McGRAW-HILL SERIES IN

Control Systems Engineering

SYNTHESIS OF

Optimum Control Systems

SHELDON S. L. CHANG

PROFESSOR OF ELECTRICAL ENGINEERING
NEW YORK UNIVERSITY

McGRAW-HILL BOOK COMPANY, INC.

New York Toronto London

SYNTHESIS OF OPTIMUM CONTROL SYSTEMS

 Library of Congress Catalog Card Number 61-9111

10515 THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

There are two major directions in the development of automatic control systems:

1. Improvements in components
2. Making the best use of the components

The second aspect, which is generally known as system design, is the main topic of this book.

Up until 10 years ago, most literature on control systems regarded components as no more than transfer functions. Admittedly, one of the most essential aspects of a component is its transfer function. Admittedly also, experienced system designers do not regard components as transfer functions alone. They use them merely for studying a closed-loop system's stability and response to medium-sized signals, but in selecting components to be used, a fair amount of attention is also paid to the system's power capacity, stability of its characteristics, internal noise, linearity, etc. While this double-pronged approach—linear stability theory plus common sense—has been highly successful and is still being used by most servo designers, it is not adequate to cope with a complex system in which ultimate response characteristics are desired. To design such a system, experience is insufficient, and experiments are likely to be too slow.

In the past decade, systematic treatments of the optimum-design problem, which considers one or the other limiting features of the so-called fixed (or unalterable) component in addition to its transfer function, have appeared in the literature. In the author's opinion, there are four basic methods in this category:

1. Wiener's least-square optimization with quadratic constraint and its manifestations in nonstationary and sampled-data systems
2. The maximum principle and its forerunner, the "bang-bang" servo principle
3. Self-optimizing systems of various types
4. Computer optimization and control of nonlinear systems

The present volume consists of a treatment of the above four basic methods plus two auxiliary topics:

1. Estimation and measurement of power spectra and correlation functions
2. An analysis of the changes in a closed-loop system's response because of component inaccuracies

The material selected is based on the author's opinion of what additional knowledge is most useful to a servo designer in his work, assuming that he has mastered the basic tools of the trade such as Laplace transforms, Bode plot, Nyquist plot, Evans's root-locus method, etc. This additional knowledge is treated in the simplest possible way (known to the author), but the reader is assumed to be fairly well prepared at least in undergraduate mathematics. Knowledge of contour integration is essential and of probability theory and statistics desirable but not necessary, as the prerequisite knowledge on this subject is treated briefly in Appendixes A and C. Matrix algebra is used in some parts of the book, but a reader can understand most of the book without it.

The book may be used as a graduate text by itself or as an auxiliary text in a graduate course based on one of the standard texts. It will also be useful as a reference book for servo designers and research engineers.

It is the author's pleasure to acknowledge the encouragement and good suggestions from Dean John R. Ragazzini of New York University, Dr. J. G. Truxal of Brooklyn Polytechnic Institute, Professor T. J. Higgins of the University of Wisconsin, and Dr. J. H. Chadwick of the Sperry Gyroscope Company. Many thanks are due to Miss Maryann Regan for typing most of the manuscript and also to Mrs. Mary Rooney, Mrs. Eleanor Gilmore, Mrs. Marie Trotta, and Mrs. Beatrice Schwartz for helping to make various revisions and corrections on the manuscript.

Sheldon S. L. Chang

CONTENTS

CHAPTER 1

INTRODUCTION

1-1. Some Remarks on the Technological Development of Feedback Control Systems. The modern theoretical development of feedback control systems started in the 1920s or 1930s and is marked by Minorsky's paper on the steering of ships (1922), Nyquist's paper Regeneration Theory (1932), and Hazen's paper Theory of Servomechanisms (1934). Before that, the development of feedback controls was mainly in the hands of inventors. While there were isolated instances of successful applications of the concept of feedback control, such as Watt's applications of the flyball governor to the steam engine (1788), Whitehead's torpedo control (1866), and Sperry's gyro stabilizer (1915), there were many, many more attempts that were left unrecorded because they failed. The lack of theory prevented consistent success and economical design toward a prespecified objective.

The theoretical developments made it possible for engineers to design satisfactory feedback control systems as daily routine, using such now classical methods as the Nyquist plot, Bode diagram, Nichols chart, and Evans's root-locus method. However, one common denominator of these design techniques is that they have been developed without critical consideration. Each method leads to one way or another of compensating a system so that it is stable and satisfies a set of more or less arbitrary performance requirements, e.g., rise time, bandwidth, error coefficients, peak overshoot, etc. As these requirements can be satisfied in many ways, the selection of system configuration as well as time constants of the compensating networks is left largely to the discretion or experience of the designer. There is no place in the above-mentioned techniques for many factors which are known by experience to be significant, e.g., the torque-to-inertia ratio of a servomotor, the noise in the sensing elements, etc. Consequently, the question of what makes the best system under actual operating environment and component limitations cannot be answered by these techniques alone.

In the past one or two decades, while the classical design techniques have been reaching their fruition, the trend of research work has veered toward optimization. This has been for good reason: The problem is no longer how to design one of many systems that work but to design the system that works best. In many applications, notably fire-control and inertial-guidance systems, no degree of accuracy is too good; in missile-steering systems, no response is too fast. When we try to improve the accuracy and speed of response of a system, the ultimate limits, which are reached sooner or later, are noise and saturation. Any design technique, if it is to be realistic, must take due account of these factors. The outstanding work in this direction is represented by Wiener's theory of optimum filtering and prediction. However, as Wiener's theory was not intended originally for control systems, it did not give due consideration to the power limitations of the components. This significant addition to the theory was due to Newton. Other developments of wide engineering implications include the phase-plane technique or predictor control of

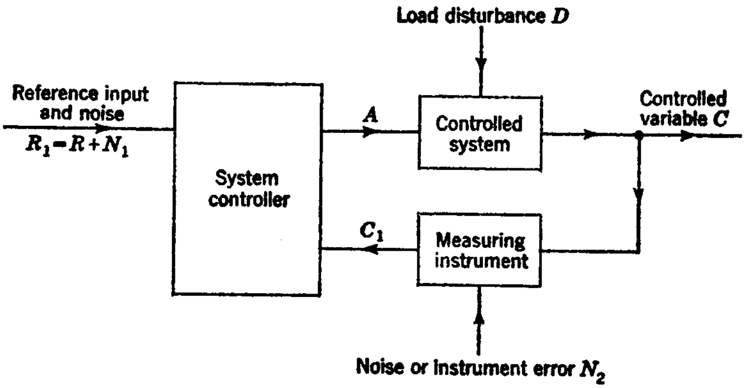

FIG. 1-1. Basic components in a feedback control system.

systems with simple saturation, the development of sampled-data systems, and the recent developments in self-optimizing systems. The last is perhaps the culminating point of the trend toward optimization. An ideal self-optimizing system learns about its environment and adjusts itself to optimum expected performance in a continual process of measuring and adjusting. While developments in this area are still in their infancy, the importance of the self-optimizing concept cannot be overemphasized.

1-2. The Given Conditions in a Servo-design Problem. Perhaps the compelling need or reason for the recent developments and trends in servo theory can be understood by an examination of the problems facing a servo designer today. In doing so, we shall also define a number of terms that will be useful later.

A general representation of a control system is shown in Fig. 1-1. The controlled system may be a stabilized platform including a torque motor, a steering system including the control surface and its hydraulic drive, or a combination of a servomotor, gears, and marking pen. We assume that we may install instruments to measure whichever system variable we like. However, there is always the error or noise in measurement. The reference input may be transmitted from remote places, with unavoidable noise in the transmission link. Finally there are disturbing forces and moments applied to the load system. The elements possess varying degrees of inalterability. While a bigger and faster hydraulic drive, a better torque-to-inertia ratio in a motor-driven system, or a better signal-to-noise ratio in a measuring instrument is sometimes possible by using more expensive hardware, there is a limit on what one can do beyond any cost consideration. Furthermore, in order not to use expensive hardware unnecessarily, there is a point in investigating what is the best one can do with the given hardware.

By contrast, the system controller with its amplifiers and compensating networks, linear or nonlinear, is almost entirely up to the designer. While an increase in gain that requires an additional stage amplifier incurs some added expense, it is nowhere nearly comparable to the cost of a bigger hydraulic system. The noise in well-designed electronics is usually an order of magnitude lower than that of the measuring instruments and does not constitute a valid limitation on system gain. The saturation of amplifiers, if it is by noise of various types, can usually be alleviated by proper prefiltering. If an amplifier is saturated by an actuating error signal before the controlled system is saturated, economic considerations usually dictate a change of the setup. We can safely say that, at least in systems where high performance is at a premium, no limitation of any kind should be imposed on the controller itself.

Thus we may classify the elements in a control system into three categories:

1. The controlled system, which is also called the plant, the fixed component, or the unalterable component, meaning that its choice is usually not up to the servo designer
2. The measuring or sensing elements
3. The system controller

The first two items are more or less given, and a servo designer's job is essentially to design the system controller itself. However, that is only the first step. If we find it possible to design a controller to meet some given performance specification, the problem does not stop there. We are usually asked what is the best one can do with the given hardware (items 1 and 2). Sometimes we find it very difficult to meet the specified performance; then we must be prepared to answer the question: Is it at all possible to meet the specified performance with the hardware on hand?

Or, what improvement in the hardware is necessary in order to obtain the specified performance?

1-3. Some Classical Fallacies. These questions are difficult or impossible to answer if one knows only the classical techniques or theory. Compounding this, we shall see that some classical notions of performance and of what one can do with a system no longer represent the absolute truth. We shall confine our discussion to three areas:

1. System response to input
2. System response to load disturbance
3. Effect of variations in plant transfer function

Perhaps the best-known notion is the following: *The most desirable system response is the one with the shortest rise time and adequate damping.*

The above is no longer true when noise is considered. A shorter rise time also implies a larger passband for noise, which increases the over-all error. Another point is that what one can obtain on paper is different from what one would obtain in an actual system, because of plant saturation and a number of other factors. A shorter rise time usually means larger transient as well as noise input to the plant for the same reference input. Once the plant is saturated, the system is likely to be more sluggish than one which has a longer rise time on paper but operates in the linear range. A third point is that two systems with widely different pole-zero locations can have the same rise time and peak overshoot but a different order of magnitude of peak value of transient plant input. Considering the small-signal transient response alone is not enough, since a system with a lower peak value of plant input performs better when the reference input signal is large.

The best response to load disturbance is the stiffest one. In other words, the most undisturbable system is the best. This is not always true, depending on the available torque, rate, or power of the controlled system. For instance, in a platform-stabilization system we nearly always provide sufficient torque in the torque motor to balance out the disturbing moments. This is not so in the roll stabilization of an airplane or a ship and is even less so in a steering or attitude-control system. Consequently we try to balance out the disturbances as much as possible in a platform-stabilization system, the only limit to system bandwidth being instrument noise. In the roll stabilization of a ship we try to balance out nearly all the disturbing roll moment due to ocean waves when the sea state is not very heavy, so that the passengers can enjoy their voyage. However, in a really stormy sea, the roll moment is many times larger than the available stabilizing moment of the fins (or activated tanks, etc.), and all one can do is to use the fins to damp out the predominant resonant mode of the ship response to lessen the danger of capsizing. If the servo is designed with unduly large bandwidth, it would simply be jammed and would not serve any useful purpose at all. In steering and attitude-control systems, the rudder or control surface is used to keep a mean course or attitude, as its moment and speed of movement are not adequate to keep up with the instantaneous disturbances due to ocean waves or atmospheric turbulence.

From the frequency-response point of view, because of load inertia and other integration or resonant effects, some Fourier components of the load disturbance cause far more change in the controlled variable than others. When the stabilizing forces or moments at our disposal are limited, our problem is to design not the stiffest system but one which is properly selective, so that these forces or moments can be most effectively utilized.

The closed-loop response of a control system can be made as independent of plant transfer characteristics as one likes by using shunt compensation and increasing the gain of the inner loop. The problem of the sensitivity of closed-loop response to plant variations is one of the greatest current interest. An example of its practical application is the control of a ballistic missile. Within a relatively short time of less than a minute the Mach number changes from zero to 10 or higher and then to something undefined as the air pressure changes from 1 atm to practically zero. During the same period the mass changes as much as 10:1, with a corresponding shift of the center of mass. Obviously there is a drastic change in plant transfer characteristics; however, the closed-loop system response is required not only to be well damped but also to stay close to some specified performance at all times.

At first glance, one would think that the problem could be readily solved by using shunt compensaton with a large inner-loop gain. The inner loop, can be stabilized by introducing a sufficient number of zeros in the vicinity of the plant poles and keeping the poles of the compensating network far enough to the left. A close examination reveals that there is a definite relationship between the sensitivity function and the system's responses to load disturbance and instrument noise. This, together with the possible existence of transportation lag or nonminimum-phase effect in the plant, imposes a limit on how far we can go with inner-loop feedback.

Various ingenious schemes such as conditional feedback or the use of a model have been suggested, tried out, and published in the literature; however, from the analytical point of view, they are not different from the simple series-shunt compensation system. These points will become obvious from a discussion in the next section.

1-4. Basic Relations of Linearly Compensated Systems. With reference to Fig. 1-1, a control system is called linearly compensated if the controller is linear. It does not matter whether the controlled system and measuring instrument are linear or not. In other words, we can write an equation between R_1, C_1, and A:

$$A(s) = H_1(s)R_1(s) - H_2(s)C_1(s) \tag{1-1}$$

where s is the Laplace-transform variable.

The above definition of a linearly compensated system can be generalized to include systems that have nonlinear elements in cascade with the controlled system, since these cascade elements can be viewed as part of the controlled system.

Our first basic relationship is that of equivalence: A linear system controller is completely specified by the transfer functions $H_1(s)$ and $H_2(s)$. Two system controllers with different configurations but the same $H_1(s)$

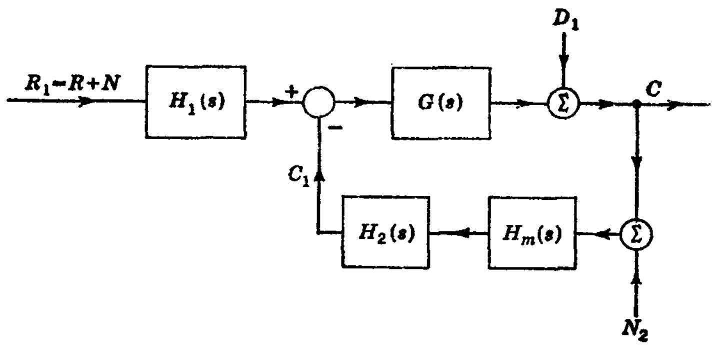

FIG. 1-2. Block-diagram representation of a linear system.

and $H_2(s)$ are completely equivalent as far as every aspect of the system response is concerned (such as stability and responses to input, to noise, to load disturbance, etc.).

When the problem is formulated in the form of Fig. 1-1, the equivalence relation is almost self-evident. The controlled system and measuring instrument are described by an equation relating $c_1(t)$ to all previous values of $d(t)$, $n_2(t)$, and $a(t)$:

$$c_1(t) = f[d(t'),n_2(t''),a(t'''),t] \tag{1-2}$$

There is no need to discuss Eq. (1-2) except to note that it is completely independent of the system controller. Because Eqs. (1-1) and (1-2) determine $a(t)$, and consequently $c(t)$, the dependence of the controlled variable $c(t)$ on the system controller is only through the network functions $H_1(s)$ and $H_2(s)$. This proves the equivalence relation.

In case the controlled system and measuring instruments are linear, a basic relationship exists between the sensitivity function, system response to load disturbance, and system response to instrument noise. It can be derived as follows:

In Fig. 1-2, $G(s)$ and $H_m(s)$ are transfer functions of the controlled system and measuring instrument, respectively. The load disturbance is represented by its equivalent value $d_1(t)$ at the output end. Physically, $d_1(t)$ is the output variable if the control is completely cut off. The noise in the measuring instrument is represented by its equivalent value $n_2(t)$ of the measured variable. The closed-loop transfer functions C/R, C/D_1, and C/N_2 are defined as the response in C due to R, D_1, and N_2 alone. From an inspection of Fig. 1-2, we have

$$\frac{C}{R}(s) = \frac{H_1(s)G(s)}{1 + H_2(s)H_m(s)G(s)} \tag{1-3}$$

$$\frac{C}{D_1}(s) = \frac{1}{1 + H_2(s)H_m(s)G(s)} \tag{1-4}$$

$$\frac{C}{N_2}(s) = \frac{-H_2(s)H_m(s)G(s)}{1 + H_2(s)H_m(s)G(s)} \tag{1-5}$$

For an infinitesimal variation in $G(s)$ or $\Delta G(s)$, Eq. (1-3) gives

$$\frac{\Delta \dfrac{C}{R}(s)}{\dfrac{C}{R}(s)} = \frac{\Delta G(s)}{G(s)} \frac{1}{1 + H_2(s)H_m(s)G(s)} \tag{1-6}$$

The sensitivity function S is defined as

$$S = \frac{\Delta \dfrac{C}{R}(s) \Big/ \dfrac{C}{R}(s)}{\Delta G(s)/G(s)} \tag{1-7}$$

It represents the ratio of per-unit change in C/R to per-unit change in G and gives a quantitative measure of the dependence of the system transfer function to the plant transfer function. Equations (1-4) to (1-7) give

$$S = \frac{C}{D_1}(s) = 1 + \frac{C}{N_2}(s) \tag{1-8}$$

Because of the equivalence relation, Eq. (1-8) is completely independent of the system configuration. It holds as long as the system is linear.

Equations (1-3) and (1-4) imply that the transfer function C/R and C/D_1 can be independently selected. For any given $G(s)$, we can find H_1 and H_2 to give the desired C/R and C/D. However, Eq. (1-8) shows that the sensitivity of C/R to variations in G is closely tied to C/D_1 and C/N_2. In order to make C/R insensitive to variations in G, the loop gain

H_2H_mG must be maintained at some fairly high value for all possible values of G. A sketch of the various gain functions is given in Fig. 1-3. The C/R function is approximately H_1/H_2H_m. The loop gain (H_2H_mG) varies as G varies and must be allowed to decrease gradually beyond C/R on account of system stability. Consequently, the system bandwidth to instrument noise is considerably higher than that of C/R, and the system's stiffness to load disturbance is also necessarily high over a wide band. It is obvious that how far the noise bandwidth extends depends on how insensitive to plant variations the C/R function is made to be.

Thus we see that instrument noise and load disturbance are limiting factors to the degree of insensitivity that one can achieve in the system response function to plant variations. Of course there are other limiting

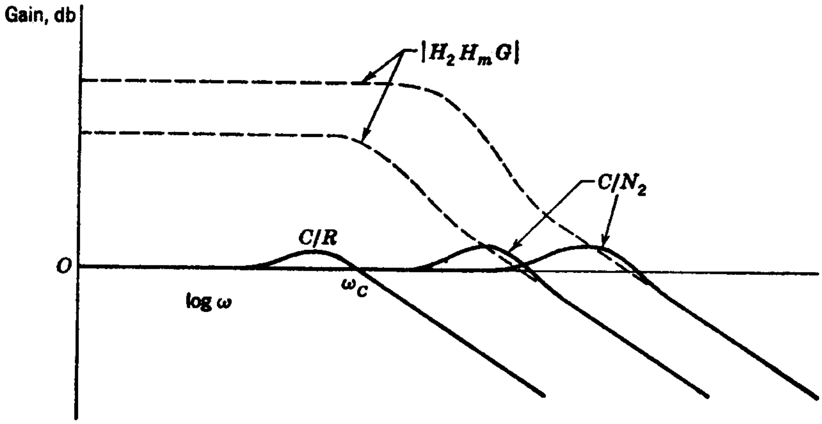

Fig. 1-3. Gain versus log ω of various transfer functions in a system with changing plant.

factors, e.g., transportation lag and high-frequency resonant modes of the plant. While these high-frequency effects are not very significant in the passband of C/R itself, they nevertheless limit the extended gain bandwidth of H_mH_2G.

1-5. Scope of This Book. The above is an illustration of the various factors entering into a control problem and how they are interrelated. Present control theory does not give a cookbook solution of every problem that may arise, but it does provide a number of basic tools for solving these problems. To be more explicit, we have a number of idealized situations for which exact mathematical solutions are feasible. Sometimes a control problem falls within one of these, and we have an immediate solution. However, many problems do not fall into one idealized situation alone, and a direct mathematical solution is either too cumbersome or impossible with our present knowledge of mathematics. The