametrami
- povedenie pri ljubom vosdejtvii
- peredatochnaja funktsija
- koeffisient usilenija
- raspredelenie kornej i polusov
- (sveno) [1.3]
- chastotnaja kharakteristika, amplitudno-fasovaja kharakteristika; godografi
- amplitudnaja kharakteristika
- fazovaja kharakteristika
- logarifmicheskaja kharakteristika
- sveno s ravnornym propuskaniem
- neitral'noe sveno, astaticheskoe sveno
- stabiliziruigubohee sveno
- obystrojkhpolusnik
- peredatochnaja funktsija [samknutoj sistemy]
- peredatochnaja funktsija rasomknutoj testoi
- staticheskaja oshibka
- astaticheskaja sistema
- sistema s odnim integriruiguhohim svenom
- (vvedenie integriruhohovo svena)
- (vvedenie stabiliziruhohovo svena)
- kornivnoj godografi
- kriterij khozestva
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>motor stage</td>
<td>órgano motor</td>
<td>l’organe moteur</td>
</tr>
<tr>
<td>motor</td>
<td>motor</td>
<td>le moteur</td>
</tr>
<tr>
<td>servomotor</td>
<td>servomotor</td>
<td>le servomoteur</td>
</tr>
<tr>
<td>direct-current motor</td>
<td>motór de corriente continua</td>
<td>le moteur à courant continu</td>
</tr>
<tr>
<td>alternating-current motor</td>
<td>motor de corriente alterna</td>
<td>le moteur à courant alternatif</td>
</tr>
<tr>
<td>prescribed current</td>
<td>corriente (o intensidad) nominal</td>
<td>le courant nominal</td>
</tr>
<tr>
<td>prescribed speed</td>
<td>velocidad nominal</td>
<td>la vitesse nominale</td>
</tr>
<tr>
<td>driving torque</td>
<td>par motor</td>
<td>le couple moteur</td>
</tr>
<tr>
<td>opposing torque</td>
<td>par resistente</td>
<td>le couple résistant</td>
</tr>
<tr>
<td>load</td>
<td>carga</td>
<td>la charge extérieure</td>
</tr>
<tr>
<td>unloaded characteristic</td>
<td>característica en vacío</td>
<td>la caractéristique à vide</td>
</tr>
<tr>
<td>characteristic under load</td>
<td>característica con carga</td>
<td>la caractéristique sous charge</td>
</tr>
<tr>
<td>separately excited</td>
<td>excitación independiente</td>
<td>à excitation séparée</td>
</tr>
<tr>
<td>field current</td>
<td>corriente de excitación</td>
<td>le courant d’excitation</td>
</tr>
<tr>
<td>field-controlled</td>
<td>controlado por el inductor</td>
<td>à contrôle d’inducteur</td>
</tr>
<tr>
<td>armature</td>
<td>armatura, inducido</td>
<td>l’induit</td>
</tr>
<tr>
<td>armature-controlled</td>
<td>controlado por el inducido</td>
<td>à contrôle d’induit</td>
</tr>
<tr>
<td>slip</td>
<td>dealisamiento, rebalamiento</td>
<td>le glissement</td>
</tr>
<tr>
<td>two-phase motor</td>
<td>motor bifásico</td>
<td>le moteur diphasé</td>
</tr>
<tr>
<td>hydraulic servomotor</td>
<td>servo motor hidráulico</td>
<td>le servo moteur hydraulique</td>
</tr>
<tr>
<td>pump</td>
<td>bomba</td>
<td>la pompe</td>
</tr>
<tr>
<td>pneumatic servomotor</td>
<td>servo motor neumático</td>
<td>le servo moteur pneumatique</td>
</tr>
<tr>
<td>power</td>
<td>potencia</td>
<td>la puissance</td>
</tr>
<tr>
<td>efficiency</td>
<td>rendimiento</td>
<td>le rendement</td>
</tr>
<tr>
<td>German</td>
<td>Russian</td>
<td>Page</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>(die Motorstufe)</td>
<td>ispolnitel’noe ustrojstvo</td>
<td>215</td>
</tr>
<tr>
<td>der Motor</td>
<td>dvigatel’, motor</td>
<td>216</td>
</tr>
<tr>
<td>der Stellmotor</td>
<td>servovdvigatel’, servomotor</td>
<td>217</td>
</tr>
<tr>
<td>der Gleichstrommotor</td>
<td>dvigatel’ postojannovo toka</td>
<td>218</td>
</tr>
<tr>
<td>der Wechselstrommotor</td>
<td>dvigatel’ peremennovo toka</td>
<td>219</td>
</tr>
<tr>
<td>der Nennstrom</td>
<td>nominal’nyj tok</td>
<td>220</td>
</tr>
<tr>
<td>die Nenndrehzahl</td>
<td>nominal’noe snaschene skorosti</td>
<td>221</td>
</tr>
<tr>
<td>das Antriebsmoment</td>
<td>moment dvishushchij</td>
<td>222</td>
</tr>
<tr>
<td>das Gegenmoment</td>
<td>moment soprotivlenij</td>
<td>223</td>
</tr>
<tr>
<td>die Belastung</td>
<td>vneshnaja nagruka</td>
<td>224</td>
</tr>
<tr>
<td>die Leerlaufkennlinie</td>
<td>kharakteristika khloestovo khoda</td>
<td>225</td>
</tr>
<tr>
<td>die Lastkennlinie</td>
<td>nagrusochnaja kharakteristika</td>
<td>225</td>
</tr>
<tr>
<td>mit Fremderregung</td>
<td>s nesavisimyem vosbushdeniem</td>
<td>227</td>
</tr>
<tr>
<td>der Erregerstrom</td>
<td>potok vosbushdenija</td>
<td>228</td>
</tr>
<tr>
<td>mit Feldsteuerung</td>
<td>s vosbushdeniem na potok vosbushdenija</td>
<td>229</td>
</tr>
<tr>
<td>der Anker</td>
<td>jakor’</td>
<td>230</td>
</tr>
<tr>
<td>mit Ankerstromsteuerung</td>
<td>s vosbushdeniem na napriashejen jakorja</td>
<td>231</td>
</tr>
<tr>
<td>der Slutpf</td>
<td>skol’shenije</td>
<td>232</td>
</tr>
<tr>
<td>der Zweiphasenmotor</td>
<td>dvukhfasnyj elektricheskij dvigatel’</td>
<td>233</td>
</tr>
<tr>
<td>der hydraulische Stellmotor</td>
<td>gidavraleskij servomotor</td>
<td>234</td>
</tr>
<tr>
<td>die Pumps</td>
<td>nasoe</td>
<td>235</td>
</tr>
<tr>
<td>der pneumatische Stellmotor</td>
<td>pnuamaticeskij servomotor</td>
<td>236</td>
</tr>
<tr>
<td>die Leistung</td>
<td>mozhchnost’</td>
<td>227</td>
</tr>
<tr>
<td>der Wirkungskoefizient</td>
<td>koeffizient poleznovo deistvija</td>
<td>238</td>
</tr>
</tbody>
</table>
ALPHABETICAL LIST OF GLOSSARY WORDS

A
Absolute stability margin, 104 Acceleration, 25 Accuracy, 25 All-pass element, 142 Alternating current, 45 Alternating-current motor, 219 Amplification without (or with) power, 7 Amplitude response, 135 Angular frequency, 39 Angular velocity, 27 Approximate method, 70 Argument, 90 Armature, 250 Armature-controlled, 231 Autocorrelation function, 167 Autocorrelator, 168 Automatic control, 2 Autopilot, 20 Average value, 160 Axis, imaginary, 91

B
Behavior for any input, 129 Bit (binit), 175 Block diagram, 6

C
Capacitance, 51 Capacity [of a channel], 176 Cascade system, 8 Center, 194 Characteristic equation, 98 Characteristic under load, 226 Characteristic, unloaded, 225 Charge, 53 Circuit, 48 Closed-loop transfer function, 146 Closed-loop system, 11 Closed path, 92 Command, 14 Commutation, 183 Compensating network, 204 Compensation network, 204 Compensation by integral control, 152 Component, 24, 199 Configuration, pole-zero, 132 Conjugate roots, 94 Constant coefficients, with, 64 Control, automatic, 2 Control input, 14 Control system, 1 Controlled weapon, 21 Convolution integral, 85 Coulomb friction, 178 Couple, 30 Criterion of performance, 156 Current, 47

D
Damping, 99 Damped oscillation, 120 Damping ratio, 122 Dead zone, 179 Decrement, 121 Delay, 115 Denominator, 74 Deviation (see error) Derivative with respect to time, 61 Describing function, 187 Describing-function approach, 186 Design, 22 Difference equation, 189 Differential equation, 62 Dimensionless [reduced] variable, 43 Direct current, 44 Direct-current generator, 57 Direct-current motor, 218 Discontinuity, 67 Discontinuous, 177 Distortion, 188 Disturbance, 15 Distributed parameters, with, 128 Driving torque, 222

E
Efficiency, 59, 238 Electromotive force (emf), 56 Equation, differential, 62 Equilibrium, 97 Error, 16 Error-sensing device, 200

F
Feedback, 9 Feedback amplifier, 214 Feedback system, 11 Field-controlled, 229 Field current, 228 Final-value theorem, 84 First-order factor, 112 First-order system, 111 Focus, 194 Follow-up system, 18 Force, 29 Forced oscillation, 185 Forced state, 36 Fourier expansion, 82 Fraction, rational, 72 Frequency, 38 Friction, viscous, 35 Function, 60

G
Gain, 131 Gain margin, 155 Gaussian distribution, 164 Gear, 207 Gear ratio, 205 Guided missile, 21

H
Harmonic function of time, 37 Hunting, 110 Hurwitz's criterion, 106 Hydraulic servomotor, 234 Hysteresis, 180

788
### Alphabetical List of Glossary Words

**I**
- Imaginary axis, 91
- Imaginary part, 89
- Impedance, 54
- Inductance, 50
- Information quantity, 174
- Information theory, 179
- Initial conditions, 66
- Initial-value theorem, 84
- Input, 12
- Instability, 101
- Integration, 143
- Inverse transform, 81

**K**
- Kochenburger approach, 186

**L**
- Lag, 115
- Lag theorem, 83
- Laplace transform, 80
- Laplace transformation, 79
- Left half plane, 103
- Left-hand criterion, 108
- Level regulator, 19
- Limit, 68
- Limit cycle, 197
- Linear, 63
- Load, 224
- Locus, transfer, 134
- Loop, 10
- Lumped-constant [system], 127

**M**
- Magnetic amplifier, 211
- Magnetic flux, 55
- Many-degree-of-freedom system, 125
- Mass, 32
- Minimum-phase-shift system, 141
- Mode, 133
- Moment of inertia, 31
- Motor, 216
- Motor stage, 215
- Multiple root, 95

**N**
- Natural frequency, 118
- Network, 48
- Nichols chart, 148
- Nichols locus, 187
- Node, 195
- Noise, 158
- Nonlinear, 63
- Numerator [of a fraction], 73
- Nyquist criterion, 107
- Nyquist plot, 134

**O**
- One-degree-of-freedom system, 124
- On-off, 182
- Open-loop stable, 103
- Open-loop transfer function, 147
- Opposing torque, 223
- Optimization, 172
- Order of multiplicity, 96
- Oscillation, forced, 185
- Output, 13
- Overshoot, 123

**P**
- Parameters, distributed, 128
- Partial-fraction expansion, 75
- Passband, 139
- Peak value of magnification, 138
- Performance criterion, 156
- Performance operator [see transfer function]
- Periodic solution, 76
- Phase angle, 40
- Phase extension plane, 190
- Phase lead, 144
- Phase-leading compensation, 153
- Phase margin, 155
- Phase portrait, 198
- Phase plane, 190
- Phase response, 136
- Phase space, 191
- Phase trajectory, 192
- Pick-off, 201
- Pneumatic servomotor, 236
- Poisson distribution, 165
- Pole-zero configuration, 132
- Position error, 149
- Potentiometer, 203
- Power, 58, 237
- Prescribed current, 220
- Prescribed speed, 221
- Principle of superposition, 136
- Probability, 159
- Probability distribution, 163
- Pump, 235

**Q**
- Quadratic factor, 117
- Quadripole, 145
- Quality factor, 52

**R**
- Radius vector, 90
- Random, 157
- Rational fraction, 72
- Real part, 88
- Rectifier, 205
- Regulator, 4
- Regulator, level, 19
- Relay, 206
- Relay amplifier, 209
- Residue theorem, 93
- Resistance, 49
- Resonance ratio, [Q], 138
- Resonant frequency, 119
- Response time to within 5%, 114
- Root locus, 154
- Root mean square (rms), 161
- Root-mean-square error criterion, 171
- Rotating amplifier, 212
- Routh's criterion, 105a

**S**
- Saddle, 196
- Saturation, 161
- Second-order system, 116
- Self-adjusting system, 5
- Self-oscillation, 184
- Selsyn, 203
- Separately excited, 267
- Servomechanism, 3
- Servomotor, 217
- Servomotor, hydraulic, 234
- Servomotor, pneumatic, 236
- Servo system, 3
- Simplified, 71
- Singular point, 193
- Slip, 233
- Slope of 6 db per octave, 140
- Smoothing, 170
- Smooth, to, 170
- Specifications, 23
- Spectral density, 166
- a plane, 87
- Spring, 28
- Stability, 100
- Stability domain, 109
- Stable, 100
- State,
ALPHABETICAL LIST OF GLOSSARY WORDS

forced, 36 Stationary, 162 Step response, 42 Stiffness, 34 Stochastic, 157 Superposition, principle of, 126 Synchronous machine, 203 System function (see closed-loop transfer function) System to be controlled, 17 System with one integration, 161 System with zero position error, 150

T
Threshold, 179 Time constant, 113 Time domain, 86 Tend toward infinity, 69 Torque, 30 Transfer function, 130 Transfer locus, 134 Transient, 41 Two-phase motor, 233 Two-stage amplifier, 213

U
Unit impulse, 78 Unit-step, 77 Unit-step response, 42 Unloaded characteristic, 225 Unstable, 101

V
Vacuum-tube amplifier, 210 Variable, 80 Variable coefficients, with, 65 Velocity, 26 Viscous friction, 35 Voltage, 48

W
White noise, 169
Note 1. Block diagrams in German and Russian literature

1. Whereas American and French authors use block diagrams of the servo type (Fig. 1-32a), most German authors use block diagrams of the regulator type, as shown

![Block diagram](image)

**Fig. G1.** Block diagrams of a simple feedback control system: (a) American, (b) German.

in Fig. 1-32b, in which the forward path represents the Regelstrecke and the feedback path represents the Regler. Furthermore, German authors often insert in each block a sketch of the step response (Übergangsfunktion or Sprungantwort) of the corresponding component, rather than the transfer function of the latter. Figure G1a and b shows typical American and German block diagrams for a simple system.

2. It is important to note that most German authors consider the deviation (Regelabweichung \( x_w \)) to be the output minus the input, i.e., the quantity opposite to the error \( e = e - r \) usually considered in the United States, England, and France. As a result, the closed-loop transfer function of a servo system becomes \( KG/(KG - 1) \). Thus the open-loop transfer locus is symmetric with the American one with respect to the origin, and the critical point lies at +1 (Fig. G2).

3. Besides ordinary block diagrams (block-scheme), Russian authors sometimes consider a more abstract type of diagram, called a *structural diagram* (strukturnaja

![Diagram](image)

**Fig. G2.** Typical open-loop transfer locus showing critical point: (a) American, (b) German.
schema). In a structural diagram each block does not correspond to a physical component, but expresses a transfer function that has the form of an elementary factor called \( \text{weno} \) (in the plural: \( \text{weno}s \)), i.e., literally, a link. Examples of \( \text{weno}s \) are \( 1/s \) or integrating \( \text{weno} \); \( 1/(1 + Ts) \), or aperiodic \( \text{weno} \); \( 1/(1 + Ts + T^2s^2) \) or oscillatory \( \text{weno} \); \( (1 + as) \) or stabilizing \( \text{weno} \); \( (1 - as)/(1 + as) \) or all-pass \( \text{weno} \); etc. Figure 9-8 is an example of a structural diagram.

Note 2. Auxiliary loops and compensation in German literature

Rückführungen. In German literature Rückführungen are the feedback paths of secondary loops introduced within the controlling system (Fig. G3). Their denominations implicitly assume (a) that the functions \( G_1(s) \) and \( G_4(s) \) in Fig. G3 are unity and (b) that the forward path of the internal loop introduced by the Rückführung has infinite gain, so that the transfer function of the controlling system (Regler) is inversely proportional to the transfer function \( F \) of the Rückführung itself. Ideally,

\[
K_1G_1K_2G_2K_5G_5 = \frac{K_1}{1 + F} \cdot \frac{K_3}{F(s)}
\]

The technique of servo compensation is thought of by German authors in terms of introducing adequate Rückführungen. The latter are characterized either by the corresponding transfer function \( 1/F(s) \) of the controlling system under the above assumptions or by step response (Übergangsfunktion or Sprungantwort) \( q(t) = \mathcal{L}^{-1}[1/sF(s)] \).

The following cases are important (Fig. G4 to e).

a. P-Regler, i.e., proportional control. The step response \( q(t) = k_p \) is obtained when \( F(s) = 1/k_p \).

b. I-Regler, i.e., integral control. The step response is \( q(t) = kd \), the Rückführung is \( F(s) = s/k_i \).

c. PI-Regler, i.e., proportional-plus-integral control. The step response

\[
q(t) = k_p + kd
\]

is obtained by introducing an auxiliary feedback \( F(s) = s/(k_p + k_d) \) which is called nachgegebende (literally: "giving in") Rückführung.

d. PD-Regler, i.e., proportional-plus-derivative control. A step response of the type \( q(t) = k_p + k_d \delta(t) \), where \( \delta(t) \) is the unit-impulse function, is obtained by introducing a versüßernde (literally, delaying) Rückführung \( F(s) = 1/(k_p + k_d) \).
e. PDI-Regler. This is a combination of proportional, derivative, and integral control:

\[ q(t) = k_p + k_d \delta(t) + k_i t \quad \text{and} \quad F(s) = \frac{s}{(k_d s^2 + k_p s + k_i)} \]

More elaborate combinations can be obtained, for instance, PD₁-Regler with

\[ q(t) = k_p + k_d u_1(t) \]

(see, for example, Fig. 4-15a), PD₁₁-Regler, etc.

Note 3. Stability in Russian literature

1. Concerning stability, Russian authors use a few concepts that are not so commonly used by Western engineers. The most important of these concepts are the following:

a. Russian authors always make the distinction between ordinary stability (устойчивость) and asymptotic stability (асимптотическая устойчивость). Ordinary steady-state stability implies that, after a disturbance, the variables that characterize a system should remain arbitrarily close to their steady-state values provided their initial discrepancies from these steady-state values are small enough. Asymptotic stability implies the additional condition that the variables should for infinite time actually tend toward their steady-state values. For example, a linear system possessing poles on the imaginary axis is stable without being asymptotically stable.

b. When the stability condition is satisfied only for a given range of possible initial conditions, the system is said to be устойчив в малом, i.e., "stable in the small." On the contrary, if the system is stable whatever the values of the initial conditions compatible with the assumptions that define the system, the latter is said to be устойчив в большом, i.e., "stable in the large" (Figs. G5 and G6). The case in

Fig. G5. (a) "Global" and (b) "local" stability for the case of a simple mechanical system.
which a system is stable whatever the initial values of the variable all over the phase
space is called "unlimited stability" (neogranichennaja ustojchivost'), e.g., a stable
linear system. For adaptation of these Russian expressions, we suggest in English
local stability for stability "in the small" and global stability for stability "in the large."

Fig. G6. (a) "Global" and (b) "local" stability for the case of a nonlinear system
with stable focus and saddle. Rectangle is domain of values for x and y that are
compatible with the problem under consideration.

2. For the study of stability, Russian authors use the Routh-Hurwitz criterion
(mostly in the Hurwitz form), the Mikhailov criterion (Nyquist's criterion being con-
sidered as a particular case of Mikhailov's), and the D-razbienie (in German, D-Zerle-
gung), a special representation that could be called in English the \( j \omega \)-breakdown or
D-decomposition (see Prob. 23).

For nonlinear systems one of the methods which seem most popular in Russia is the
second (or direct) method of Liapunov (vtoroj ili prjamoj metod Ljapunova), which is
briefly outlined in Sec. 27.2.3. It should be remembered that linearization as defined
in Sec. 27.1.4 is for Russian authors the first method of Liapunov (pervyj metod Ljapunova
or metod o pervom pribli\( z\)henii).
SELECTED BIBLIOGRAPHY

The theory and the techniques of automatic control have given rise, in recent years, to a proliferation of books and publications of all kind; the number\(^1\) makes it practically impossible to draw up a complete bibliography. The list of bibliographical references that follows does not aim at presenting a complete view of the literature, as the present authors do not claim to have read everything on the subject. A certain number of works of general\(^2\) interest on automatic control are merely listed below; they are works that the present authors have read and used and that they recommend as useful references for students and engineers in the field. In each of them are listed numerous other bibliographical references, which easily enables the reader to compile for himself a more comprehensive bibliography\(^3\) on the subject in which he is more especially interested.

The works have been classified according to the subjects they cover and sometimes according to the language in which they are written. A few comments have been added on most of them in order to show their relationship to the subjects covered in this book.

1. GENERAL WORKS ON AUTOMATIC CONTROL

1.1. Works in English

A. C. HALL: "The Analysis and Synthesis of Linear Servomechanisms" (Technology Press, Cambridge, Mass., 1943) is one of the first attempts at a systematic and realistic application of feedback theory to servomechanisms. The Hall-Sartorius criterion is outlined in chap. 2. The bulk of the work deals with compensation, including the use of cascade lead controllers, impedance matching, and compensating-network synthesis.

Leroy McColl: "Fundamental Theory of Servomechanisms" (Van Nostrand, Princeton, N.J., 1945) is the first widely circulated book on the general theory of automatic control published in English. All the essentials of linear servo theory are outlined, including sampled-data systems. A great number of subsequent works are merely rewritings or developments of specific parts of this fundamental work. Published in French (S.D.I.T., Ministère de l'Air, no. 4208) and in Russian (Iinozdat, 1947).

H. LaVer, R. Lesnick, and L. E. Matson: "Servomechanism Fundamentals" (McGraw-Hill, New York, 1947) presents a valuable, instructive study of simple servo systems (the frequency-response approach is not much developed). Of special value are the chapter devoted to follow-up links and the physical interpretations given for derivative and integral control.

G. Brown and D. Campbell: "Principles of Servomechanisms" (Wiley, New York, 1948) is an excellent introductory textbook for linear servomechanisms and presents all the essentials of servomechanism theory and technique in a clear and practical manner. Floyd's method for obtaining the transient response from the frequency response is outlined in appendix A.

\(^1\) Four hundred publications in English alone in the month of July, 1953!

\(^2\) References more specifically related to certain particular aspects of feedback control have been cited throughout the book (for example, references on sampled-data servo systems on page 335). As a rule, they are not cited again in this bibliography.

\(^3\) A very comprehensive bibliography on automatic control has been published under the title "Bibliography in Feedback Control," Applications and Industry, pp. 430-460, January, 1954. A complete list of Russian references on automatic control will be found in Automatika i Telemekhanika, vol. 9, no. 5, 1948, and vol. 10, nos. 5 and 6, 1949.
H. M. James, N. B. Nichols, and R. S. Phillips: "Theory of Servomechanisms" (McGraw-Hill, New York, 1947) outlines the general theory and design techniques of linear servo systems. A systematic study is made of steady-state errors (chap. 4). Three excellent chapters are devoted to statistical considerations in feedback control and to the rms criterion. Published also in Russian.

I. A. Greenwood, Jr., J. V. Holdam, Jr., and D. Macrae, Jr.: "Electronic Instruments" (McGraw-Hill, New York, 1948) offers in part 2 a highly instructive presentation of the theory of linear servo systems, a chapter being devoted to servo components and, especially, to servomotor control.

W. R. Ahrendt and J. F. Taplin: "Automatic Feedback Control" (McGraw-Hill, New York, 1951) provides an excellent presentation of linear servo theory with discussions on saturation and discontinuities. Numerous examples from the technical field and problems with solutions are given.


J. G. Truxal: "Automatic Feedback Control System Synthesis" (McGraw-Hill, New York, 1955) is a work of great importance with many original viewpoints. It has introduced a great number of new ideas and methods into the servo field. Of special value in this connection are the chapters on signal-flow diagrams, compensating-network synthesis, and the Guillemin-Truxal approach for compensation in the s-plane.

Under the title "Frequency Response" (Macmillan, New York, 1956), R. Oldenburger has collected the papers given at the ASME Conference in New York, in 1955, and a number of classical papers concerning the fundamentals of automatic-control theory. Many interesting viewpoints and examples will be found in this valuable volume.

O. J. M. Smith: "Feedback Control Systems" (McGraw-Hill, New York, 1958) is an advanced book that affords thorough and often original studies of many linear and nonlinear feedback control problems. It is of special interest to those interested in research and discussions on advanced feedback control problems.

"Control Engineers' Handbook," edited by J. Truxal (McGraw-Hill, New York, 1958) contains a summary of the theoretical aspects of servo theory plus twelve chapters on servocomponents. It presents to the practicing engineer one of the fullest existing views on present-day American technology in the automatic-control field.

Wiley (New York) is publishing a "Handbook of Automation, Computation and Control," by E. Gansbe, S. Ramo, and D. Woolridge. Volume 1 ("Control Fundamentals") presents a vast number of mathematical results and formulas useful in the fields of control and communication engineering.

1.2. General works in French:

P. Naslin: "Les systèmes asservis" (Revue d'Optique, Paris, 1951) covers the subject of linear-servo theory (with emphasis on compensation techniques), including a-c compensating networks, statistical considerations and sampled-data systems. "Technologie et calcul pratique des systèmes asservis" (Dunod, Paris, 1957), by the same author, provides a condensed presentation of frequency-response methods and some chapters on components.

J. Gille, M. Pélegrin, and P. Decaulne: "Théorie et technique des asservissements" (Dunod, Paris, 1956, 1958) is the French equivalent of the present book, with a few minor changes. Published in German (Oldenbourg, Munich, 1959), in Russian (Masgiz, Moscow, 1959), and in Polish (Panstwowe Wydawnictwa Techniczne, Warsaw, 1959). Problems with solutions are published in French and German as a separate book, "Problèmes d'Asservissements avec solutions."

M. Bonamy: "Servomécanismes, théorie et technologie" (Masson, Paris, 1957) consists of a very clear treatment of linear servo theory by frequency-response methods followed by chapters on components.
1.3. General works in German

The first works of general interest on automatic-control theory and technique were written by Germans during World War II.


R. Oldenbourg and H. Sartorius: "Dynamik selbsttätiger Regelungen" (1944, 1951) is a work of first-rank importance that can be considered as the German counterpart of the American works by Hall and McColl. It provides a systematic general treatment of linear-servo theory, with emphasis on the choice of parameters, and also presents a study of certain types of nonlinearities. Published in English (ASME, New York, 1948), in Russian, and in Japanese.

F. Steckler: "Praktische Stabilitätsprüfung mittels Ortskurven und numerischer Verfahren" (Springer, Berlin, 1950) gives a systematic presentation of the harmonic approach from the different viewpoints of complex variable theory, and incorporates some material on nonlinear problems.


1.4. General works in Russian

To the authors' knowledge, the most important Russian work on automatic control in general is "Osnovy avtomaticheskovo regulirovaniya. Teorija" (Masgiz, Moscow, 1954) by seventeen authors under the direction of V. Solodovnikov. This work contains a very clear presentation of linear theory (700 pages, with emphasis on stability, transients, and statistical methods) and nonlinear theory (350 pages, with emphasis on oscillations and transients). Published in German (Oldenbourg and Technik, 1958).

M. Ajzerman: "Teorija avtomaticheskovo regulirovaniya dvigatelej: uravnenija dvizhenija i ustoichivost'" (Gostekhizdat, Moscow, 1952) provides a presentation of the essentials of automatic control, with emphasis on stability. The method of the D-rasbiejente for linear stability (see Prob. 23) is explained in chap. 8, and applications of Liapunov's criterion for nonlinear stability are developed in chap. 13. By the same author: "Lektssi po teorii avtomaticheskovo regulirovaniya" (Gostekhizdat, Moscow, 1956).

M. Meerov: "Vvedenie v dimaniu avtomaticheskovo regulirovaniya elektricheskikh mashin" (Academy of Sciences, Moscow, 2d ed., 1956) contains a thorough treatment of electric servomotor control. Published in German (Technik, Berlin, 1954).

2. MATHEMATICAL BACKGROUND

Chapters 4, 5, and 8 of M. Gardner and F. Barnes. "Transients in Linear Systems" (Wiley, New York, 1942) provide the essentials of Laplace-transform theory and are perfectly adequate for most servo engineers. A more advanced theory of the Laplace transform can be found in G. Doetsch: "Handbuch der Laplace Transformation" (Birkhäuser, Stuttgart, 1950, 1955).

For questions of probability and noise the authors strongly recommend the excellent work by A. Blanc-Lapierre and J. Fortet: "Théorie des fonctions aléatoires" (Masson, Paris, 1953). For noise problems see also "Le bruit de fond," by P. Grivet and A. Blaquiére (Masson, Paris, 1958).

To our knowledge the most complete work on probability and random processes with emphasis on automatic control applications is "Teorija sluchajnykh funktsij i ejo primenenije k zadacham avtomaticheskovo regulirovaniya," by V. Pugachlov (Gostekhizdat, Moscow, 1957; under publication in French by Dunod, Paris).
For nonlinear systems an introduction to the theory of nonlinear differential equations is presented by E. Picard: "Traité d'analyse," vol. 3, chaps. 8, 10 (Gauthier-Villars, Paris, 1927). A very up-to-date presentation of the theorems relative to nonlinear differential equations will be found in G. Sansone and R. Conti: "Equazioni differenziali non lineari" (Edizioni Cremonese, Rome, 1956).

3. LINEAR SYSTEMS IN GENERAL

Works on linear systems in general can be useful as reference books for servo engineers because of the close connection between the automatic-control field and mechanical and electrical engineering. In the United States a basic reference work on linear dynamics is M. Gardner and F. Barnes: "Transients in Linear Systems" (Wiley, New York, 1942). In Europe, Y. Rocard: "Dynamique des vibrations" (Masson, Paris, 1943) is considered the classic work on the subject.

The two following works on electrical engineering are important because of the influence they had on the development of servo theory: H. Bodie: "Network Analysis and Feedback Amplifier Design" (Van Nostrand, Princeton, N.J., 1945); E. Guillemin: "Communication Networks" (Wiley, New York, 1951).

Finally, a very great number of useful diagrams and charts concerning linear systems of the first, second, and third order will be found in C. S. Draper, W. McKay, and S. Lees: "Instrument Engineering," especially vol. 2 (McGraw-Hill, New York, 1952).

4. STATISTICS

4.1. Applications of statistical theories to servo design

The application of statistical considerations to the design of servo systems was dealt with for the first time in the James, Nichols, and Phillips 'paper (Sec. 1.1, Bibliography).

N. Wiener: "The Interpolation, Extrapolation and Smoothing of Stationary Time Series" (Wiley, New York, 1950), although it involves advanced mathematics and has little appeal for the average servo engineer, became the basis of many subsequent works on optimization.

One of the best of such works is G. Newton, L. Gould, and J. Kaiser: "Analytical Design of Linear Feedback Controls" (Wiley, New York, 1957), which has the merit of quantitatively specifying the farthest theoretical limits of accuracy of linear feedback control.


To those who read Russian we recommend V. Solodovnikov: "Vvedenie v statisticheskii dinamiku sistem avtomaticheskogo upravlenija" (Gostekhizdat, Moscow, 1952), which presents an elegant synthesis of the applications of statistical considerations to the servo field. Published in German (Oldenbourg, Munich, 1957).

4.2. Cybernetics

The idea of feedback control and especially its statistical aspect has tempted many authors to apply it to human (especially physiological and social) problems. This is the vogue of cybernetics, introduced by N. Wiener: "Cybernetics, or Control and Communication in the Animal and the Machine" (Hermann, Paris, 1949) and amplified by a great number of more recent publications. Many such works lack both technical seriousness and philosophical solidity. The present authors strongly recommend the valuable discussion of related problems by P. Cossa: "La cybernetique: du cerveau humain aux cerveaux artificiels" (Masson, Paris, 1955).

4.3. Information theory


1 After the Greek word kubernés, pilot. In Modern Greek, kubernésis means Government.

Especially interesting applications of information theory are presented by M. Nichols and L. Rauch in their excellent work "Radio Telemetry" (Wiley, New York, 1954, 1956).


5. NONLINEAR SYSTEMS

This field is extremely comprehensive. Contrary to books on linear systems, the different works published on this subject generally stress very different aspects of the question and may have scarcely anything in common with one another.

5.1. Fundamental works

The great majority of works published in the nonlinear field derive from the two fundamental publications of H. Poincaré: "Sur les courbes définies par des équations différentielles" (Journal de mathématiques pures et appliquées, Gauthier-Villars, Paris, 1881, 1882), and "Les méthodes nouvelles de la mécanique céleste" (Gauthier-Villars, Paris, 1892, 1893, and 1899).

The fundamental work of Lyapunov should also be cited: "Obshchaja zadacha ustojchivosti dvizhenija" (Kharkov Mathematical Society, 1892). This publication was translated by E. Davaux, published as "Problème général de la stabilité du mouvement" by the Faculté des Sciences of Toulouse [Annales, 9:203–469, (1907)] and reprinted in 1949 by Princeton University Press.

5.2. General works on nonlinear mechanics

Works on nonlinear mechanics are numerous, but few of them are directly adaptable to automatic-control problems. Thus, their interest for servo engineers is somewhat similar to that of Bode's or Guillemin's works in the linear field.

An elementary, but clear and didactic, presentation of some important nonlinear concepts is provided by N. Minorsky: "Introduction to Nonlinear Mechanics" (Edwards, Ann Arbor, 1947). Another valuable English reference is J. Stoker: "Nonlinear Vibrations in Mechanical and Electrical Systems" (Interscience, New York, 1950).

Two fundamental Russian works on the subject have been adapted by S. Lefschetz and published by Princeton University Press, in 1943 and 1949. They are respectively N. Krylov and N. Bogolyubov: "Introduction to Nonlinear Mechanics" (Kiev, 1937), and A. Andronov and C. Chaikin: "Theory of Oscillations" (Moscow, 1937).

The Russian classic work on stability is "Ustojchivost' dvizhenija," by N. Chetaev (Gostekhizdat, Moscow, 1946).

Two more recent Russian works by N. Malkin should also be cited: "Teorija ustoichivosti dvizhenija" (Gostekhizdat, Moscow, 1952) provides a very clear and complete treatment of the question of stability (stability of equilibrium and self- and forced oscillations, including the critical cases); "Nekotorye zadachi teorii nelinejnykh kolebanij" (Gostekhizdat, Moscow, 1957) is a more advanced book that presents a thorough application of Poincaré's perturbation method to the oscillations of nonlinear systems.

In "Introduction to Nonlinear Analysis" (McGraw-Hill, New York, 1958), W. Cunningham offers an excellent treatment of graphical and analytical methods for studying nonlinear systems, with chapters on difference equations, equations with variable coefficients, and numerous interesting examples.
5.3. Works on nonlinear feedback systems


Different aspects of this method are discussed in various publications of interest; for example, E. C. Johnson, "Sinusoidal Analysis of Feedback-control Systems Containing Nonlinear Elements," *Trans. AIEE*, 71:169–181 (1951), and J. Loeb, "Recent Advances in Nonlinear Servo Theory" in "Frequency Response," pp. 260–268, edited by R. Oldenburger (Macmillan, New York, 1956). Applications are outlined in different works; one of the most interesting is to be found in the work already cited (Sec. 1.1, Bibliography) by Chestnut and Mayer (vol. 2, chap. 8). Also see the publications referred to in our Chap. 24.

An original method based on energy considerations was developed by K. Klotter and I. Flugge-Lotz in the *ZWB Untersuchung u. Mitteilung* 1326 "Über Bewegungen eines Schwingers unter dem Einfluss von Schwarz-weiss-steuerungen" (Berlin, 1943), reprinted in English (NACA Technical Memorial no. 1237, 1949). See also I. Flugge-Lotz: "Discontinuous Control" (Princeton University Press, 1953).


The two following Russian books provide rigorous methods for nonlinear servo systems. J. Z. Tsyplin: "Teorija relejnych sistem avtomatichesko regulirovanija" (Gostekhizdat, Moscow, 1956) offers a synthesis of on-off control systems, emphasis being laid on self- and forced oscillations with original methods for the discussion of their existence and stability. This book also deals with the linearization and optimization of on-off control systems. Published in German (Oldenbourg and Technik, 1958).

A. Isetov: 1 "Ustoichivost' nelinejnykh reguliruemikh sistem" (Gostekhizdat, Moscow, 1955) develops the application of Liapunov's criterion (or Liapunov's direct method) to general servo systems incorporating one or two nonlinear components for wide classes of nonlinearities.

Other methods applicable to nonlinear feedback systems will be found in the nonlinear part of Soledovnikov's work (Sec. 1.4, Bibliography). For other works devoted to Liapunov's direct method, see the publications referred to in our Sec. 27.2.3.

R. Cosgriff: "Nonlinear Control Systems" (McGraw-Hill, New York, 1958) provides a clear review of practical methods used in feedback control engineering with special emphasis on nonlinear servos.


6. SERVOCOMPONENTS

Most of the literature devoted to servomechanisms in past years concerns the theoretical aspects of automatic control, and it is to be regretted that few important works have been written on servo-system components.

Concerning electric servomotors, a noteworthy exception is the excellent work of A. Tustin: "Direct Current Machines for Control Systems" (Spon, London, 1952). Much interesting technical data relating to a-c servomotors can be found in J. Fallou: "Lecons d'électrotechnique générale" (Gauthier-Villars, Paris, 1948).


So far as hydraulic servomotors are concerned, the most important book published is J. Horne, J. Shearer, J. Blackburn, G. Reethof, and S. Lee: "Fundamentals

1 Or Liotov. Pronounce: Liotof.
of Fluid Power Control” (Wiley, New York, 1958), which presents a thorough study of fluid power control from first principles and provides a great deal of useful data for the designer.

Finally, it should be recalled that J. Truxal’s “Control Engineers’ Handbook” (mentioned in Sec. 1.1, Bibliography) contains twelve valuable chapters on servo and regulator components.

7. VARIOUS PROCEEDINGS AND PERIODICAL PUBLICATIONS

A very great number of congresses and symposia in recent years have been at least partially devoted to problems of automatic control. The interest of the corresponding proceedings and reports of discussions varies. We recommend the following, in which valuable contributions can be found, especially on nonlinear problems. (The more important ones are marked with an asterisk.)


“Trudy vtorovo vsesojuznovo sovshchanija po teorii avtomaticheskovo regulirovanija” (3 vols., Akademija Nauk, Moscow, 1955) contains the proceedings of a congress organised by the Institute for Automatic Control of the Academy of Sciences and held in Moscow, 1953.


Finally, valuable papers on servo systems can be found in various technical magazines. The following journals treat questions of automatic control exclusively.


In Russian: Avtomatika i Telemekhanika, edited by the Institut Avtomatiki i Telemechekaniki, Academy of Sciences, Moscow.

In German: Regelungstechnik, edited by H. Sartorius, Oldenbourg, Munich.

INDEX

A-c motor, 570–571, 591–600
  comparison of, with d-c motor, 571
  principle of, 591–592
  (See also Two-phase motor)

Acceleration constant, 120
  from pole-zero configuration, 171–172
  of servo system, 264

Acceleration error, 50, 120
  of servo system, 263–264

Acceleration gain, 119

Accelerometer, 535–537
  equation of, 83–84
  orders of magnitude for, 106–107

Additivity (superposition), 26n.

Aircraft in longitudinal motion, equations of, 127–129
  root locus of, 250
  simulation of, 368–369
  transfer locus of, 132–136
  (See also Autopiloting)

Aircraft stabilization, longitudinal, 277–278
  in roll, 255–258
  (See also Autopilot for guided missile; Autopiloting)

All-pass element, 154

Amplidyne, 655–658
  principle of, 655–656
  transfer function of, 657

Amplifier, classification of, 643
  comparison of different types of, 687, 694–696
  definition of, 641
  rotating (see Rotating amplifier)

Amplitude, 48n.

Amplitude-phase plot (see Nichols locus)

Amplitude response, of first-order system, 72, 74–75
  of second-order system, 84, 90–91

Analog computer, 367–370

Analogy, 37–42, 367–370
  electro-acoustic, 37n.
  electro-hydraulic, 726
  electro-mechanical, 38–40

Analogy, electrostatic, 176
  hydraulic, 175

Angle condition, 239, 240n.

Anticipating path, 316–317

Armature-controlled motor, 586–590
  characteristic curves of, 588–589
  comparison of, with field-controlled motor, 589–590
  principle of, 586–587
  transfer function of, 587

Armature design, example of, 717–718

Asymptotic stability (see Stability)

Autocorrelation function, 204–206, 210

Autocorrelator, 205–206

Automatic control, 9

Autopilot for guided missile, basic design of, 710–718
  choice of motor, 714–715, 717–718
  compensation, 715–717

Autopiloted aircraft, root locus of, 250–251

Autopiloting, 12–14, 16, 265–267

Axial-piston motor, 608–609

Ballast resistor, 581

Bandwidth, choice of, 208–209
  of second-order system, 104–106
  of servo system, 228, 287

Barrel-cylinder motor, 608–609

Basic design of servo system, examples of, 699–710, 711–718
  principles for, 698–699

Bendixon's criterion, 491

Bernoulli's law, 613

Block diagram, 1
  of multiple servo system, 365
  of regulator, 11
  of servo system, 11, 297
  with one nonlinear component, 391–392
  (See also Structural diagram)

Bode plot, 142–143

Bode's equation, 155
Booster, 628–631
Borel’s theorem (see Convolution integral)
Branch of electric network, 32
Breakaway point, 242–243

Capacitance, 31
Capacitance sensing device, 515–516, 526
Carson transform, 67
Cascade system, sampled, 340
   synthesis of, 187–190, 312–314
   transfer function of, 130, 179
   transfer matrix of, 182
Cauchy’s theorem, 125
Center, 485
Characteristic curves, of servomotor, 549–551, 552–553
   armature-controlled motor, 588
   field-controlled motor, 585
   internal impedance read from, 553–554
   positioning of, 554–563
      harmonic-response method, 555–561
      modulus method, 562–563
   two-phase motor, 595–600
   of transistor, 682–684
Characteristic equation, 52, 116, 222
Characteristic impedance, 183, 192
Clamping, 337
Closed-loop control, 9
Closed-loop system (see Feedback system)
Closed-loop transfer function, 215
Command, 5
Commutator-type dynamo, 523–524
Commutatorless dynamo, 524–525
Compensating network, 11
   electrical, 298–299, 309–310, 314
   mechanical, 299–300
Compensation, 295–317
   by addition, of dipole, 322–323
   of zero, 321–322
   by anticipating path, 316–317
   German terminology for, 772–773
   by integral control, 307–312, 423
   of nonlinear servo, 422–423
   by phase lead, 297–307
   of time lags, 304–305, 423
Compound excitation, 569
Conditional stability, 254, 259, 276

Conformal mapping, 126, 280
Constant (see Time constant)
Control, 1–4
   integral (see Integral control)
   Control input (command), 5
   Controller (see Compensating network)
   Controlling mode, 171
Convolution integral, 61–62
   application of, 400
   interpretation of, 124
   proof of, 65–66
Copying servo system, example of, 677–679
Correlation function, 205–210
Coulomb friction, 194, 435–436
Cremer-Leonhard criterion, 175n.
Critical amplitude, for on-off servo system, 429–430, 476–477
   for Van der Pol’s equation, 378
Critical case of Liapunov’s approximation, 489–490
Critical locus, 406
Critical point, 227
   (See also Singular points)
Cross correlation, 205
Curvature, 381, 384, 549–550
Cut-off amplitude (see Critical amplitude)
Cybernetics, 778

D-c motor, 569
   comparison of, with a-c motor, 571
   numerical data for, 602–603
   steady-state equations of, 574–576
   (See also Armature-controlled motor; Field-controlled motor)
Damping, 99–101
   effect of, on frequency response, 99–100
   on transient response, 100–101
   of hydraulic actuator, 628
   of hydraulic transmission, 622
Damping ratio, 89, 99, 106–107
   orders of magnitude for, 106–107
   relation of, to pole configuration, 110–111
Dangerous stability boundary, 418–419, 489–490
Dead zone, 381–384
   describing function of element with, 411–412
Dead zone, influence of, on on-off servo, 416–417
   on-off servo with, self-oscillations of, 416–417, 443, 449, 464–475
   transient of, 397–398
Decade, 139n.
Decibel, 138
Decilog, 139n.
Delay, 81
   of nonlinear element, 393
   of relay, 647
   (See also Lag)
Derivative control, 297–298
   (See also Phase-lead compensation)
Describing function, of arbitrary nonlinear element, 412–414
   of common nonlinear elements, 406–412
   definition of, 403–406
Detector (see Error-sensing device; Rectifier)
Differencer (differencing operator), 7, 503–504
Differentiation theorem, 54–55, 65
Dimensionless variable, for aircraft, 128–129
   for first-order systems, 73
   for second-order systems, 89
Dipole, in pole-zero configuration, 322–323
   as two-terminal box, 178
Distributing valve, 605, 610–612
Distributor, 615
Disturbance, 5–7, 220–222, 261–262
Dither effect, 199, 511, 612
Dominant mode, 171
Drum winding, 573
Dual network, 35–37
Duhamel’s theorem (see Convolution integral)

Eddy currents, influence of, on time constants, 584–585
Efficiency, of electric servomotor, 571–573, 589–590
   of servomotor, 545–547
Electric servomotor, a-c, 591–600
   armature-controlled, 586–590
   compared to hydraulic motor, 633–635
   efficiency of, 571–573, 589–590
   field-controlled, 581–586
   Electric servomotor, numerical data for, 600–603
   special features of, 571
   Electrometer tube, 527–528
   Equilibrium, 46, 485
   Equivalent gain (see Describing function)
   Ergodic hypothesis, 202–203
   Error (see Velocity error)
   Error-sensing device, choice of, 503, 537–538
   classification of, 511–512
   definition of, 20, 503–504
   Experimental determination, of linearity, from sinusoidal tests, 151
   from step response, 152
   of transfer locus, 150–154
   from response to arbitrary input, 152–153
   from sinusoidal tests, 150–151
   from step response, 151–152
   (See also Describing function)
Feedback path, compensating network in, 315
   transfer function in, 219–220
Feedback system, 9
   philosophy of, 19
   properties of, 17–19
   transfer function of, 130–131
Field-circuit calculations, 576–577, 578–579
Field-controlled motor, advantages of, 585
   characteristic curves of, 585
   comparison of, with armature-controlled motor, 589–590
   principle of, 581–582
   transfer function of, 582–584
   variations of, 585–586
Filter, 18
   first-order system as, 103
   second-order system as, 103
Filtering, 77, 674–675
   out of high frequencies, 84, 133–136
   consequences of, 357, 404–405, 461
   sampled-data, 355–356
Final-value theorem, 58, 65
First-harmonic approximation, discussion of, 431, 446, 463
   principle of, 403–404
First-order system, equation of, 68–69
  examples of, 78–80
  response of, frequency, 72–76
    impulse, 70–71
    to sampled input, 345–348
    step, 71
  transfer function of, 77–78
Flow graph, 222–223
Floyd’s method, 145n.
Fluctuating energy, 541
Focus, 486
Follow-up system, 10
Frequency, angular, 48n.
Frequency entrainment (see Van der Pol’s equation)
Frequency response, 48–49
  of first-order system, 72–76
  of nonlinear system, 84–87
  relation of, to transfer function, 87, 121–122, 126
  of second-order system, 84–87
Frequency-response locus (see Transfer locus)
Frequency spectrum (see Spectrum)
Friction damper for stabilization, 316, 735

Gain, acceleration, 119
  of second-order system, 99
  static, 116–118
  velocity, 119
  (See also Open-loop gain)
Gain crossover frequency, 228
Gain margin, 281–286
Gear ratio, selection of, 542–543
Gyroscope, device, 532–535
  gyropole, 299–300

Hall’s chart, 224
Hall-Sartorius criterion, 290–291
Harmonic approach, 126, 158
Harmonic response (see Frequency response)
Harmonic-response method for servomotors, 555–561
  analytical solution for, 556
  example for, 564–568
  graphical solution for, 557–561
  principle of, 555–556

Hectopascal (kg/cm²), 617a.
Holding, 337
Homeostasis, 18, 19
Homogeneity (superposition), 26n., 151, 152
Hopkinson’s coefficient, 580–581
Human operator, 4, 290
Hydraulic amplifier, 680
Hydraulic motor, comparison with electric motor, 633–635
  description of, 605–609
  equations of, 623–624
  numerical data for, 635–638
  transfer function of, 628–628
Hydraulic transmission, 610, 619–623

Impedance of servomotor, internal, 553–554
Impedance matching, 187–192
  for cascade synthesis, 187–190
  for maximum power transfer, 190–192
Impedances, considered in amplifiers, 642
  of quadrupole, 183
Impulse function, 58, 60–61
  Laplace transform of, 59–61
Impulse response, 48
  determination of, from pole-zero configuration, 173
  of first-order system, 70–71
  relation of, to transfer function, 152
Indication time, 70
Inductance, 30–31, 32
  calculation of, 579–581, 584
  effect of, on time constant, 79–80, 577–579
Inertia of servomotor, 541–542, 543, 626
Information theory, basic ideas, 210–214
  capacity, 199, 213
  flow, 211–212
  informational efficiency, 213
  informational matching, 213–214
  quantity, 210–211
Initial conditions, for electric networks, 33
  for linear systems, 64, 124–125
Initial-value theorem, 58, 65
Input, control, 5
  unwanted (disturbance), 5–7, 220–222, 261–262
Integral control, 307–314
  adaptation of, 310–312
  combination of, to phase lead, 312–313
  network for, 309–310
  overcompensated, 274, 309
  principle of, 307–308
  undercompensated, 264, 308, 309
Integral-squared-error criterion, 290–291
Integration in transfer function, 116
  effect of, on servo error, 225–226
  on steady state, 118–119
  on transfer locus, 137–138, 141
Inverse transfer locus (see Transfer locus)
Isoclines, method of, 433
ITAE criterion, 291

jω-breakdown, 731
Jump phenomenon, 430

Kirchhoff’s laws, 27, 32–33

Lag, comparison of, to time constant, 80–92, 196
  compensation of, 303–304
  determination of tolerable, 289–290, 703
  influence of, on limit cycle, 418, 440–441
  on servo performance, 234–236, 289–290
  (See also Integral control)
Laplace transform, of common functions, 54, 59
  definition of, 53–54, 64–65
  differentiation theorem for, 54–55, 65
  inverse, 66
  properties of, 61–62
  of sampled functions, 339–340
Lead factor, 168, 255
Leaking valve, 612
Left-hand criterion, 279–280
  application of, to nonlinear servos, 414–415
Leonhard’s criterion, 175n., 282
Liapunov, first-degree approximation of, 488–489
Liapunov’s criterion (direct method), 493–497
  application of, example, 494–496
  discussion of, 496–497
  principle of, 493–494
Liénard’s construction, 433–434
Limit cycle, 416–417
  existence of, 491
  stability of, 419–421, 491–493
  (See also On-off servo system)
Linear actuator, 605
Linear system, 24, 26–29
Linearity, conditions for, 26n., 151
  of motor, 549, 624–625
Linearity domain, 198–199
  in amplitude, 193–195
  in frequency, 27–29, 195–198
Linearization, 488–489, 498–501
  of hydraulic motor, 624–625
  by noise, 509–510, 513n.
Liouville’s theorem, 125–126
Liapunov (see Liapunov)
Loeb’s criterion for stability of limit cycle, 419–421
Logarithmic coordinates, 138–141
  for first-order systems, 73–75
  for second-order systems, 89–91
Ludbrook’s theorems, 122, 148
Lumped-constant systems, electric, 29–37
  components of, 29–32
  equations of, 32–33
  mechanical, 38–40
Lurje transform, 496

Magnesyn, 520
Magnetic amplifier, 665–668
Magnetic bridge, 522
Magnetostriction-type sensing device, 526
Magnification, peak value of (see Resonance ratio)
Magnitude condition, 235, 240n.
Magnitude response (see Amplitude response)
Margin, stability (see Stability margin)
Matching, impedance (see Impedance matching)
Matrix (see Transfer matrix)
Mean-square theorem, 203
Mesh equation, 32, 35
Metadyne, 655–658
  as constant-current source, 581–582, 659–660
  principle of, 655–656
  transfer function of, 657–658
Mikhailov’s criterion, 175n.
Minimum-phase-shift system, 154
Missile, basic design of autopilot for
  (see Autopilot for guided missile)
Mode, 111, 161
  dominant, 171
Modulation, compared to sampling,
  338, 355–356
  compared to vibrating relay, 499
Modulus method, example of, 564–568
  principle of, 562–563
Multiple feedback systems, 364–370
  definition of, 364, 366
  equations of, 370–371
  stability of, 371–372

Natural frequency, 51, 103, 144, 287
  of hydraulic actuator, 628
  orders of magnitude for, 107, 289
  of second-order system, 101–102, 111
Needle valve, 612
Network, electric, 29–37
  components of, 29–32
  equations of, 32–33
  mechanical, 40–41
Nichols’ chart, 224
Nichols locus, 139–141
  of first-order system, 76
  practical determination of, 142–143
  of second-order system, 91–95
Node, 486
Node equation, 32, 35
Noise, 200
  in amplifiers, 511, 643
  linearization by, 509–510, 513n.
  output, 207–208
  in potentiometers, 382, 510–511
  in sensing devices, 505–511
  dynamic, 509
  macroscopic, 508–509
  microscopic, 506–508
  flicker, 508
  shot, 507–508
  thermal, 507
  threshold due to, 386–388
Nonlinear component, servo system with
  one, 390–393
Nonlinearity, classification of, 380–386
  definition of, 379–381
  incidental, 388
  purposeful, 388–389
Nonminimum-phase-shift system, 154–157
Nonsynchronous motor, 570
Nonunity feedback system, 219–220, 276
Nozzle, 615
Numerical data, for electric servomotors,
  600–603
  for hydraulic servomotors, 635–638
Nyquist’s criterion, 268–276
  application of, to inverse locus, 275–276
  practical, 272–274
  for sampled-data systems, 359–360
  statement of, 271–272
  theory of, 269–271
Nyquist locus (see Transfer locus)

Octave, 138
On-off component, describing function
  of, 406–410
  with hysteresis, 409–410
  without hysteresis, 406–409
On-off servo system, forced oscillations
  of, 429–430, 475–477
  existence of, 429, 476–477
  stability of, 429n., 480–481
  limit cycles of, 416, 436–439, 449–475
  French approach for, 449–454, 465n.
  influence on, of dead zone, 416–417, 418
  of hysteresis, 418
  of lag, 418, 440–441
  Russian approach for, 455–463, 464–475
  stability of, 416–417, 481–483
  optimisation of, 443–445
  transient of, 394–398, 436–440
  with dead zone, 397–398
  without dead zone, 394–397
  with hysteresis, 397, 398
Open-loop control, 1–4, 9
Open-loop gain, 216–217
  adjustment of, 229–234
  on Hall’s chart, 231–232
  on Nichols’ chart, 232–234
  as performance criterion, 263–264, 287
Open-loop transfer function, 215–217
Optimisation, 319, 329, 333
of on-off servo, 443–445
Overshoot, 50
of second-order system, 100

Patin transmission, 521
Peak value of magnification (see Resonance ratio)
Peltier effect, 529
Performance of second-order system, 98–106
bandwidth, 103–105
damping, 99–101
natural and resonant frequencies, 101–103
response time, 106
Performance criteria, 284–294
general-purpose, 286–289
for multiple servo, 372–373
use of, 294
(See also Hall–Sartorius criterion; RMS-error criterion)
Performance operator, 116
Performance specifications, 285
Periodic function, Fourier expansion of, 62
Laplace transform of, 56–58, 723
Permanent input, 46
Permanent state (see Steady state)
Phase, 48n.
Phase demodulator, 672–674
Phase-lead compensation, 297–307
(See also Lead factor)
Phase margin, 281, 286
Phase plane, 433
with hysteresis, 443n.
use of, for on-off servo, 436–441
for saturated linear servo, 441–443
Phase response, of elementary factors, 143
of first-order system, 75–76
of second-order system, 85
Phase trajectory, 433
Phase velocity, 434
Photoelectric cell, 528
Piezoelectrical instruments, 527–528
Pneumatic amplifier, 680
Poincaré's theorems, 490
Poison dynamo, 524

Pole-zero configuration, 169–177
compensation from, 319–326
frequency response from, 169–170
of second-order system, 109–112
stability from, 160–161
steady state from, 171–172
transient from, 172–174
Position error, 49, 119
for servo system, 226, 260–261
Position piloting (autopiloting), 12–14, 16, 265–267
Potentiometric sensing device, 510, 513–514
(See also Variable-resistance sensing device)
Prediction, 301–302, 329–330
Preliminary design (see Basic design of servo system)
Pressure-head loss, 614
computation of, 617–618
distributed, 616–617
local, 614–615
Proust's method for transient, 145–148
Pulse-width modulating sensing device, 530–531
Pump, 605, 609–610
Pure phase-shift element, 154

Quadrupole, 180–187
electric, 180–182
electromechanical, 186–187
matrix of, 182
mechanical, 184–186
symmetrical, 182
Quadrupole connection, 183
Quadrupole impedances, 183
Quartz, use of, as sensing element, 527–536

Radial-piston motor, 607–608
Ramp response, 48
of first-order system, 71–72
of second-order system, 98
of servo system, 262–263
Random function, 202–203
as input to nonlinear system, 423–426
Random response of power stage, 547–549
Rate-of-climb indicator, 80
Reciprocal networks, 35–37
Rectifier, 671–672
Separate excitation, 569
(See also Armature-controlled motor; Field-controlled motor)
Series excitation, 569
Servo system, basic design (see Basic design of servo system)
components of, 20
definition of, 1, 10–11
eamples of, 11–15
relative character of, 15–17
Servocontrol, 628–631
Servodyne, 660
Servomechanism, 10
Servomotor, 20, 628a.
internal impedance of, 553–554
transfer function of, 550–551, 563–564
(See also Electric servomotor; Hydraulic motor)
Shunt excitation, 569
Signal-flow diagram, 222–223
Simulator, 366–370
Singular points, 484–490
classification of, 485–486
definition of, 485
nature of, 486–487
Sinusoidal tests, 150–151, 403–404, 431
Slip, 595–596
Source, current, 30
voltage, 30
Spectrum, frequency, 201–206
definition of, 203
determination of, 204–206
Spirule, 244a.
Split-series motor, 569, 585
Spoiler, 384, 499–500
Squirrel-cage-type motor, 591
Stability, asymptotic, 159–160, 488
condition for linear, 160–161
conditional, 254, 259, 276
of forced oscillation, 480–481
global, 773–774
local, 773–774
of multiple feedback system, 371–372
of nonlinear servo, 414–417
as performance criterion, 286–287, 288
Russian terminology for, 773–774
of sampled-data system, 349–350, 359–361
structural, 167–168
Stability-accuracy dilemma, 230–231, 295
Stability boundary, dangerous, 418–419, 489–490
Stability criteria, for feedback systems, 268–276, 279–280
for limit cycles, 419–421, 481–482, 491–493
linear, 163–165, 174–175
nonlinear, 493–497
Stability margin, 162, 254
gain, 281, 286
phase, 281, 286
Stabilization, 254–255, 303
of airplane, longitudinal, 277–278
in roll, 255–258, 282–283
friction damper for, 316, 735
Star, 486
Static friction (Coulomb friction), 194, 435–436
Static gain (see Gain)
Steady state, 46
of d-c motor, 574–576
effect of integration on, 118–119
figures of merit for, 49–50, 51
from pole-zero configuration, 171–172
Step response, 47–48, 122–124, 151–152
of first-order system, 69–70
from frequency response, 145–148
from pole-zero configuration, 172–174
from transfer function, 122
Sthène, 738
Strain gauge, 514–515
Structural diagram, 771–772
Structural stability, 167–168
Summation operator, 7
Superposition, principle of, 26a., 375
Synchronisation threshold (see Critical amplitude)
Synchronous motor, 570
use of, as synchro detector, 522–523
Synthesis of servo system, 23–24, 334

Tachometric feedback, 315
Thermoelectric device, 528–530
Threshold, in amplifiers, 643
due to noise, 194, 386–388
do thyatron, 663–665
(See also Dead zone)
Thyatron tube amplifier, 661–665
Time constant, compared to lag, 80–82, 196
due to inductance, 79–80, 577–579
Time constant, of electric servomotor, 82, 577–578, 583–584
of first-order system, 68
determination of, 72, 78
influence of, on transient, 69–70

Time-constant ratio, 299

Toroidal winding, 573

Transfer function, of aircraft in symmetric flight, 127–129
of amplidyne, 657
of armature-controlled motor, 587–589
derivation of, 132–138
determination of, experimental, 150–154

Vernet’s method for, 152–153
of feedback control system, 215, 219–220
open-loop, 215–217
of field-controlled motor, 582–584, 586
of first-order system, 77–78
generalized (see Describing function)
of metadyne, 657
philosophy of, 125–126
of regulator, 221–222
in relation with frequency response, 87, 121–122
of second-order system, 88
of servomotor, conditions for, 563–564
determination of, 550–551
of two-phase motor, 594–595
of Ward-Leonard system, 653

Transfer locus, 132–138
of aircraft in symmetric flight, 132
experimental determination of, 150–154
of first-order system, 77–78
inverse, 138
application of Nyquist’s criterion to, 275–276
gain adjustment using, 232
for second-order system, 95
practical construction of, 142–143
of second-order system, 88–89
(See also Transfer function)

Transfer matrix, 182
of servomotor, 551–552
(See also Quadripole)

Transient, 46
figures of merit for, 50–51
importance of, 46–47
of linear servo, 228–229

Transient, of nonlinear servo, 398–402, 421–422
of on-off servo, 394–398
with dead zone, 394–397
without dead zone, 397–398
(See also Impulse response; Ramp response; Step response)

Transistor, 681–693, 696
arrangements of, 686–689
for a-c amplification, 689–690
for d-c amplification, 690–691
characteristic curves of, 682–684
equations of, 684–686

Transmittance, 116

Tube (see Thyatron-tube amplifier;
Vacuum tube as sensing device;
Vacuum-tube amplifier)

Two-phase motor, 593–600
characteristic curves of, 595–600
construction of, 593–594
transfer function of, 594–595
use of, as generator, 525

Undamped natural frequency, 89, 102
Unit impulse (see Impulse function;
Impulse response)

Unit step (see Step response)

Unlimited stability, 774

Unstable system, 159–161
transfer function of, 150

Vacuum tube as sensing device, 527–528, 531–532
Vacuum-tube amplifier, 644–645, 660–661
Van der Pol’s equation, notions on, 376–379
frequency entrainment, 378
physical applications, 378–379
synchronization, 377–378

Variable-capacitance sensing device, 515–516

Variable-inductance sensing device, 516–523
adjustable-core inductance, 516
saturable-core reactor, 516–517
synchronous transducer, 517–523
magnesy, 520
magnetic bridge, 522
Patin transmission, 521
principle of, 517–518
Variable-inductance sensing device, synchronous transducer, use of, for remote control, 519–520
for remote indication, 518–519
Variable-resistance sensing device, 513–516
potentiometer, 510, 513–514
strain gauge, 514–515
thermistor, 515
Variometer (see Variable-inductance sensing device)
Velocity constant, 120
expression of, from phase response, 141
from pole-zero configuration, 171
from transfer function, 120–121
of servo system, 263
Velocity error, 49, 119–121
of first-order system, 71
of servo system, 262–263
(See also Velocity constant)
Velocity gain, 119
Velocity piloting, 265–267
Verge of instability, behavior on, 418–419, 489–490
Vergeset's method for determining transfer function, 152–153
Ward-Leonard system, 651–654
with compensating windings, 654
machines derived from (see Amplidyne; Metadyne)
transfer function of, 653
White noise, 204
Wiener's method (see RMS-error criterion)
z transfer function, 342
z transform, 341–349
of common functions, 342–345
definition of, 341
properties of, 351–352
Zero (see Pole-zero configuration)