

1.1 Introduction

Control systems are an integral part of modern society. Numerous applications are all around us: The rockets fire, and the space shuttle lifts off to earth orbit; in splashing cooling water, a metallic part is automatically machined; a self-guided vehicle delivering material to workstations in an aerospace assembly plant glides along the floor seeking its destination. These are just a few examples of the automatically controlled systems that we can create.

We are not the only creators of automatically controlled systems; these systems also exist in nature. Within our own bodies are numerous control systems, such as the pancreas, which regulates our blood sugar. In time of “fight or flight,” our adrenaline increases along with our heart rate, causing more oxygen to be delivered to our cells. Our eyes follow a moving object to keep it in view; our hands grasp the object and place it precisely at a predetermined location.

Even the nonphysical world appears to be automatically regulated. Models have been suggested showing automatic control of student performance. The input to the model is the student’s available study time, and the output is the grade. The model can be used to predict the time required for the grade to rise if a sudden increase in study time is available. Using this model, you can determine whether increased study is worth the effort during the last week of the term.

Control System Definition

A control system consists of *subsystems* and *processes* (or *plants*) assembled for the purpose of obtaining a desired *output* with desired *performance*, given a specified *input*. Figure 1.1 shows a control system in its simplest form, where the input represents a desired output.

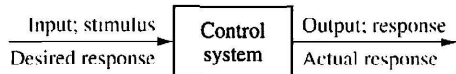


FIGURE 1.1 Simplified description of a control system

For example, consider an elevator. When the fourth-floor button is pressed on the first floor, the elevator rises to the fourth floor with a speed and floor-leveling accuracy designed for passenger comfort. The push of the fourth-floor button is an *input* that represents our desired *output*, shown as a step function in Figure 1.2. The *performance* of the elevator can be seen from the elevator response curve in the figure.

Two major measures of performance are apparent: (1) the transient response and (2) the steady-state error. In our example, passenger comfort and passenger patience are dependent upon the transient response. If this response is too fast, passenger comfort is sacrificed; if too slow, passenger patience is sacrificed. The steady-state error is another important performance specification since passenger safety and convenience would be sacrificed if the elevator did not properly level.

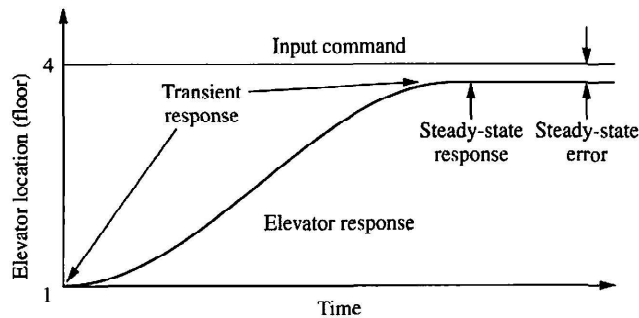


FIGURE 1.2 Elevator response

compensator parameters required to yield a desired response can be made by changes in software rather than hardware. The computer can also perform supervisory functions, such as scheduling many required applications. For example, the space shuttle main engine (SSME) controller, which contains two digital computers, alone controls numerous engine functions. It monitors engine sensors that provide pressures, temperatures, flow rates, turbopump speed, valve positions, and engine servo valve actuator positions. The controller further provides closed-loop control of thrust and propellant mixture ratio, sensor excitation, valve actuators, spark igniters, as well as other functions (Rockwell International, 1984).

1.4 Analysis and Design Objectives

In Section 1.1 we briefly alluded to some control system performance specifications, such as transient response and steady-state error. We now expand upon the topic of performance and place it in perspective as we define our analysis and design objectives.

Analysis is the process by which a system's performance is determined. For example, we evaluate its transient response and steady-state error to determine if they meet the desired specifications. *Design* is the process by which a system's performance is created or changed. For example, if a system's transient response and steady-state error are analyzed and found not to meet the specifications, then we change parameters or add additional components to meet the specifications.

A control system is *dynamic*: It responds to an input by undergoing a transient response before reaching a steady-state response that generally resembles the input. We have already identified these two responses and cited a position control system (an elevator) as an example. In this section, we discuss three major objectives of systems analysis and design: **producing the desired transient response, reducing steady-state error, and achieving stability.** We also address some other design concerns, such as cost and the sensitivity of system performance to changes in parameters.

Transient Response

Transient response is important. In the case of an elevator, a slow transient response makes passengers impatient, whereas an excessively rapid response makes them uncomfortable. If the elevator oscillates about the arrival floor for more than a second, a disconcerting feeling can result. Transient response is also important for structural reasons: Too fast a transient response could cause permanent physical damage. In a computer, transient response contributes to the time required to read from or write to the computer's disk storage (see Figure 1.7). Since reading and writing cannot take place until the head stops, the speed of the read/write head's movement from one track on the disk to another influences the overall speed of the computer.



FIGURE 1.7 Computer hard disk drive, showing disks and read/write head

In this book, we establish quantitative definitions for transient response. We then analyze the system for its *existing* transient response. Finally, we adjust parameters or design components to yield a *desired* transient response—our first analysis and design objective.

Steady-State Response

Another analysis and design goal focuses on the steady-state response. As we have seen, this response resembles the input and is usually what remains after the transients have decayed to zero. For example, this response may be an elevator stopped near the fourth floor or the head of a disk drive finally stopped at the correct track. We are concerned about the accuracy of the steady-state response. An elevator must be level enough with the floor for the passengers to exit, and a read/write head not positioned over the commanded track results in computer errors. An antenna tracking a satellite must keep the satellite well within its beamwidth in order not to lose track. In this text we define steady-state errors quantitatively, analyze a system's steady-state error, and then design corrective action to reduce the steady-state error—our second analysis and design objective.

Stability

Discussion of transient response and steady-state error is moot if the system does not have *stability*. In order to explain stability, we start from the fact that the total response of a system is the sum of the *natural response* and the *forced response*. When you studied linear differential equations, you probably referred to these responses as the *homogeneous and the particular solutions, respectively*. Natural response describes the way the system dissipates or acquires energy. The form or nature of this response is dependent only on the system, not the input. On the other hand, the form or nature of the forced response is dependent on the input. Thus, for a *linear* system, we can write

$$\text{Total response} = \text{Natural response} + \text{Forced response} \quad (1.1)^2$$

For a control system to be useful, the natural response must (1) eventually approach zero, thus leaving only the forced response, or (2) oscillate. In some systems, however, the natural response grows without bound rather than diminish to zero or oscillate. Eventually, the natural response is so much greater than the forced response that the system is no longer controlled. This condition, called *instability*, could lead to self-destruction of the physical device if limit stops are not part of the design. For example, the elevator would crash through the floor or exit through the ceiling; an aircraft would go into an uncontrollable roll; or an antenna commanded to point to a target would rotate, line up with the target, but then begin to oscillate about the target with *growing* oscillations and *increasing* velocity until the motor or amplifiers reached their output limits or until the antenna was damaged structurally. A time plot of an unstable system would show a transient response that grows without bound and without any evidence of a steady-state response.

Control systems must be designed to be stable. That is, their natural response must decay to zero as time approaches infinity, or oscillate. In many systems the transient response you see on a time response plot can be directly related to the natural response. Thus, if the natural response decays to zero as time approaches infinity, the transient response will also die out, leaving only the forced response. If the system is stable, the proper transient response and steady-state error characteristics can be designed. Stability is our third analysis and design objective.

² You may be confused by the words *transient* vs. *natural*, and *steady-state* vs. *forced*. If you look at Figure 1.2, you can see the transient and steady-state portions of the total response as indicated. The transient response is the sum of the natural and forced responses, while the natural response is large. If we plotted the natural response by itself, we would get a curve that is different from the transient portion of Figure 1.2. The steady-state response of Figure 1.2 is also the sum of the natural and forced responses, but the natural response is small. Thus, the transient and steady-state responses are what you actually see on the plot; the natural and forced responses are the underlying mathematical components of those responses.

Other Considerations

The three main objectives of control system analysis and design have already been enumerated. However, other important considerations must be taken into account. For example, factors affecting hardware selection, such as motor sizing to fulfill power requirements and choice of sensors for accuracy, must be considered early in the design.

Finances are another consideration. Control system designers cannot create designs without considering their economic impact. Such considerations as budget allocations and competitive pricing must guide the engineer. For example, if your product is one of a kind, you may be able to create a design that uses more expensive components without appreciably increasing total cost. However, if your design will be used for many copies, slight increases in cost per copy can translate into many more dollars for your company to propose during contract bidding and to outlay before sales.

Another consideration is *robust* design. System parameters considered constant during the design for transient response, steady-state errors, and stability change over time when the actual system is built. Thus, the performance of the system also changes over time and will not be consistent with your design. Unfortunately, the relationship between parameter changes and their effect on performance is not linear. In some cases, even in the same system, changes in parameter values can lead to small or large changes in performance, depending on the system's nominal operating point and the type of design used. Thus, the engineer wants to create a robust design so that the system will not be sensitive to parameter changes. We discuss the concept of system sensitivity to parameter changes in Chapters 7 and 8. This concept, then, can be used to test a design for robustness.

Case Study

Introduction to a Case Study

Now that our objectives are stated, how do we meet them? In this section we will look at an example of a feedback control system. The system introduced here will be used in subsequent chapters as a running case study to demonstrate the objectives of those chapters. A colored background like this will identify the case study section at the end of each chapter. Section 1.5, which follows this first case study, explores the design process that will help us build our system.

Antenna Azimuth: An Introduction to Position Control Systems

A position control system converts a position input command to a position output response. Position control systems find widespread applications in antennas, robot arms, and computer disk drives. The radio telescope antenna in Figure 1.8 is one example of a system that uses position control systems. In this section, we will look in detail at an antenna azimuth position control system that could be used to position a radio telescope antenna. We will see how the system works and how we can effect changes in its performance. The discussion here will be on a qualitative level, with the objective of getting an intuitive feeling for the systems with which we will be dealing.

An antenna azimuth position control system is shown in Figure 1.9(a), with a more detailed layout and schematic in Figures 1.9(b) and 1.9(c), respectively. Figure 1.9(d) shows a *functional block diagram* of the system. The functions are shown above the blocks, and the required hardware is indicated inside the blocks. Parts of Figure 1.9 are repeated on the front endpapers for future reference.

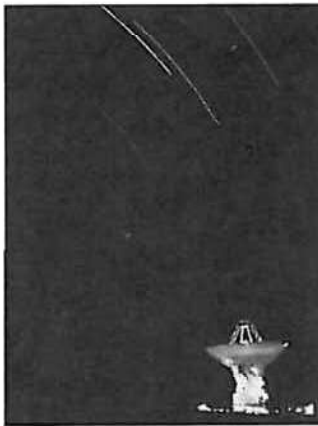


FIGURE 1.8 The search for extraterrestrial life is being carried out with radio antennas like the one pictured here. A radio antenna is an example of a system with position controls.

function can be used only for linear systems, it yields more intuitive information than the *differential equation*. We will be able to change system parameters and rapidly sense the effect of these changes on the system response. The transfer function is also useful in modeling the interconnection of subsystems by forming a block diagram similar to Figure 1.9(d) but with a mathematical function inside each block.

Still another model is the *state-space representation*. One advantage of state-space methods is that they can also be used for systems that cannot be described by linear differential equations. Further, state-space methods are used to model systems for simulation on the digital computer. Basically, this representation turns an n th-order differential equation into n simultaneous first-order differential equations. Let this description suffice for now; we describe this approach in more detail in Chapter 3.

Finally, we should mention that to produce the mathematical model for a system, we require knowledge of the parameter values, such as equivalent resistance, inductance, mass, and damping, which is often not easy to obtain. Analysis, measurements, or specifications from vendors are sources that the control systems engineer may use to obtain the parameters.

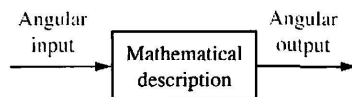


FIGURE 1.12 Equivalent block diagram for the antenna azimuth position control system

Step 5: Reduce the Block Diagram

Subsystem models are interconnected to form block diagrams of larger systems, as in Figure 1.9(d), where each block has a mathematical description. Notice that many signals, such as proportional voltages and error, are internal to the system. There are also two signals—angular input and angular output—that are external to the system. In order to evaluate system response in this example, we need to reduce this large system's block diagram to a single block with a mathematical description that represents the system from its input to its output, as shown in Figure 1.12. Once the block diagram is reduced, we are ready to analyze and design the system.

Step 6: Analyze and Design

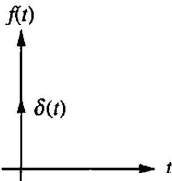
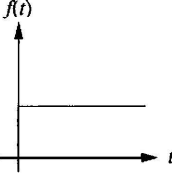
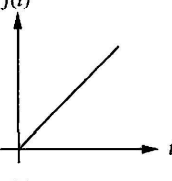
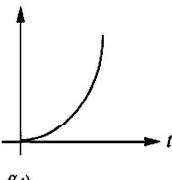
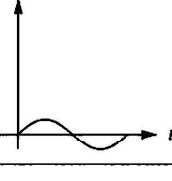
The next phase of the process, following block diagram reduction, is analysis and design. If you are interested only in the performance of an individual subsystem, you can skip the block diagram reduction and move immediately into analysis and design. In this phase, the engineer analyzes the system to see if the response specifications and performance requirements can be met by simple adjustments of system parameters. If specifications cannot be met, the designer then designs additional hardware in order to effect a desired performance.

Test input signals are used, both analytically and during testing, to verify the design. It is neither necessarily practical nor illuminating to choose complicated input signals to analyze a system's performance. Thus, the engineer usually selects standard test inputs. These inputs are impulses, steps, ramps, parabolas, and sinusoids, as shown in Table 1.1.

An *impulse* is infinite at $t = 0$ and zero elsewhere. The area under the unit impulse is 1. An approximation of this type of waveform is used to place initial energy into a system so that the response due to that initial energy is only the transient response of a system. From this response the designer can derive a mathematical model of the system.

A *step* input represents a *constant command*, such as position, velocity, or acceleration. Typically, the step input command is of the same form as the output. For example, if the system's output is position, as it is for the antenna azimuth position control system, the **step input represents** a desired position, and the output represents the actual position. If the system's output is velocity, as is the spindle speed for a video disc player, the step input represents a constant desired speed, and the output represents the actual speed. **The designer uses step inputs because both the transient response and the steady-state response are clearly visible and can be evaluated.**

TABLE 1.1 Test waveforms used in control systems

Input	Function	Description	Sketch	Use
Impulse	$\delta(t)$	$\delta(t) = \infty$ for $0- < t < 0+$ $= 0$ elsewhere $\int_{0-}^{0+} \delta(t) dt = 1$		Transient response Modeling
Step	$u(t)$	$u(t) = 1$ for $t > 0$ $= 0$ for $t < 0$		Transient response Steady-state error
Ramp	$tu(t)$	$tu(t) = t$ for $t \geq 0$ $= 0$ elsewhere		Steady-state error
Parabola	$\frac{1}{2}t^2u(t)$	$\frac{1}{2}t^2u(t) = \frac{1}{2}t^2$ for $t \geq 0$ $= 0$ elsewhere		Steady-state error
Sinusoid	$\sin \omega t$			Transient response Modeling Steady-state error

The *ramp* input represents a *linearly increasing command*. For example, if the system's output is position, the input ramp represents a linearly increasing position, such as that found when tracking a satellite moving across the sky at constant speed. If the system's output is velocity, the input ramp represents a linearly increasing velocity. The response to an input ramp test signal yields additional information about the steady-state error. The previous discussion can be extended to *parabolic* inputs, which are also used to evaluate a system's steady-state error.

Sinusoidal inputs can also be used to test a physical system to arrive at a mathematical model. We discuss the use of this waveform in detail in Chapters 10 and 11.

We conclude that one of the basic analysis and design requirements is to evaluate the time response of a system for a given input. Throughout the book you will learn numerous methods for accomplishing this goal.

The control systems engineer must take into consideration other characteristics about feedback control systems. For example, control system behavior is altered by fluctuations in component values or system parameters. These variations can be

6.1 Introduction

In Chapter 1, we saw that three requirements enter into the design of a control system: transient response, stability, and steady-state errors. Thus far we have covered transient response, which we will revisit in Chapter 8. We are now ready to discuss the next requirement, stability.

Stability is the most important system specification. If a system is unstable, transient response and steady-state errors are moot points. An unstable system cannot be designed for a specific transient response or steady-state error requirement. What, then, is stability? There are many definitions for stability, depending upon the kind of system or the point of view. In this section, we limit ourselves to linear, time-invariant systems.

In Section 1.5, we discussed that we can control the output of a system if the steady-state response consists of only the forced response. But the total response of a system is the sum of the forced and natural responses, or

$$c(t) = c_{\text{forced}}(t) + c_{\text{natural}}(t) \quad (6.1)$$

Using these concepts, we present the following definitions of stability, instability, and marginal stability:

A linear, time-invariant system is *stable* if the natural response approaches zero as time approaches infinity.

A linear, time-invariant system is *unstable* if the natural response grows without bound as time approaches infinity.

A linear, time-invariant system is *marginally stable* if the natural response neither decays nor grows but remains constant or oscillates as time approaches infinity.

Thus, the definition of stability implies that only the forced response remains as the natural response approaches zero.

These definitions rely on a description of the natural response. When one is looking at the total response, it may be difficult to separate the natural response from the forced response. However, we realize that if the input is bounded and the total response is not approaching infinity as time approaches infinity, then the natural response is obviously not approaching infinity. If the input is unbounded, we see an unbounded total response, and we cannot arrive at any conclusion about the stability of the system; we cannot tell whether the total response is unbounded because the forced response is unbounded or because the natural response is unbounded. Thus, our alternate definition of *stability*, one that regards the total response and implies the first definition based upon the natural response, is this:

A system is stable if *every* bounded input yields a bounded output.

We call this statement the bounded-input, bounded-output (BIBO) definition of stability.

Let us now produce an alternate definition for instability based on the total response rather than the natural response. We realize that if the input is bounded but the total response is unbounded, the system is unstable, since we can conclude that the natural response approaches infinity as time approaches infinity. If the input is unbounded, we will see an unbounded total response, and we cannot draw any conclusion about the stability of the system; we cannot tell whether the total response is unbounded because the forced response is unbounded or because the

natural response is unbounded. Thus, our alternate definition of *instability*, one that regards the total response, is this:

A system is unstable if *any* bounded input yields an unbounded output.

These definitions help clarify our previous definition of *marginal stability*, which really means that the system is stable for some bounded inputs and unstable for others. For example, we will show that if the natural response is undamped, a bounded sinusoidal input of the same frequency yields a natural response of growing oscillations. Hence, the system appears stable for all bounded inputs except this one sinusoid. Thus, marginally stable systems by the natural response definitions are included as unstable systems under the BIBO definitions.

Let us summarize our definitions of stability for linear, time-invariant systems. Using the natural response:

1. A system is stable if the natural response approaches zero as time approaches infinity.
2. A system is unstable if the natural response approaches infinity as time approaches infinity.
3. A system is marginally stable if the natural response neither decays nor grows but remains constant or oscillates.

Using the total response (BIBO):

1. A system is stable if *every* bounded input yields a bounded output.
2. A system is unstable if *any* bounded input yields an unbounded output.

Physically, an unstable system whose natural response grows without bound can cause damage to the system, to adjacent property, or to human life. Many times systems are designed with limited stops to prevent total runaway. From the perspective of the time response plot of a physical system, instability is displayed by transients that grow without bound and, consequently, a total response that does not approach a steady-state value or other forced response.¹

How do we determine if a system is stable? Let us focus on the natural response definitions of stability. Recall from our study of system poles that poles in the left half-plane (lhp) yield either pure exponential decay or damped sinusoidal natural responses. These natural responses decay to zero as time approaches infinity. Thus, if the closed-loop system poles are in the left half of the plane and hence have a negative real part, the system is stable. That is, *stable systems have closed-loop transfer functions with poles only in the left half-plane.*

Poles in the right half-plane (rhp) yield either pure exponentially increasing or exponentially increasing sinusoidal natural responses. These natural responses approach infinity as time approaches infinity. Thus, if the closed-loop system poles are in the right half of the s -plane and hence have a positive real part, the system is unstable. Also, poles of multiplicity greater than 1 on the imaginary axis lead to the sum of responses of the form $At^n \cos(\omega t + \phi)$, where $n = 1, 2, \dots$, which also approaches infinity as time approaches infinity. Thus, *unstable systems have closed-loop transfer functions with at least one pole in the right half-plane and/or poles of multiplicity greater than 1 on the imaginary axis.*

¹ Care must be taken here to distinguish between natural responses growing without bound and a forced response, such as a ramp or exponential increase, that also grows without bound. A system whose forced response approaches infinity is stable as long as the natural response approaches zero.



Transient and Steady-State Response Analyses

5-1 INTRODUCTION

It was stated in Chapter 3 that the first step in analyzing a control system was to derive a mathematical model of the system. Once such a model is obtained, various methods are available for the analysis of system performance.

In practice, the input signal to a control system is not known ahead of time but is random in nature, and the instantaneous input cannot be expressed analytically. Only in some special cases is the input signal known in advance and expressible analytically or by curves, such as in the case of the automatic control of cutting tools.

In analyzing and designing control systems, we must have a basis of comparison of performance of various control systems. This basis may be set up by specifying particular test input signals and by comparing the responses of various systems to these input signals.

Many design criteria are based on the response to such signals or on the response of systems to changes in initial conditions (without any test signals). The use of test signals can be justified because of a correlation existing between the response characteristics of a system to a typical test input signal and the capability of the system to cope with actual input signals.

Typical Test Signals. The commonly used test input signals are those of step functions, ramp functions, acceleration functions, impulse functions, sinusoidal functions, and the like. With these test signals, mathematical and experimental analyses of control systems can be carried out easily since the signals are very simple functions of time.

Which of these typical input signals to use for analyzing system characteristics may be determined by the form of the input that the system will be subjected to most frequently under normal operation. If the inputs to a control system are gradually changing functions of time, then a ramp function of time may be a good test signal. Similarly, if a system is subjected to sudden disturbances, a step function of time may be a good test signal; and for a system subjected to shock inputs, an impulse function may be best. Once a control system is designed on the basis of test signals, the performance of the system in response to actual inputs is generally satisfactory. The use of such test signals enables one to compare the performance of all systems on the same basis.

Transient Response and Steady-State Response. The time response of a control system consists of two parts: the transient response and the steady-state response. By transient response, we mean that which goes from the initial state to the final state. By steady-state response, we mean the manner in which the system output behaves as t approaches infinity. Thus the system response $c(t)$ may be written as

$$c(t) = c_{tr}(t) + c_{ss}(t)$$

where the first term on the right-hand side of the equation is the transient response and the second term is the steady-state response.

Absolute Stability, Relative Stability, and Steady-State Error. In designing a control system, we must be able to predict the dynamic behavior of the system from a knowledge of the components. The most important characteristic of the dynamic behavior of a control system is absolute stability, that is, whether the system is stable or unstable. A control system is in equilibrium if, in the absence of any disturbance or input, the output stays in the same state. A linear time-invariant control system is stable if the output eventually comes back to its equilibrium state when the system is subjected to an initial condition. A linear time-invariant control system is critically stable if oscillations of the output continue forever. It is unstable if the output diverges without bound from its equilibrium state when the system is subjected to an initial condition. Actually, the output of a physical system may increase to a certain extent but may be limited by mechanical "stops," or the system may break down or become nonlinear after the output exceeds a certain magnitude so that the linear differential equations no longer apply.

Important system behavior (other than absolute stability) to which we must give careful consideration includes relative stability and steady-state error. Since a physical control system involves energy storage, the output of the system, when subjected to an input, cannot follow the input immediately but exhibits a transient response before a steady state can be reached. The transient response of a practical control system often exhibits damped oscillations before reaching a steady state. If the output of a system at steady state does not exactly agree with the input, the system is said to have steady-state error. This error is indicative of the accuracy of the system. In analyzing a control system, we must examine transient-response behavior and steady-state behavior.

Outline of the Chapter. This chapter is concerned with system responses to aperiodic signals (such as step, ramp, acceleration, and impulse functions of time). The outline of the chapter is as follows: Section 5-1 has presented introductory material for the chapter. Section 5-2 treats the response of first-order systems to aperiodic inputs. Section 5-3 deals with the transient response of the second-order systems. Detailed

analyses of the step response, ramp response, and impulse response of the second-order systems are presented. Section 5-4 discusses the transient response analysis of higher-order systems. Section 5-5 gives an introduction to the MATLAB approach to the solution of transient response problems. Section 5-6 gives an example of a transient-response problem solved with MATLAB. Section 5-7 presents Routh's stability criterion. Section 5-8 discusses effects of integral and derivative control actions on system performance. Finally, Section 5-9 treats steady-state errors in unity-feedback control systems.

5-2 FIRST-ORDER SYSTEMS

Consider the first-order system shown in Figure 5-1(a). Physically, this system may represent an RC circuit, thermal system, or the like. A simplified block diagram is shown in Figure 5-1(b). The input-output relationship is given by

$$\frac{C(s)}{R(s)} = \frac{1}{Ts + 1} \quad (5-1)$$

In the following, we shall analyze the system responses to such inputs as the unit-step, unit-ramp, and unit-impulse functions. The initial conditions are assumed to be zero.

Note that all systems having the same transfer function will exhibit the same output in response to the same input. For any given physical system, the mathematical response can be given a physical interpretation.

Unit-Step Response of First-Order Systems. Since the Laplace transform of the unit-step function is $1/s$, substituting $R(s) = 1/s$ into Equation (5-1), we obtain

$$C(s) = \frac{1}{Ts + 1} \frac{1}{s}$$

Expanding $C(s)$ into partial fractions gives

$$C(s) = \frac{1}{s} - \frac{T}{Ts + 1} = \frac{1}{s} - \frac{1}{s + (1/T)} \quad (5-2)$$

Taking the inverse Laplace transform of Equation (5-2), we obtain

$$c(t) = 1 - e^{-t/T}, \quad \text{for } t \geq 0 \quad (5-3)$$

Equation (5-3) states that initially the output $c(t)$ is zero and finally it becomes unity. One important characteristic of such an exponential response curve $c(t)$ is that at $t = T$ the value of $c(t)$ is 0.632, or the response $c(t)$ has reached 63.2% of its total change. This may be easily seen by substituting $t = T$ in $c(t)$. That is,

$$c(T) = 1 - e^{-1} = 0.632$$

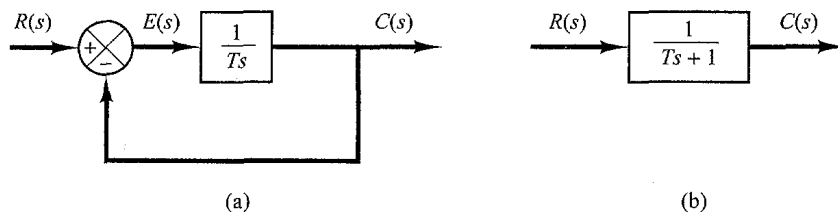
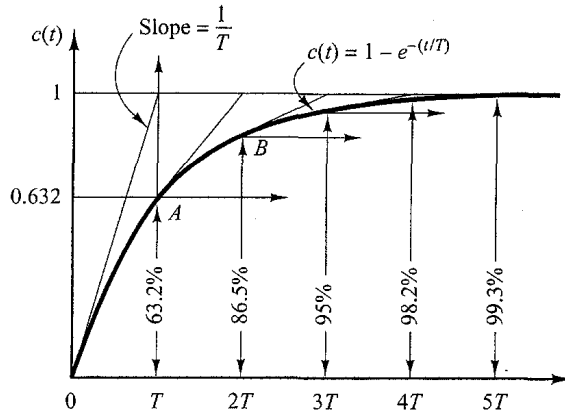


Figure 5-1
(a) Block diagram of a first-order system;
(b) simplified block diagram.

Figure 5-2
Exponential
response curve.



Note that the smaller the time constant T , the faster the system response. Another important characteristic of the exponential response curve is that the slope of the tangent line at $t = 0$ is $1/T$, since

$$\left. \frac{dc}{dt} \right|_{t=0} = \left. \frac{1}{T} e^{-t/T} \right|_{t=0} = \frac{1}{T} \quad (5-4)$$

The output would reach the final value at $t = T$ if it maintained its initial speed of response. From Equation (5-4) we see that the slope of the response curve $c(t)$ decreases monotonically from $1/T$ at $t = 0$ to zero at $t = \infty$.

The exponential response curve $c(t)$ given by Equation (5-3) is shown in Figure 5-2. In one time constant, the exponential response curve has gone from 0 to 63.2% of the final value. In two time constants, the response reaches 86.5% of the final value. At $t = 3T, 4T$, and $5T$, the response reaches 95%, 98.2%, and 99.3%, respectively, of the final value. Thus, for $t \geq 4T$, the response remains within 2% of the final value. As seen from Equation (5-3), the steady state is reached mathematically only after an infinite time. In practice, however, a reasonable estimate of the response time is the length of time the response curve needs to reach and stay within the 2% line of the final value, or four time constants.

Unit-Ramp Response of First-Order Systems. Since the Laplace transform of the unit-ramp function is $1/s^2$, we obtain the output of the system of Figure 5-1(a) as

$$C(s) = \frac{1}{Ts + 1} \frac{1}{s^2}$$

Expanding $C(s)$ into partial fractions gives

$$C(s) = \frac{1}{s^2} - \frac{T}{s} + \frac{T^2}{Ts + 1} \quad (5-5)$$

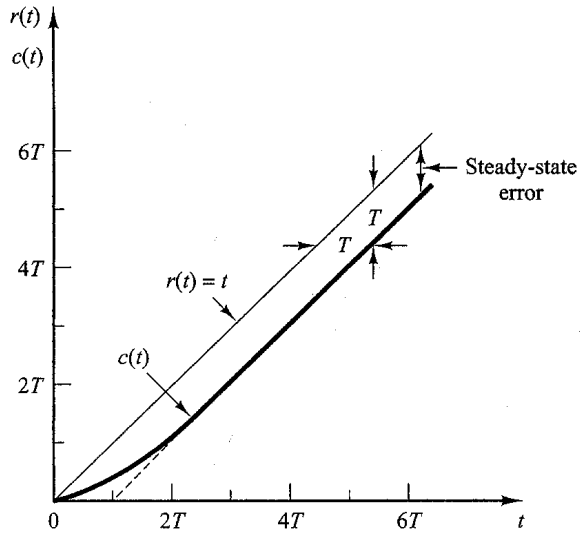
Taking the inverse Laplace transform of Equation (5-5), we obtain

$$c(t) = t - T + Te^{-t/T}, \quad \text{for } t \geq 0 \quad (5-6)$$

The error signal $e(t)$ is then

$$\begin{aligned} e(t) &= r(t) - c(t) \\ &= T(1 - e^{-t/T}) \end{aligned}$$

Figure 5-3
Unit-ramp response
of the system shown
in Figure 5-1(a).



As t approaches infinity, $e^{-t/T}$ approaches zero, and thus the error signal $e(t)$ approaches T or

$$e(\infty) = T$$

The unit-ramp input and the system output are shown in Figure 5-3. The error in following the unit-ramp input is equal to T for sufficiently large t . The smaller the time constant T , the smaller the steady-state error in following the ramp input.

Unit-Impulse Response of First-Order Systems. For the unit-impulse input, $R(s) = 1$ and the output of the system of Figure 5-1(a) can be obtained as

$$C(s) = \frac{1}{Ts + 1} \quad (5-7)$$

The inverse Laplace transform of Equation (5-7) gives

$$c(t) = \frac{1}{T} e^{-t/T}, \quad \text{for } t \geq 0 \quad (5-8)$$

The response curve given by Equation (5-8) is shown in Figure 5-4.

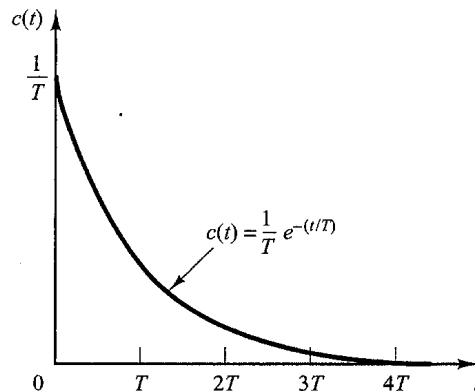


Figure 5-4
Unit-impulse
response of the
system shown in
Figure 5-1(a).

An Important Property of Linear Time-Invariant Systems. In the analysis above, it has been shown that for the unit-ramp input the output $c(t)$ is

$$c(t) = t - T + Te^{-t/T}, \quad \text{for } t \geq 0 \quad [\text{See Equation (5-6).}]$$

For the unit-step input, which is the derivative of unit-ramp input, the output $c(t)$ is

$$c(t) = 1 - e^{-t/T}, \quad \text{for } t \geq 0 \quad [\text{See Equation (5-3).}]$$

Finally, for the unit-impulse input, which is the derivative of unit-step input, the output $c(t)$ is

$$c(t) = \frac{1}{T} e^{-t/T}, \quad \text{for } t \geq 0 \quad [\text{See Equation (5-8).}]$$

Comparing the system responses to these three inputs clearly indicates that the response to the derivative of an input signal can be obtained by differentiating the response of the system to the original signal. It can also be seen that the response to the integral of the original signal can be obtained by integrating the response of the system to the original signal and by determining the integration constant from the zero output initial condition. This is a property of linear time-invariant systems. Linear time-varying systems and nonlinear systems do not possess this property.

5-3 SECOND-ORDER SYSTEMS

In this section, we shall obtain the response of a typical second-order control system to a step input, ramp input, and impulse input. Here we consider a servo system as an example of a second-order system.

Servo System. The servo system shown in Figure 5-5(a) consists of a proportional controller and load elements (inertia and viscous friction elements). Suppose that we wish to control the output position c in accordance with the input position r .

The equation for the load elements is

$$J\ddot{c} + B\dot{c} = T$$

where T is the torque produced by the proportional controller whose gain is K . By taking Laplace transforms of both sides of this last equation, assuming the zero initial conditions, we obtain

$$Js^2C(s) + BsC(s) = T(s)$$

So the transfer function between $C(s)$ and $T(s)$ is

$$\frac{C(s)}{T(s)} = \frac{1}{s(Js + B)}$$

By using this transfer function, Figure 5-5(a) can be redrawn as in Figure 5-5(b), which can be modified to that shown in Figure 5-5(c). The closed-loop transfer function is then obtained as

$$\frac{C(s)}{R(s)} = \frac{K}{Js^2 + Bs + K} = \frac{K/J}{s^2 + (B/J)s + (K/J)}$$

Such a system where the closed-loop transfer function possesses two poles is called a second-order system. (Some second-order systems may involve one or two zeros.)

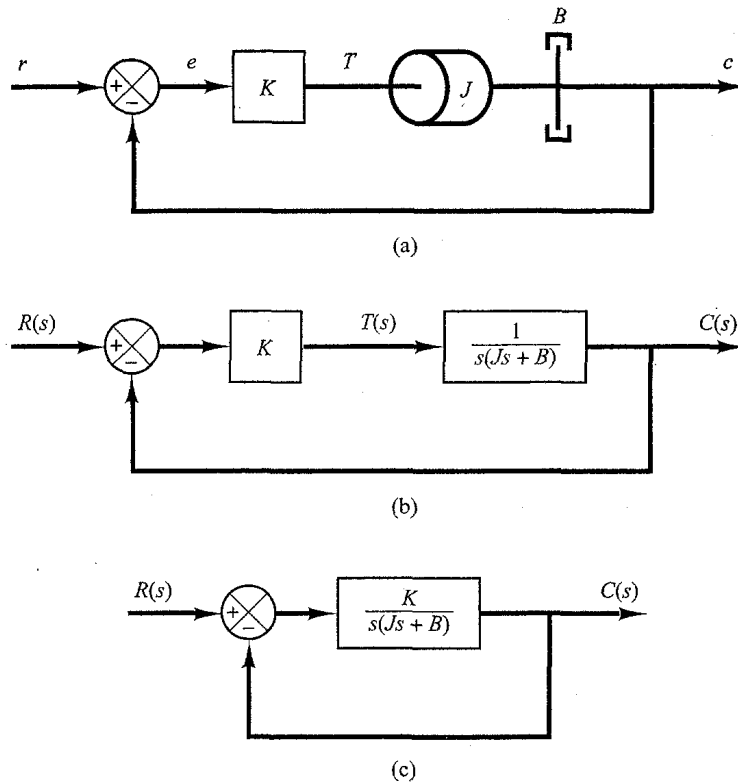


Figure 5-5
 (a) Servo system;
 (b) block diagram;
 (c) simplified block diagram.

Step Response of Second-Order System. The closed-loop transfer function of the system shown in Figure 5-5(c) is

$$\frac{C(s)}{R(s)} = \frac{K}{Js^2 + Bs + K} \quad (5-9)$$

which can be rewritten as

$$\frac{C(s)}{R(s)} = \frac{\frac{K}{J}}{\left[s + \frac{B}{2J} + \sqrt{\left(\frac{B}{2J}\right)^2 - \frac{K}{J}} \right] \left[s + \frac{B}{2J} - \sqrt{\left(\frac{B}{2J}\right)^2 - \frac{K}{J}} \right]}$$

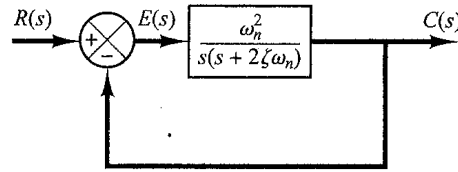
The closed-loop poles are complex conjugates if $B^2 - 4JK < 0$ and they are real if $B^2 - 4JK \geq 0$. In the transient-response analysis, it is convenient to write

$$\frac{K}{J} = \omega_n^2, \quad \frac{B}{J} = 2\zeta\omega_n = 2\sigma$$

where σ is called the *attenuation*; ω_n , the *undamped natural frequency*; and ζ , the *damping ratio* of the system. The damping ratio ζ is the ratio of the actual damping B to the critical damping $B_c = 2\sqrt{JK}$ or

$$\zeta = \frac{B}{B_c} = \frac{B}{2\sqrt{JK}}$$

Figure 5–6
Second-order system.



In terms of ζ and ω_n , the system shown in Figure 5–5(c) can be modified to that shown in Figure 5–6, and the closed-loop transfer function $C(s)/R(s)$ given by Equation (5–9) can be written

$$\frac{C(s)}{R(s)} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (5-10)$$

This form is called the *standard form* of the second-order system.

The dynamic behavior of the second-order system can then be described in terms of two parameters ζ and ω_n . If $0 < \zeta < 1$, the closed-loop poles are complex conjugates and lie in the left-half s plane. The system is then called underdamped, and the transient response is oscillatory. If $\zeta = 0$, the transient response does not die out. If $\zeta = 1$, the system is called critically damped. Overdamped systems correspond to $\zeta > 1$.

We shall now solve for the response of the system shown in Figure 5–6 to a unit-step input. We shall consider three different cases: the underdamped ($0 < \zeta < 1$), critically damped ($\zeta = 1$), and overdamped ($\zeta > 1$) cases.

(1) *Underdamped case* ($0 < \zeta < 1$): In this case, $C(s)/R(s)$ can be written

$$\frac{C(s)}{R(s)} = \frac{\omega_n^2}{(s + \zeta\omega_n + j\omega_d)(s + \zeta\omega_n - j\omega_d)}$$

where $\omega_d = \omega_n\sqrt{1 - \zeta^2}$. The frequency ω_d is called the *damped natural frequency*. For a unit-step input, $C(s)$ can be written

$$C(s) = \frac{\omega_n^2}{(s^2 + 2\zeta\omega_n s + \omega_n^2)s} \quad (5-11)$$

The inverse Laplace transform of Equation (5–11) can be obtained easily if $C(s)$ is written in the following form:

$$\begin{aligned} C(s) &= \frac{1}{s} - \frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \\ &= \frac{1}{s} - \frac{s + \zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d^2} - \frac{\zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d^2} \end{aligned}$$

In Chapter 2 it was shown that

$$\begin{aligned} \mathcal{L}^{-1}\left[\frac{s + \zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d^2}\right] &= e^{-\zeta\omega_n t} \cos \omega_d t \\ \mathcal{L}^{-1}\left[\frac{\omega_d}{(s + \zeta\omega_n)^2 + \omega_d^2}\right] &= e^{-\zeta\omega_n t} \sin \omega_d t \end{aligned}$$

Hence the inverse Laplace transform of Equation (5-11) is obtained as

$$\begin{aligned}\mathcal{L}^{-1}[C(s)] &= c(t) \\ &= 1 - e^{-\zeta\omega_n t} \left(\cos \omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \omega_d t \right) \\ &= 1 - \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}} \sin \left(\omega_d t + \tan^{-1} \frac{\sqrt{1-\zeta^2}}{\zeta} \right), \quad \text{for } t \geq 0 \quad (5-12)\end{aligned}$$

This result can be obtained directly by using a table of Laplace transforms. From Equation (5-12), it can be seen that the frequency of transient oscillation is the damped natural frequency ω_d and thus varies with the damping ratio ζ . The error signal for this system is the difference between the input and output and is

$$\begin{aligned}e(t) &= r(t) - c(t) \\ &= e^{-\zeta\omega_n t} \left(\cos \omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \omega_d t \right), \quad \text{for } t \geq 0\end{aligned}$$

This error signal exhibits a damped sinusoidal oscillation. At steady state, or at $t = \infty$, no error exists between the input and output.

If the damping ratio ζ is equal to zero, the response becomes undamped and oscillations continue indefinitely. The response $c(t)$ for the zero damping case may be obtained by substituting $\zeta = 0$ in Equation (5-12), yielding

$$c(t) = 1 - \cos \omega_n t, \quad \text{for } t \geq 0 \quad (5-13)$$

Thus, from Equation (5-13), we see that ω_n represents the undamped natural frequency of the system. That is, ω_n is that frequency at which the system output would oscillate if the damping were decreased to zero. If the linear system has any amount of damping, the undamped natural frequency cannot be observed experimentally. The frequency that may be observed is the damped natural frequency ω_d , which is equal to $\omega_n \sqrt{1-\zeta^2}$. This frequency is always lower than the undamped natural frequency. An increase in ζ would reduce the damped natural frequency ω_d . If ζ is increased beyond unity, the response becomes overdamped and will not oscillate.

(2) *Critically damped case* ($\zeta = 1$): If the two poles of $C(s)/R(s)$ are equal, the system is said to be a critically damped one.

For a unit-step input, $R(s) = 1/s$ and $C(s)$ can be written

$$C(s) = \frac{\omega_n^2}{(s + \omega_n)^2 s} \quad (5-14)$$

The inverse Laplace transform of Equation (5-14) may be found as

$$c(t) = 1 - e^{-\omega_n t} (1 + \omega_n t), \quad \text{for } t \geq 0 \quad (5-15)$$

This result can also be obtained by letting ζ approach unity in Equation (5-12) and by using the following limit:

$$\lim_{\zeta \rightarrow 1} \frac{\sin \omega_d t}{\sqrt{1-\zeta^2}} = \lim_{\zeta \rightarrow 1} \frac{\sin \omega_n \sqrt{1-\zeta^2} t}{\sqrt{1-\zeta^2}} = \omega_n t$$

(3) *Overdamped case* ($\zeta > 1$): In this case, the two poles of $C(s)/R(s)$ are negative real and unequal. For a unit-step input, $R(s) = 1/s$ and $C(s)$ can be written

$$C(s) = \frac{\omega_n^2}{(s + \zeta\omega_n + \omega_n\sqrt{\zeta^2 - 1})(s + \zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1})s} \quad (5-16)$$

The inverse Laplace transform of Equation (5-16) is

$$\begin{aligned} c(t) &= 1 + \frac{1}{2\sqrt{\zeta^2 - 1}(\zeta + \sqrt{\zeta^2 - 1})} e^{-(\zeta + \sqrt{\zeta^2 - 1})\omega_n t} \\ &\quad - \frac{1}{2\sqrt{\zeta^2 - 1}(\zeta - \sqrt{\zeta^2 - 1})} e^{-(\zeta - \sqrt{\zeta^2 - 1})\omega_n t} \\ &= 1 + \frac{\omega_n}{2\sqrt{\zeta^2 - 1}} \left(\frac{e^{-s_1 t}}{s_1} - \frac{e^{-s_2 t}}{s_2} \right), \quad \text{for } t \geq 0 \end{aligned} \quad (5-17)$$

where $s_1 = (\zeta + \sqrt{\zeta^2 - 1})\omega_n$ and $s_2 = (\zeta - \sqrt{\zeta^2 - 1})\omega_n$. Thus, the response $c(t)$ includes two decaying exponential terms.

When ζ is appreciably greater than unity, one of the two decaying exponentials decreases much faster than the other, so the faster decaying exponential term (which corresponds to a smaller time constant) may be neglected. That is, if $-s_2$ is located very much closer to the $j\omega$ axis than $-s_1$ (which means $|s_2| \ll |s_1|$), then for an approximate solution we may neglect $-s_1$. This is permissible because the effect of $-s_1$ on the response is much smaller than that of $-s_2$, since the term involving s_1 in Equation (5-17) decays much faster than the term involving s_2 . Once the faster decaying exponential term has disappeared, the response is similar to that of a first-order system, and $C(s)/R(s)$ may be approximated by

$$\frac{C(s)}{R(s)} = \frac{\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}}{s + \zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}} = \frac{s_2}{s + s_2}$$

This approximate form is a direct consequence of the fact that the initial values and final values of both the original $C(s)/R(s)$ and the approximate one agree with each other.

With the approximate transfer function $C(s)/R(s)$, the unit-step response can be obtained as

$$C(s) = \frac{\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}}{(s + \zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1})s}$$

The time response $c(t)$ is then

$$c(t) = 1 - e^{-(\zeta - \sqrt{\zeta^2 - 1})\omega_n t}, \quad \text{for } t \geq 0$$

This gives an approximate unit-step response when one of the poles of $C(s)/R(s)$ can be neglected.

A family of unit-step response curves $c(t)$ with various values of ζ is shown in Figure 5-7, where the abscissa is the dimensionless variable $\omega_n t$. The curves are functions

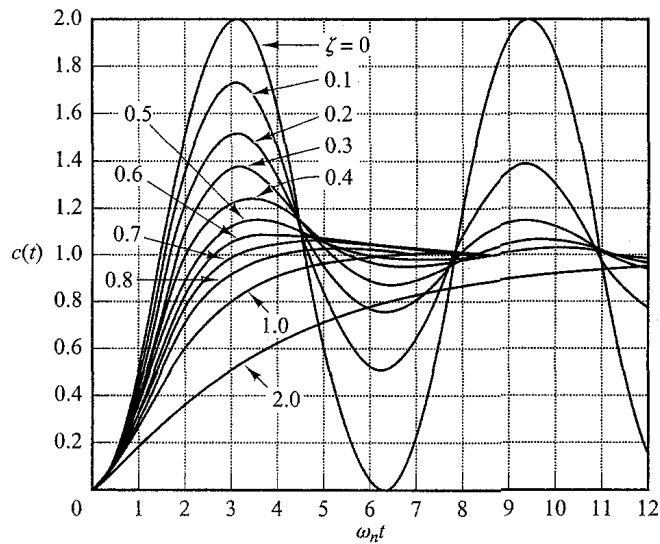


Figure 5-7
Unit-step response curves of the system shown in Figure 5-6.

only of ζ . These curves are obtained from Equations (5-12), (5-15), and (5-17). The system described by these equations was initially at rest.

Note that two second-order systems having the same ζ but different ω_n will exhibit the same overshoot and the same oscillatory pattern. Such systems are said to have the same relative stability.

It is important to note that, for second-order systems whose closed-loop transfer functions are different from that given by Equation (5-10), the step-response curves may look quite different from those shown in Figure 5-7.

From Figure 5-7, we see that an underdamped system with ζ between 0.5 and 0.8 gets close to the final value more rapidly than a critically damped or overdamped system. Among the systems responding without oscillation, a critically damped system exhibits the fastest response. An overdamped system is always sluggish in responding to any inputs.

Definitions of Transient-Response Specifications. In many practical cases, the desired performance characteristics of control systems are specified in terms of time-domain quantities. Systems with energy storage cannot respond instantaneously and will exhibit transient responses whenever they are subjected to inputs or disturbances.

Frequently, the performance characteristics of a control system are specified in terms of the transient response to a unit-step input since it is easy to generate and is sufficiently drastic. (If the response to a step input is known, it is mathematically possible to compute the response to any input.)

The transient response of a system to a unit-step input depends on the initial conditions. For convenience in comparing transient responses of various systems, it is a common practice to use the standard initial condition that the system is at rest initially with the output and all time derivatives thereof zero. Then the response characteristics of many systems can be easily compared.

The transient response of a practical control system often exhibits damped oscillations before reaching steady state. In specifying the transient-response characteristics of a control system to a unit-step input, it is common to specify the following:

1. Delay time, t_d
2. Rise time, t_r
3. Peak time, t_p
4. Maximum overshoot, M_p
5. Settling time, t_s

These specifications are defined in what follows and are shown graphically in Figure 5–8.

1. Delay time, t_d : The delay time is the time required for the response to reach half the final value the very first time.
2. Rise time, t_r : The rise time is the time required for the response to rise from 10% to 90%, 5% to 95%, or 0% to 100% of its final value. For underdamped second-order systems, the 0% to 100% rise time is normally used. For overdamped systems, the 10% to 90% rise time is commonly used.
3. Peak time, t_p : The peak time is the time required for the response to reach the first peak of the overshoot.
4. Maximum (percent) overshoot, M_p : The maximum overshoot is the maximum peak value of the response curve measured from unity. If the final steady-state value of the response differs from unity, then it is common to use the maximum percent overshoot. It is defined by

$$\text{Maximum percent overshoot} = \frac{c(t_p) - c(\infty)}{c(\infty)} \times 100\%$$

The amount of the maximum (percent) overshoot directly indicates the relative stability of the system.

5. Settling time, t_s : The settling time is the time required for the response curve to reach and stay within a range about the final value of size specified by absolute percentage of the final value (usually 2% or 5%). The settling time is related to the largest time constant of the control system. Which percentage error criterion to use may be determined from the objectives of the system design in question.

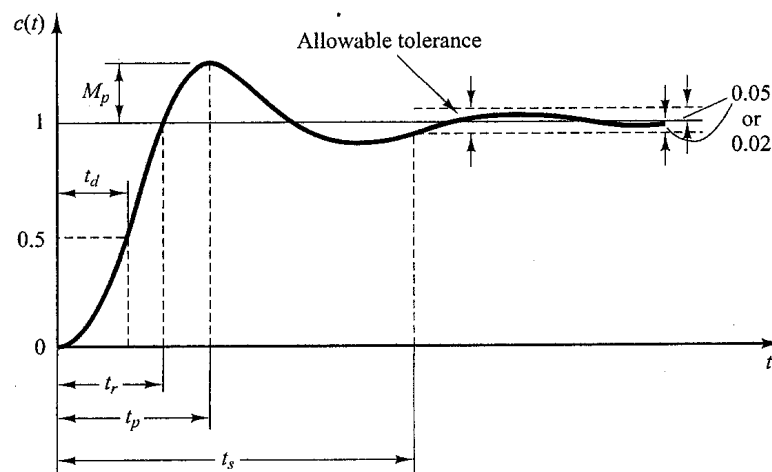


Figure 5–8
Unit-step response curve showing t_d , t_r , t_p , M_p , and t_s .

The time-domain specifications just given are quite important since most control systems are time-domain systems; that is, they must exhibit acceptable time responses. (This means that, the control system must be modified until the transient response is satisfactory.)

Note that not all these specifications necessarily apply to any given case. For example, for an overdamped system, the terms peak time and maximum overshoot do not apply. (For systems that yield steady-state errors for step inputs, this error must be kept within a specified percentage level. Detailed discussions of steady-state errors are postponed until Section 5-9.)

A Few Comments on Transient-Response Specifications. Except for certain applications where oscillations cannot be tolerated, it is desirable that the transient response be sufficiently fast and be sufficiently damped. Thus, for a desirable transient response of a second-order system, the damping ratio must be between 0.4 and 0.8. Small values of ζ ($\zeta < 0.4$) yield excessive overshoot in the transient response, and a system with a large value of ζ ($\zeta > 0.8$) responds sluggishly.

We shall see later that the maximum overshoot and the rise time conflict with each other. In other words, both the maximum overshoot and the rise time cannot be made smaller simultaneously. If one of them is made smaller, the other necessarily becomes larger.

Second-Order Systems and Transient-Response Specifications. In the following, we shall obtain the rise time, peak time, maximum overshoot, and settling time of the second-order system given by Equation (5-10). These values will be obtained in terms of ζ and ω_n . The system is assumed to be underdamped.

Rise time t_r : Referring to Equation (5-12), we obtain the rise time t_r by letting $c(t_r) = 1$.

$$c(t_r) = 1 = 1 - e^{-\zeta\omega_n t_r} \left(\cos \omega_d t_r + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \omega_d t_r \right) \quad (5-18)$$

Since $e^{-\zeta\omega_n t_r} \neq 0$, we obtain from Equation (5-18) the following equation:

$$\cos \omega_d t_r + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \omega_d t_r = 0$$

or

$$\tan \omega_d t_r = -\frac{\sqrt{1-\zeta^2}}{\zeta} = -\frac{\omega_d}{\sigma}$$

Thus, the rise time t_r is

$$t_r = \frac{1}{\omega_d} \tan^{-1} \left(\frac{\omega_d}{-\sigma} \right) = \frac{\pi - \beta}{\omega_d} \quad (5-19)$$

where β is defined in Figure 5-9. Clearly, for a small value of t_r , ω_d must be large.

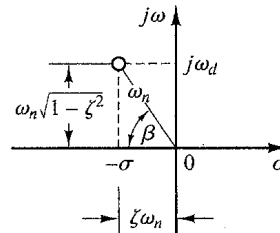


Figure 5-9
Definition of the angle β .

Peak time t_p : Referring to Equation (5-12), we may obtain the peak time by differentiating $c(t)$ with respect to time and letting this derivative equal zero. Since

$$\begin{aligned} \frac{dc}{dt} &= \zeta \omega_n e^{-\zeta \omega_n t} \left(\cos \omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \omega_d t \right) \\ &\quad + e^{-\zeta \omega_n t} \left(\omega_d \sin \omega_d t - \frac{\zeta \omega_d}{\sqrt{1-\zeta^2}} \cos \omega_d t \right) \end{aligned}$$

and the cosine terms in this last equation cancel each other, dc/dt , evaluated at $t = t_p$, can be simplified to

$$\left. \frac{dc}{dt} \right|_{t=t_p} = (\sin \omega_d t_p) \frac{\omega_n}{\sqrt{1-\zeta^2}} e^{-\zeta \omega_n t_p} = 0$$

This last equation yields the following equation:

$$\sin \omega_d t_p = 0$$

or

$$\omega_d t_p = 0, \pi, 2\pi, 3\pi, \dots$$

Since the peak time corresponds to the first peak overshoot, $\omega_d t_p = \pi$. Hence

$$t_p = \frac{\pi}{\omega_d} \quad (5-20)$$

The peak time t_p corresponds to one-half cycle of the frequency of damped oscillation.

Maximum overshoot M_p : The maximum overshoot occurs at the peak time or at $t = t_p = \pi/\omega_d$. Assuming that the final value of the output is unity, M_p is obtained from Equation (5-12) as

$$\begin{aligned} M_p &= c(t_p) - 1 \\ &= -e^{-\zeta \omega_n (\pi/\omega_d)} \left(\cos \pi + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin \pi \right) \\ &= e^{-(\sigma/\omega_d)\pi} = e^{-(\zeta/\sqrt{1-\zeta^2})\pi} \end{aligned} \quad (5-21)$$

The maximum percent overshoot is $e^{-(\sigma/\omega_d)\pi} \times 100\%$.

If the final value $c(\infty)$ of the output is not unity, then we need to use the following equation:

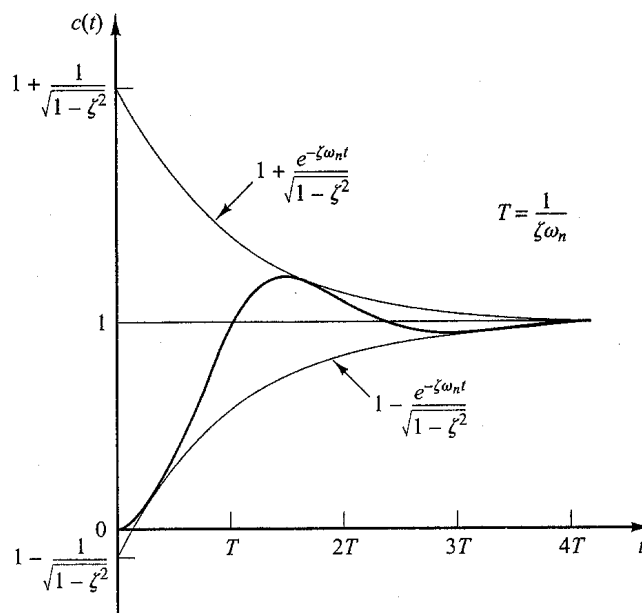
$$M_p = \frac{c(t_p) - c(\infty)}{c(\infty)}$$

Settling time t_s : For an underdamped second-order system, the transient response is obtained from Equation (5-12) as

$$c(t) = 1 - \frac{e^{-\zeta \omega_n t}}{\sqrt{1-\zeta^2}} \sin \left(\omega_d t + \tan^{-1} \frac{\sqrt{1-\zeta^2}}{\zeta} \right), \quad \text{for } t \geq 0$$

The curves $1 \pm (e^{-\zeta \omega_n t}/\sqrt{1-\zeta^2})$ are the envelope curves of the transient response to a unit-step input. The response curve $c(t)$ always remains within a pair of the envelope curves, as shown in Figure 5-10. The time constant of these envelope curves is $1/\zeta \omega_n$.

Figure 5-10
 Pair of envelope curves for the unit-step response curve of the system shown in Figure 5-6.



The speed of decay of the transient response depends on the value of the time constant $1/\zeta\omega_n$. For a given ω_n , the settling time t_s is a function of the damping ratio ζ . From Figure 5-7, we see that for the same ω_n and for a range of ζ between 0 and 1 the settling time t_s for a very lightly damped system is larger than that for a properly damped system. For an overdamped system, the settling time t_s becomes large because of the sluggish response.

The settling time corresponding to a $\pm 2\%$ or $\pm 5\%$ tolerance band may be measured in terms of the time constant $T = 1/\zeta\omega_n$ from the curves of Figure 5-7 for different values of ζ . The results are shown in Figure 5-11. For $0 < \zeta < 0.9$, if the 2% criterion is used, t_s is approximately four times the time constant of the system. If the 5% criterion is used, then t_s is approximately three times the time constant. Note that the settling time reaches a minimum value around $\zeta = 0.76$ (for the 2% criterion) or $\zeta = 0.68$ (for the 5% criterion) and then increases almost linearly for large values of ζ . The discontinuities in the curves of Figure 5-11 arise because an infinitesimal change in the value of ζ can cause a finite change in the settling time.

For convenience in comparing the responses of systems, we commonly define the settling time t_s to be

$$t_s = 4T = \frac{4}{\sigma} = \frac{4}{\zeta\omega_n} \quad (2\% \text{ criterion}) \quad (5-22)$$

or

$$t_s = 3T = \frac{3}{\sigma} = \frac{3}{\zeta\omega_n} \quad (5\% \text{ criterion}) \quad (5-23)$$

Note that the settling time is inversely proportional to the product of the damping ratio and the undamped natural frequency of the system. Since the value of ζ is usually determined from the requirement of permissible maximum overshoot, the settling time is determined primarily by the undamped natural frequency ω_n . This means that the

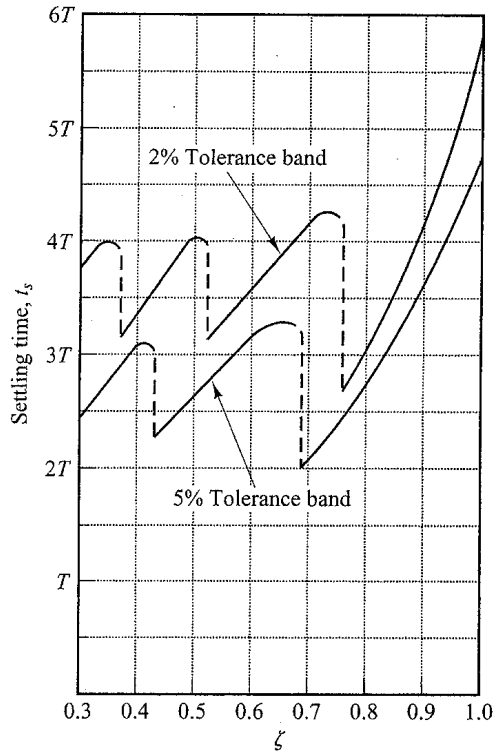


Figure 5-11
Settling time t_s
versus ζ curves.

duration of the transient period may be varied, without changing the maximum overshoot, by adjusting the undamped natural frequency ω_n .

From the preceding analysis, it is evident that for rapid response ω_n must be large. To limit the maximum overshoot M_p and to make the settling time small, the damping ratio ζ should not be too small. The relationship between the maximum percent overshoot M_p and the damping ratio ζ is presented in Figure 5-12. Note that if the damping ratio is between 0.4 and 0.7 then the maximum percent overshoot for step response is between 25% and 4%.

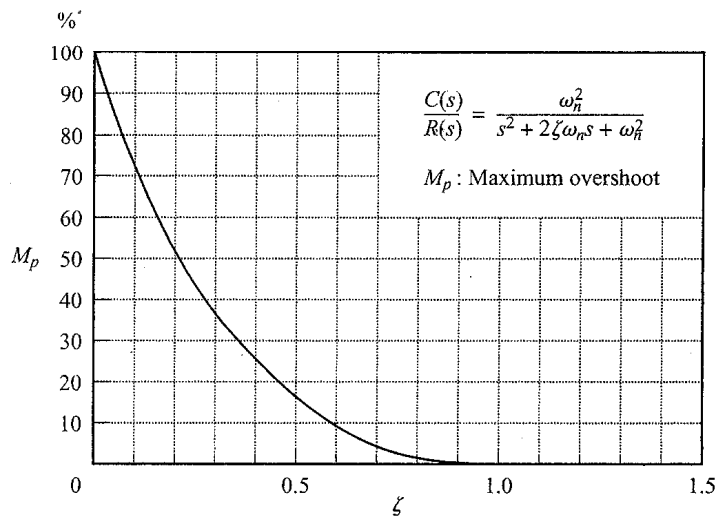


Figure 5-12
 M_p versus ζ curve.

It is important to note that the equations for obtaining the rise time, peak time, maximum overshoot, and settling time are valid only for the standard second-order system defined by Equation (5-10). If the second-order system involves a zero or two zeros, the shape of the unit-step response curve will be quite different from those shown in Figure 5-7.

EXAMPLE 5-1 Consider the system shown in Figure 5-6, where $\zeta = 0.6$ and $\omega_n = 5$ rad/sec. Let us obtain the rise time t_r , peak time t_p , maximum overshoot M_p , and settling time t_s when the system is subjected to a unit-step input.

From the given values of ζ and ω_n , we obtain $\omega_d = \omega_n \sqrt{1 - \zeta^2} = 4$ and $\sigma = \zeta \omega_n = 3$.

Rise time t_r : The rise time is

$$t_r = \frac{\pi - \beta}{\omega_d} = \frac{3.14 - \beta}{4}$$

where β is given by

$$\beta = \tan^{-1} \frac{\omega_d}{\sigma} = \tan^{-1} \frac{4}{3} = 0.93 \text{ rad}$$

The rise time t_r is thus

$$t_r = \frac{3.14 - 0.93}{4} = 0.55 \text{ sec}$$

Peak time t_p : The peak time is

$$t_p = \frac{\pi}{\omega_d} = \frac{3.14}{4} = 0.785 \text{ sec}$$

Maximum overshoot M_p : The maximum overshoot is

$$M_p = e^{-(\sigma/\omega_n)\pi} = e^{-(3/4) \times 3.14} = 0.095$$

The maximum percent overshoot is thus 9.5%.

Settling time t_s : For the 2% criterion, the settling time is

$$t_s = \frac{4}{\sigma} = \frac{4}{3} = 1.33 \text{ sec}$$

For the 5% criterion,

$$t_s = \frac{3}{\sigma} = \frac{3}{3} = 1 \text{ sec}$$

Servo System with Velocity Feedback. The derivative of the output signal can be used to improve system performance. In obtaining the derivative of the output position signal, it is desirable to use a tachometer instead of physically differentiating the output signal. (Note that the differentiation amplifies noise effects. In fact, if discontinuous noises are present, differentiation amplifies the discontinuous noises more than the useful signal. For example, the output of a potentiometer is a discontinuous voltage signal because, as the potentiometer brush is moving on the windings, voltages are induced in the switchover turns and thus generate transients. The output of the potentiometer therefore should not be followed by a differentiating element.)

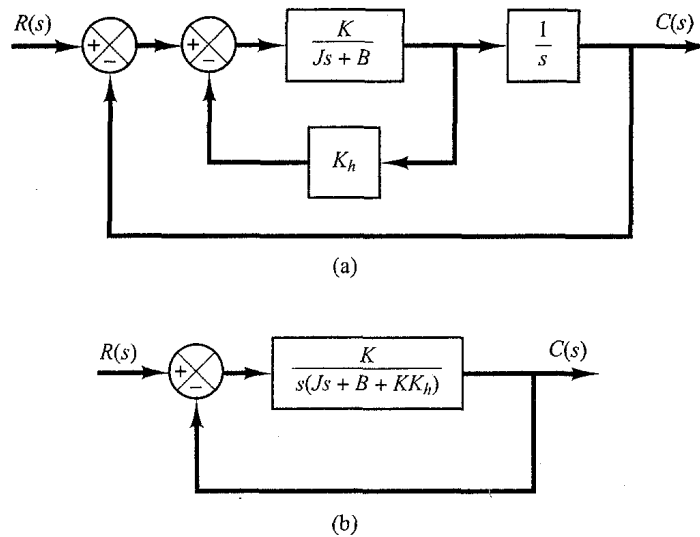


Figure 5-13
 (a) Block diagram of a servo system;
 (b) simplified block diagram.

The tachometer, a special dc generator, is frequently used to measure velocity without differentiation process. The output of a tachometer is proportional to the angular velocity of the motor.

Consider the servo system shown in Figure 5-13(a). In this device, the velocity signal, together with the positional signal, is fed back to the input to produce the actuating error signal. In any servo system, such a velocity signal can be easily generated by a tachometer. The block diagram shown in Figure 5-13(a) can be simplified, as shown in Figure 5-13(b), giving

$$\frac{C(s)}{R(s)} = \frac{K}{Js^2 + (B + KK_h)s + K} \quad (5-24)$$

Comparing Equation (5-24) with Equation (5-9), notice that the velocity feedback has the effect of increasing damping. The damping ratio ζ becomes

$$\zeta = \frac{B + KK_h}{2\sqrt{KJ}} \quad (5-25)$$

The undamped natural frequency $\omega_n = \sqrt{K/J}$ is not affected by velocity feedback. Noting that the maximum overshoot for a unit-step input can be controlled by controlling the value of the damping ratio ζ , we can reduce the maximum overshoot by adjusting the velocity feedback constant K_h so that ζ is between 0.4 and 0.7.

Remember that velocity feedback has the effect of increasing the damping ratio without affecting the undamped natural frequency of the system.

EXAMPLE 5-2

For the system shown in Figure 5-13(a), determine the values of gain K and velocity feedback constant K_h so that the maximum overshoot in the unit-step response is 0.2 and the peak time is 1 sec. With these values of K and K_h , obtain the rise time and settling time. Assume that $J = 1 \text{ kg}\cdot\text{m}^2$ and $B = 1 \text{ N}\cdot\text{m}/\text{rad}/\text{sec}$.

Determination of the values of K and K_h : The maximum overshoot M_p is given by Equation (5-21) as

$$M_p = e^{-(\zeta/\sqrt{1-\zeta^2})\pi}$$

This value must be 0.2. Thus,

$$e^{-(\zeta/\sqrt{1-\zeta^2})\pi} = 0.2$$

or

$$\frac{\zeta\pi}{\sqrt{1-\zeta^2}} = 1.61$$

which yields

$$\zeta = 0.456$$

The peak time t_p is specified as 1 sec; therefore, from Equation (5-20),

$$t_p = \frac{\pi}{\omega_d} = 1$$

or

$$\omega_d = 3.14$$

Since ζ is 0.456, ω_n is

$$\omega_n = \frac{\omega_d}{\sqrt{1-\zeta^2}} = 3.53$$

Since the natural frequency ω_n is equal to $\sqrt{K/J}$,

$$K = J\omega_n^2 = \omega_n^2 = 12.5 \text{ N-m}$$

Then, K_h is, from Equation (5-25),

$$K_h = \frac{2\sqrt{KJ}\zeta - B}{K} = \frac{2\sqrt{K}\zeta - 1}{K} = 0.178 \text{ sec}$$

Rise time t_r : From Equation (5-19), the rise time t_r is

$$t_r = \frac{\pi - \beta}{\omega_d}$$

where

$$\beta = \tan^{-1} \frac{\omega_d}{\sigma} = \tan^{-1} 1.95 = 1.10$$

Thus, t_r is

$$t_r = 0.65 \text{ sec}$$

Settling time t_s : For the 2% criterion,

$$t_s = \frac{4}{\sigma} = 2.48 \text{ sec}$$

For the 5% criterion,

$$t_s = \frac{3}{\sigma} = 1.86 \text{ sec}$$

Impulse Response of Second-Order Systems. For a unit-impulse input $r(t)$, the corresponding Laplace transform is unity, or $R(s) = 1$. The unit-impulse response $C(s)$ of the second-order system shown in Figure 5-6 is

$$C(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

The inverse Laplace transform of this equation yields the time solution for the response $c(t)$ as follows:

For $0 \leq \zeta < 1$,

$$c(t) = \frac{\omega_n}{\sqrt{1 - \zeta^2}} e^{-\zeta\omega_n t} \sin \omega_n \sqrt{1 - \zeta^2} t, \quad \text{for } t \geq 0 \quad (5-26)$$

For $\zeta = 1$,

$$c(t) = \omega_n^2 t e^{-\omega_n t}, \quad \text{for } t \geq 0 \quad (5-27)$$

For $\zeta > 1$,

$$c(t) = \frac{\omega_n}{2\sqrt{\zeta^2 - 1}} e^{-(\zeta - \sqrt{\zeta^2 - 1})\omega_n t} - \frac{\omega_n}{2\sqrt{\zeta^2 - 1}} e^{-(\zeta + \sqrt{\zeta^2 - 1})\omega_n t}, \quad \text{for } t \geq 0 \quad (5-28)$$

Note that without taking the inverse Laplace transform of $C(s)$ we can also obtain the time response $c(t)$ by differentiating the corresponding unit-step response since the unit-impulse function is the time derivative of the unit-step function. A family of unit-impulse response curves given by Equations (5-26) and (5-27) with various values of ζ is shown in Figure 5-14. The curves $c(t)/\omega_n$ are plotted against the dimensionless variable $\omega_n t$, and thus they are functions only of ζ . For the critically damped and overdamped cases, the unit-impulse response is always positive or zero; that is, $c(t) \geq 0$. This can be seen from Equations (5-27) and (5-28). For the underdamped case, the unit-impulse response $c(t)$ oscillates about zero and takes both positive and negative values.

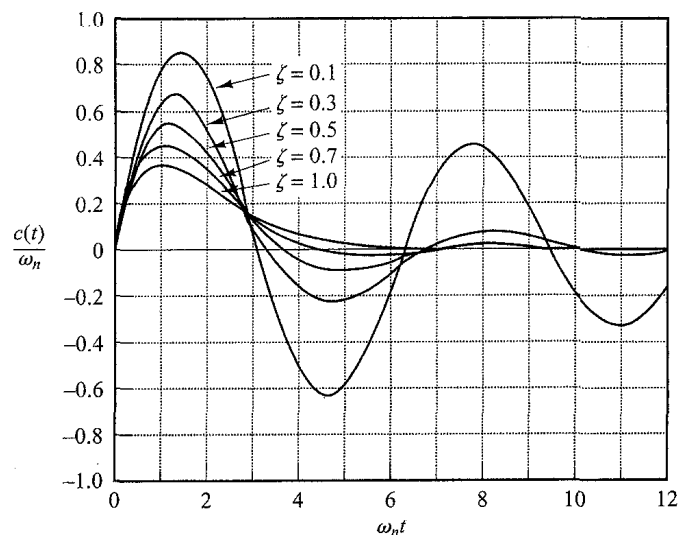
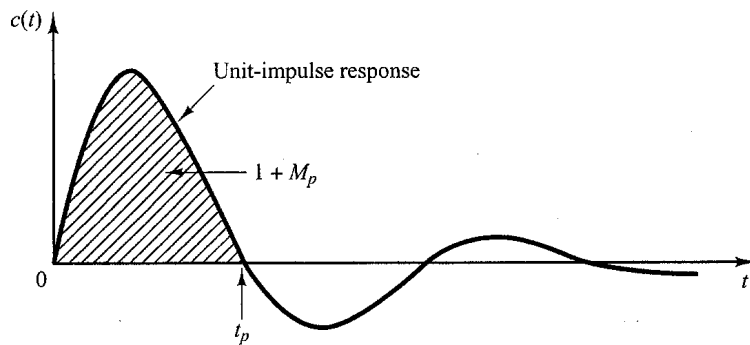


Figure 5-14
Unit-impulse
response curves of
the system shown in
Figure 5-6.

Figure 5-15
Unit-impulse
response curve of the
system shown in
Figure 5-6.



From the foregoing analysis, we may conclude that if the impulse response $c(t)$ does not change sign, the system is either critically damped or overdamped, in which case the corresponding step response does not overshoot but increases or decreases monotonically and approaches a constant value.

The maximum overshoot for the unit-impulse response of the underdamped system occurs at

$$t = \frac{\tan^{-1} \frac{\sqrt{1-\zeta^2}}{\zeta}}{\omega_n \sqrt{1-\zeta^2}}, \quad \text{where } 0 < \zeta < 1 \quad (5-29)$$

[Equation (5-29) can be obtained by equating dc/dt to zero and solving for t .] The maximum overshoot is

$$c(t)_{\max} = \omega_n \exp\left(-\frac{\zeta}{\sqrt{1-\zeta^2}} \tan^{-1} \frac{\sqrt{1-\zeta^2}}{\zeta}\right), \quad \text{where } 0 < \zeta < 1 \quad (5-30)$$

[Equation (5-30) can be obtained by substituting Equation (5-29) into Equation (5-26).]

Since the unit-impulse response function is the time derivative of the unit-step response function, the maximum overshoot M_p for the unit-step response can be found from the corresponding unit-impulse response. That is, the area under the unit-impulse response curve from $t = 0$ to the time of the first zero, as shown in Figure 5-15, is $1 + M_p$, where M_p is the maximum overshoot (for the unit-step response) given by Equation (5-21). The peak time t_p (for the unit-step response) given by Equation (5-20) corresponds to the time that the unit-impulse response first crosses the time axis.

5-4 HIGHER-ORDER SYSTEMS

In this section we shall present a transient response analysis of higher-order systems in general terms. It will be seen that the response of higher-order systems is the sum of the responses of first-order and second-order systems.