COMPANION @ WEBSITE

APPLIED ENGINEERING MATHEMATICS

Brian Vick

CRC Press
Taylor & Francis Group

Applied Engineering Mathematics

Applied Engineering Mathematics

Brian Vick

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Visit the eResources: www.crcpress.com/9780367432768

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-0-3674-3277-5 (Hardback)

Contents

Preface

This book is a practical approach to engineering mathematics with an emphasis on visualization and applications. This book is intended for undergraduate and introductory graduate courses in engineering mathematics and numerical analysis. It is aimed at students in all branches of engineering and science. This book contains a comprehensive blend of fundamental physics, applied science, mathematical analysis, numerical computation, and critical thinking. It contains both theory and application, with the applications interwoven with the theory throughout the text. The emphasis is visual rather than procedural.

This book covers some of the most important mathematical methods and tools used in applied engineering. After an introduction in Chapter 1, this book begins with a summary of the most important principles of physics in Chapter 2, followed by Chapter 3 dedicated to the proper mathematical modeling of physical processes. Then the basics of calculus are presented in Chapter 4, including a thorough treatment of numerical integration. Next the essentials of linear algebra are presented in Chapter 5. Then the topic of nonlinear algebra, with an emphasis on numerical methods, is presented in Chapter 6. The topic of the remaining five chapters is ordinary differential equations. An introduction is presented in Chapter 7, giving an overview and fundamental understanding of the origins and meaning of differential equations. Then the Laplace transform method is presented in Chapter 8. A thorough treatment of the numerical solution of ordinary differential equations is then described in Chapter 9. Chapters 10 and 11 are on first-order and second-order ordinary differential equations, respectively, and cover some important examples and characteristics of first- and second-order equations, including bifurcations. Although the chapters stand alone and can be studied in any order, the organization of this book is a logical sequence from mathematical modeling to solution methodology.

A distinctive characteristic of the text is that the visual approach is emphasized as opposed to excessive proofs and derivations. The reader will take away insight and deeper understanding with the visual images and thus have a better chance of remembering and using the mathematical methods. Many of the figures were created and computations performed with Mathematica, and the dynamic and interactive codes accompanying the examples are available for the reader to explore on their own.

My style has been developed from experience as a long-time teacher and researcher in a variety of engineering and mathematical courses. My background includes the areas of heat transfer, thermodynamics, engineering design, computer programming, numerical analysis, and system dynamics at both undergraduate and graduate levels. Also, my experiences in various research areas have motivated some of the specific topics and examples.

I would like to express thanks to my wife Linda and our children Kristen, Kelsey, Alison, and Everett for all the great times we have had and all your patience with me. I am so blessed and I love you all.

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098 USA
Tel: 508 647 7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

About the Author

Brian Vick received his bachelor's degree in 1976, master's degree in 1978, and doctoral degree in 1981, all in mechanical engineering, all from North Carolina State University. In 1982, he joined the Department of Mechanical Engineering at Virginia Tech as an assistant professor and was promoted to associate professor in 1989. Dr. Vick's main research and teaching activities have been in the areas of heat transfer, applied mathematics, numerical analysis, tribology, wave mechanics, nonlinear dynamics, and parameter estimation.

He has taught undergraduate courses in the areas of heat transfer, thermodynamics, thermal systems design, engineering design, computer programming, numerical analysis, and system dynamics. He has taught graduate level courses in the areas of conduction heat transfer, convection heat transfer, and advanced engineering mathematics.

Dr. Vick has conducted research projects in a variety of areas including the wave nature of heat transport, analysis of phase change phenomena, laminarization in highly accelerated flows, heat transfer augmentation in heat exchangers, modeling and analysis of thermal storage systems, thermal analysis of slip ring designs, development of boundary element methods, electro-thermo-mechanical analysis of shape memory alloy actuators, heat transfer in heterogeneous materials, and thermo-mechanical and thermionic emission studies in tribological processes.

He is currently working on research with NASA on the newest generation of earth radiation budget instruments. He also is actively working in the parameter estimation and inverse problem area with application to radiation instruments, estimation of blood perfusion in living tissue, and the dynamics of biofilm formation.

Chapter 1

Overview

CHAPTER OBJECTIVES

This chapter is an introduction to and overview of the educational philosophy employed in this book. The physical applications and mathematical methods are briefly summarized.

Specific objectives and topics covered are

- Objectives
- Educational philosophy
- Physical processes: conservation laws, rate equations, and property relations
- Mathematical models: algebraic equations, ordinary differential equations, and partial differential equations
- Solution methods for algebraic and differential equations
- Software

1.1 OBJECTIVES

This is a comprehensive book consisting of a blend of fundamental physics, applied science, mathematical analysis, numerical computation, and critical thinking. It contains both theory and application, with the applications interwoven with the theory throughout the text. The emphasis is visual rather than procedural.

The specific goals consist of the following.

- **Physical Processes**
  To gain a fundamental understanding of the physical processes, fundamental principles, and mathematical formulations of physical problems.
- **Mathematical Methods**
  To learn mathematical techniques for solving model equations.
- **Software**
  To learn the use of software packages, such as *Mathematica* or *MATLAB®*, to program solutions, perform calculations, and create graphics. Computational studies and graphics enhance insight into the effect of important parameters in addition to building a fundamental understanding of physical mechanisms.

- **Insight and Critical Thinking**
  To develop a sound foundation for problem-solving using critical thinking, interpretation, and reasoning skills. Applications are used extensively to foster *insight* and *intuition*.

1.2 EDUCATIONAL PHILOSOPHY

The following *guiding principles* **are** fundamental **to learning;**

- We learn by active participation, not by passive observation. That is, we learn by doing, not just by watching.
- A picture is worth a thousand words. The human brain is made to process *visual* information. More information can be assimilated in a few seconds by looking at graphics than by studying that same information for months from a printout of numerical values.
- We are all responsible for our own learning. You need to be self-motivated and have the desire to learn.

An important issue is *knowledge* **versus** *information***.** There is an old saying:

Give me a fish and I'll eat for a day. Teach me to fish and I'll eat for a lifetime.

With today's information explosion resulting from the internet, this old saying is more relevant than ever. Knowledge and fundamental reasoning skills are giving way to an unmanageable amount of information consisting of seemingly unconnected facts and figures. Hopefully, a greater emphasis can be placed on knowledge as opposed to just raw information. True progress requires a balance between raw information and basic knowledge.

In this same vein, too much *coverage of material* at the expense of *depth of understanding* can be the enemy of learning and leads to memorization and frustration. Depth of understanding is considered to be far more important than coverage of more topics.

In addition to developing an appreciation for and mastering knowledge of the fundamental physical principles and mathematical techniques, a major goal of this course is to develop basic learning skills and strategies. These include

- Reasoning and interpretive skills—the foundation of problem-solving
- Pattern recognition skills
- Adaptability—learning to recognize abstract concepts from specific applications and conversely, learning to apply abstract concepts to specific applications
- Thinking for yourself
- Motivation, enthusiasm, and passion

Regardless of your potential, there is no substitute for hard work. One must persist and struggle with difficult concepts until they are understood. Learning is a lifelong activity and is the key to success. Let's make it fun and exciting!

1.3 PHYSICAL PROCESSES

Observations of the physical world indicate that all processes are governed by a small number of fundamental principles. These are *conservation principles*, which are supplemented by *rate equations* and *property relations*. Together, they form a complete description of nature. Although there are only a handful of principles, there are countless applications and special cases. These principles are summarized in the following and are described in more detail in Chapters 2 and 3.

- The ***conservation laws*** are:
  - Conservation of mass: continuity
  - Conservation of momentum: Newton's second law
  - Conservation of energy: first law of thermodynamics
  - Conservation of chemical species
  - Conservation of electrical charge

These are general principles and are *independent* of the material.

- The ***rate equations*** supplement the conservation principles. The most important ones are:
  - Heat conduction: Fourier's law
  - Heat convection: Newton's law of cooling
  - Thermal radiation
  - Viscous fluid shear: Newton's viscosity law
  - Binary mass diffusion: Fick's law
  - Electrical conduction: Ohm's law
  - Stress-strain: Hooke's law

These are constitutive relations and are *dependent* on the material.

- The ***property relationships*** are also needed to complete the mathematical model. A few such relationships are:
  - Constant properties
  - Density: $\rho = \rho(T, P)$
  - Viscosity: $\mu = \mu(T, P)$
  - Specific heat: $c = c(T, P)$
  - Thermal conductivity: $k = k(T, P)$

These are material-dependent characteristics. Many times, it can be justified to assume constant properties.

1.4 MATHEMATICAL MODELS

The mathematical description of physical problems generally leads to an equation or set of equations involving either *algebraic expressions* or *derivatives* (i.e., differential equations). Differential equations that are a function of only one independent variable are referred to as *ordinary differential equations*. Those that depend on two or more independent variables are called *partial differential equations*. As displayed in the following figures, these equations can be classified according to the number of equations or dependent variables and whether they are linear or nonlinear.

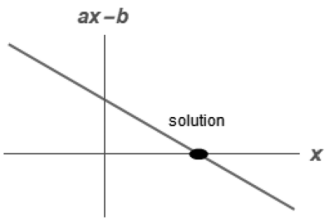
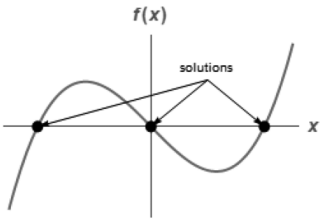
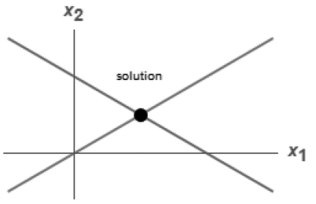
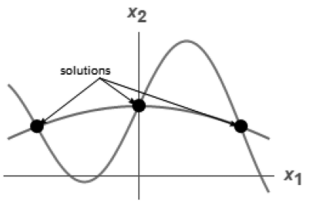
Number of Equations	Linear	Nonlinear
1	$a \cdot x = b$ 	$f(x) = 0$ 
2	$a_{11}x_1 + a_{12}x_2 = b_1$ $a_{21}x_1 + a_{22}x_2 = b_2$ 	$f_1(x_1, x_2) = 0$ $f_2(x_1, x_2) = 0$ 
n	$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$ $a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$ $\vdots$ $a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n$	$f_1(x_1, x_2, \cdots, x_n) = 0$ $f_2(x_1, x_2, \cdots, x_n) = 0$ $\vdots$ $f_n(x_1, x_2, \cdots, x_n) = 0$

Figure 1.1 Classification of algebraic equations.

1.4.1 Algebraic Equations

Algebraic equations can be classified according to the characteristics shown in Figure 1.1. Linear algebraic equations can be solved with procedures such as Gaussian elimination. Many practical problems are modeled with a well-behaved set of linear equations with a unique solution. On the other hand, nonlinear equations can have multiple solutions or no solutions at all and can be tricky to solve.

1.4.2 Ordinary Differential Equations

Differential equations are used to model the dynamical behavior of physical systems. They are rich in application and meaning. A brief summary of ordinary differential equations (ODEs) and partial differential equations (PDEs) is presented in Figures 1.2 and 1.3, respectively.

ODEs can further be classified as *initial value problems* or *boundary value problems* depending on the auxiliary conditions. Initial value problems typically involve time as the independent variable and require starting values for the dependent variables. On the other hand, boundary value problems typically have position as the independent variable and require conditions on all the boundaries of the dependent variables.

1.4.3 Partial Differential Equations

Many types of PDEs exist, exhibiting a wide variety of characteristics. The basic elliptic, parabolic, and hyperbolic equations are displayed in Figure 1.3.

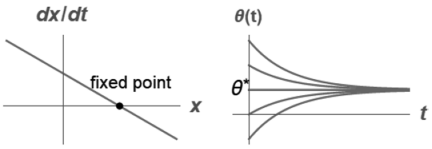
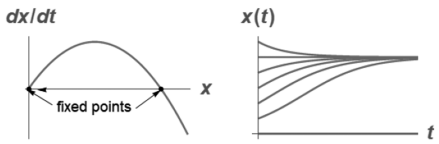
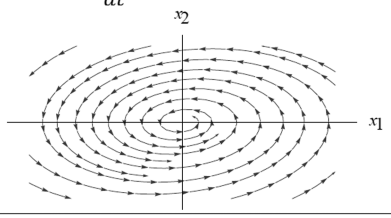
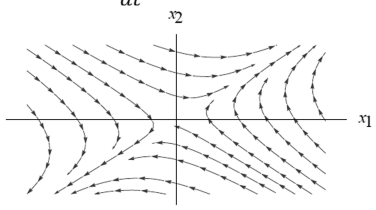
Dependent Variables	Linear	Nonlinear
1	$\frac{dx}{dt} = a \cdot x + b$ 	$\frac{dx}{dt} = f(x)$ 
2	$\frac{dx_1}{dt} = a_{11}x_1 + a_{12}x_2$ $\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2$ 	$\frac{dx_1}{dt} = f_1(x_1, x_2)$ $\frac{dx_2}{dt} = f_2(x_1, x_2)$ 
n	$\frac{dx_1}{dt} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1$ $\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2$ $\vdots$ $\frac{dx_n}{dt} = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_n$	$\frac{dx_1}{dt} = f_1(x_1, x_2, \cdots, x_n)$ $\frac{dx_2}{dt} = f_2(x_1, x_2, \cdots, x_n)$ $\vdots$ $\frac{dx_n}{dt} = f_n(x_1, x_2, \cdots, x_n)$

Figure 1.2 Classification of ordinary differential equations.

Many other variations of these basic PDEs can be formulated. In all cases, starting or initial conditions in time as well as boundary conditions in space are required.

1.5 SOLUTION METHODS

A wide variety of mathematical schemes have been proposed over the centuries to solve equations arising in engineering and applied physics. Some of the more popular and successful ones are summarized in the following.

Some commonly used methods for systems of ***linear algebraic equations*** are:

- Gaussian elimination
- *LU* decomposition
- Gauss–Seidel iteration

For ***nonlinear algebraic equations***, root find methods are employed. They include:

- Bisection
- False position
- Newton–Raphson
- Secant methods
- Golden search
- Gradient methods

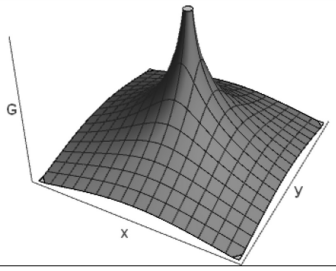
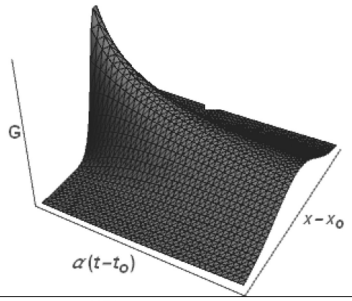
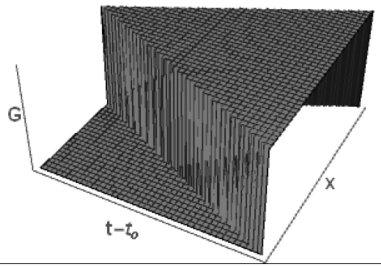
Type of Equation	Typical Equation	Fundamental Behavior
Elliptic or potential equation	$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$	
Parabolic or diffusion equation	$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$	
Hyperbolic or wave equation	$\frac{\partial^2 T}{\partial t^2} = C^2 \frac{\partial^2 T}{\partial x^2}$	

Figure 1.3 Classification of Partial differential equations.

Numerical methods are frequently used to solve ***differential equations***. Some of the most successful methods are:

- Runge–Kutta methods
- Finite difference methods
- Finite element methods
- Boundary element methods
- Cellular automata

Numerous ***analytical methods*** have been developed for ***differential equations***. Some of the most successful methods are:

- Fourier series, orthogonal function expansion, and separation of variables
- Fourier integrals and Fourier transforms
- Green's functions
- Laplace transforms
- Duhamel's method
- Integral methods
- Similarity methods

1.6 SOFTWARE

Some tremendous software packages are currently available. Two of the best and most popular choices are *Mathematica* and *MATLAB®*.

- *Mathematica* is a powerful software package and programming language, which combines numerical computations, symbolic manipulation, graphics, and text. Its symbolic manipulation capabilities are the most powerful ever developed. Mathematica is built on the powerful unifying idea that everything can be represented as a symbolic expression.
- *MATLAB®* is also a programming language that combines numerical computations with graphics. It also has symbolic manipulation capability. The basic data structure in MATLAB® (Matrix Laboratory) is the matrix.
- *Programming* sophisticated software packages can be powerful tools only if one has the necessary skill to program them. This requires logical and structured programming skills. These can only be achieved through tremendous study, practice, and patience.

The learning curve for a general, all-purpose, and powerful package such as Mathematica or MATLAB® can be steep. In the end, the rewards are well worth the effort required. Understanding and advancement of knowledge are greatly facilitated by the ability to program a computer to perform numerical computations, manipulate symbolic expressions, and visualize graphics.

Chapter 2

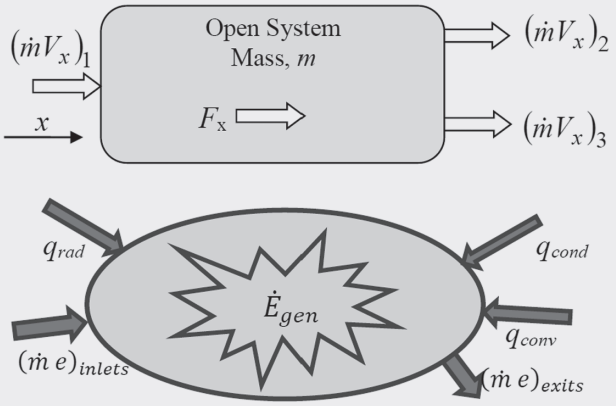
Physical Processes

CHAPTER OBJECTIVES

This chapter describes the basic principles of physics that govern processes in our world. These are postulates based on observation.

Specific objectives and topics covered are

- Physical phenomena
- Fundamental principles
- Conservation laws governing mass, momentum, and energy
- Rate equations relating potentials to flows for heat conduction, convection, radiation, viscous fluid shear, binary mass diffusion, electrical conduction, and stress
- Diffusion analogies



2.1 PHYSICAL PHENOMENA

Physical processes occurring in nature have been categorized in many different ways. Some of the more common categorizes are the following.

- Thermal
- Mechanical
- Chemical
- Electrical
- Biological

In addition, processes involving the coupling or interaction of two or more of the previous basic processes are referred to by names such as

- Thermomechanical
- Electromechanical
- Thermoelectric

Examples of coupled processes are the thermoelectric effect in thermocouples and thermomechanical effects that cause unusual behavior in shape memory alloys such as *Nitinol* (an alloy of approximately 50% nickel and 50% titanium).

Some important applications are tribology (friction, wear, and lubrication), smart materials, lasers, computers, nanostructures, aircraft design, and countless more. One of the fascinating aspects of applied mathematics is the rich and diverse number of applications, all stemming from a few basic principles.

In an attempt to understand our world, humans have classified various observed physical phenomena into these categories. Similarly, we have also departmentalized our universities and companies into categories: Mechanical Engineering Department, Chemical Engineering Department, Biological Systems Department, and so on. However, nature does not recognize these artificial divisions. During a physical process, heat and electricity flow, stresses form, friction and wear occur, and chemical reactions continually change the composition of the system. As a result, the mathematical modeling of real-life systems can be challenging. Engineers and physicists must use intuition and experience, in addition to mathematical procedures, in order to accurately model complex processes in the natural world. The process of mathematically modeling complex processes involves science and art, and perhaps a bit of luck.

2.2 FUNDAMENTAL PRINCIPLES

Observations of the physical world indicate that all processes are governed by a small number of fundamental principles. These are *conservation principles*, which are supplemented by *rate equations* and *property relations*. Together, they form a complete description of nature.

The *conservation laws* are based on the principle that the physical material making up the universe cannot be created or destroyed. The conservation laws are:

- Conservation of mass: Continuity
- Conservation of momentum: Newton's second law
- Conservation of energy: First law of thermodynamics
- Conservation of chemical species
- Conservation of electrical charge

These are general principles, *independent* of the material.

The *rate laws* relate the flow of a conserved quantity, like electric charge or energy, to a driving potential, like voltage or temperature. The rate equations are:

- Heat conduction: Fourier's law
- Heat convection: Newton's law of cooling
- Thermal radiation
- Viscous fluid shear: Newton's viscosity law
- Binary mass diffusion: Fick's law
- Electrical conduction: Ohm's law
- Stress-strain: Hooke's law

These are constitutive relations, *dependent* on the material.

Property relationships are the observed or derived relationships between physical properties. These include:

- Constant properties
- Density: $\rho = \rho(T, P)$
- Viscosity: $\mu = \mu(T, P)$
- Specific heat: $c = c(T, P)$
- Thermal conductivity: $k = k(T, P)$

Here, T and P are considered to be the independent variables. Other combinations of intrinsic material properties could also be used. It should be noted that these laws and relationships are derived from experimental evidence and observation in our natural world. As such they are postulates and thus unprovable. We accept them as truths, or at least as very good approximations. Although there are only a handful of principles, there are countless applications and special cases. These principles are described in the following sections.

2.3 CONSERVATION LAWS

The conservation principles are some of the most important and far-reaching principles of physics. We use them to try to understand the theoretical rationale for processes in nature as well as to guide us in the engineering design of structures, machines, and devices.

The conservation principles are postulates and as such, are not provable. They originate from observations of our world. We accept them on blind faith and use them as a starting point to model physical processes of interest. There are no definitive experimental observations contradicting the conservation principles.

The conservation principles can be applied to any system, ranging from the whole universe down to a differential control volume. Part of the art of mathematical analysis is to choose the system that is most useful to achieve your desired objectives. Generally, we think in terms of the following types of systems:

- Closed system or *fixed mass*—no mass flow in or out
- Open system or *control volume*—mass can cross the system boundaries

In the following sections, the conservation laws for open and closed systems will be described.

2.3.1 Conservation of Mass: Continuity

For *closed systems* no mass can enter or leave, thus conservation of mass requires no change in mass with time, or mass is constant. Thus,

$$\frac{dm}{dt} = 0 \Rightarrow m = \text{constant} \tag{2.1}$$

For *open systems*, mass can enter or leave from various inlets and exits, as shown in Figure 2.1. A mass balance on the open system gives

$$\frac{dm}{dt} = \sum_{\text{inlets}} \dot{m} - \sum_{\text{exits}} \dot{m} \tag{2.2}$$

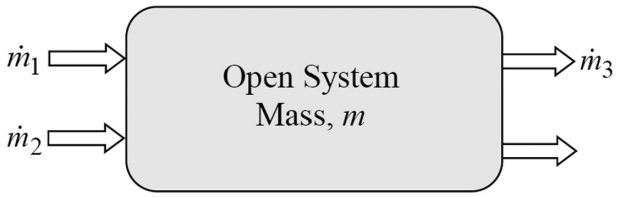


Figure 2.1 Conservation of mass in an open system.

2.3.2 Conservation of Momentum: Newton's Second Law

Newton's second law is one of the most important and widely applied principles in all of physics and engineering. It states that the momentum of an object can only be changed by an external force.

$$\frac{d}{dt}\left(m\bar{V}\right) = \sum \bar{F} \tag{2.3}$$

where

$\bar{V}$ is the velocity vector
$\bar{F}$ is a force vector

This is a vector equation with components in three directions. The x-component is

$$\frac{d}{dt}\left(mV_x\right) = \sum F_x \tag{2.4}$$

For *closed systems*, sometimes referred to as Eulerian systems, mass is constant and Equation 2.3 becomes

$$m\frac{d}{dt}\left(\bar{V}\right) = m\bar{a} = \sum \bar{F} \tag{2.5}$$

The x-component is

$$m\frac{d}{dt}\left(V_x\right) = ma_x = \Sigma F_x \tag{2.6}$$

For *open systems*, sometimes referred to as Lagrangian systems, mass can enter or leave from various inlets and exits, carrying momentum with it, as indicated in Figure 2.2. Newton's second law applied to an open system in the x-direction gives

$$\frac{d}{dt}\left(mV_x\right) = \sum F_x + \sum\left(\dot{m}V_x\right)_{\text{inlets}} - \sum\left(\dot{m}V_x\right)_{\text{exits}} \tag{2.7}$$

Similar relations hold for the y- and z-directions. The application of this principle to a differential fluid control volume produces the *Navier–Stokes* equations of fluid mechanics. The forces and velocities are related through an appropriate rate equation—Newton's law of viscous shear in this case, as discussed in Section 2.4.

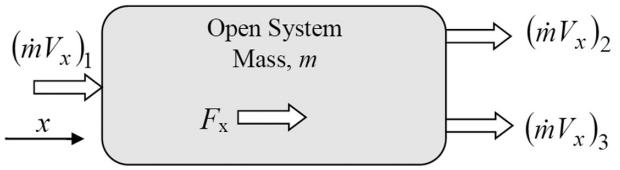


Figure 2.2 Conservation of momentum in an open system.

2.3.3 Conservation of Energy: First Law of Thermodynamics

A closed system is represented in Figure 2.3 showing various energy contributions. Conservation of energy requires that

$$\frac{dE}{dt} = \dot{E}_{\text{in}} - \dot{E}_{\text{out}} + \dot{E}_{\text{gen}} = q_{\text{cond}} + q_{\text{conv}} + q_{\text{rad}} + \dot{E}_{\text{gen}} \tag{2.8}$$

where

E = energy of the system (J)
q_{cond} = heat transfer due to conduction (W)
q_{conv} = heat transfer due to convection (W)
q_{rad} = heat transfer due to radiation (W)
$\dot{E}_{\text{gen}}$ = energy generation (resistance heating, chemical reactions, …)

Note that $\dot{E}_{\text{in}} - \dot{E}_{\text{out}}$ is the net heat transfer due to combined conduction, convection, and radiation.

The conservation of energy principle is now extended to systems where mass can enter or leave from various inlets and exits, carrying energy with it, as indicated in Figure 2.4.

$$\begin{aligned}\frac{dE}{dt} &= \dot{E}_{\text{in}} - \dot{E}_{\text{out}} + \dot{E}_{\text{gen}} + \Sigma(\dot{m}e)_{\text{inlets}} - \Sigma(\dot{m}e)_{\text{exits}} \\ &= q_{\text{cond}} + q_{\text{conv}} + q_{\text{rad}} + \dot{E}_{\text{gen}} + \Sigma(\dot{m}e)_{\text{inlets}} - \Sigma(\dot{m}e)_{\text{exits}}\end{aligned} \tag{2.9}$$

where

e = energy per mass or specific energy (J/kg)
$\dot{m}$ = mass flow rate at inlets or exits (kg/s)

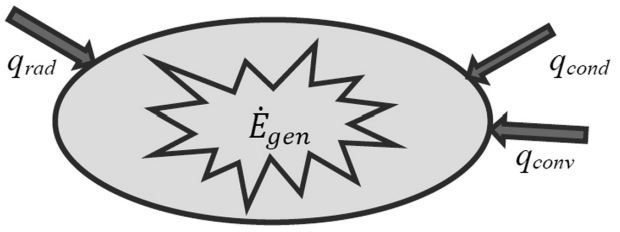


Figure 2.3 Conservation of energy in a closed system.

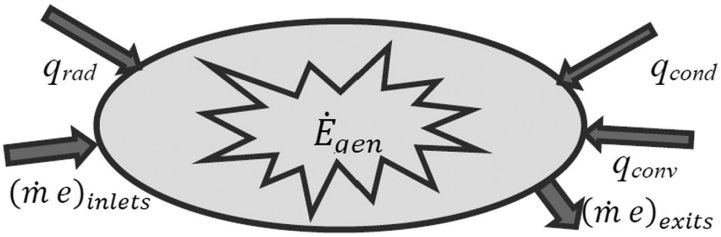


Figure 2.4 Conservation of energy in an open system.

The conduction, convection, and radiation heat flow terms can be related to temperature using rate equations. The rate of energy storage can be related to temperature using the physical property specific heat.

2.4 RATE EQUATIONS

2.4.1 Heat Conduction: Fourier's Law

Heat is conducted from high temperatures to low temperatures. Heat conduction in a plane wall is shown in Figure 2.5, where q''_{cond} is heat flux due to conduction (W/m^2), T is temperature (K or °C), k is thermal conductivity (W/m·K), and x is position (m). For one-dimensional heat conduction, experiments show

$$q''_{\text{cond}} \sim \left(\frac{T_{\text{hot}} - T_{\text{cold}}}{L} \right) \sim \frac{\Delta T}{L}$$

The proportionality may be converted to an equality by introducing the material property thermal conductivity, k. The heat flux is thus modeled as

$$q''_{\text{cond}} = -k \frac{dT}{dx} \tag{2.10}$$

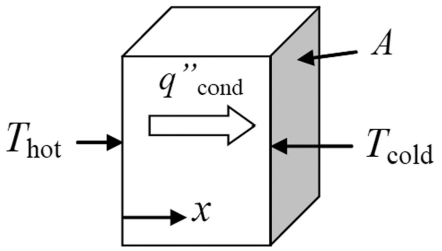


Figure 2.5 Heat flow in a plane wall.

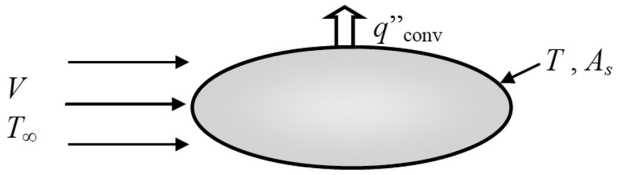


Figure 2.6 Convection heat transfer.

2.4.2 Heat Convection: Newton's Law of Cooling

Convection is the mode of heat transfer due to a combination of two mechanisms: conduction and bulk fluid motion. We are usually interested in convection heat transfer between a solid surface and a flowing fluid, shown in Figure 2.6.

For heat exchange between a solid object at temperature T and a fluid at T_∞, the heat flux is expressed in the form

$$q''_{conv} = h(T - T_\infty) \tag{2.11}$$

where

q''_{conv} = heat flux due to convection (W/m^2)
h = convective heat transfer coefficient (W/m^2·K)

h is a *flow property* that depends on thermal properties of the fluid, flow conditions (V), and geometry. Often, convection calculations are aimed at determining h.

2.4.3 Thermal Radiation

Thermal radiation is electromagnetic radiation emitted by virtue of temperature. Conduction and convection require the presence of a material for the transfer of energy, while radiation does not. A perfect or ideal emitter and absorber of thermal radiation is called a *blackbody*. The heat flux emitted from the surface of a blackbody is

$$E_b = \sigma T_s^4 \tag{2.12}$$

where

E_b = Blackbody emissive power (W/m^2)
T_s = Surface temperature (must use K, not C)
σ = Stefan–Boltzmann constant = 5.67×10^{-8} W/m^2 K^4

A real surface emits some fraction of the radiation of a blackbody at the same temperature; thus,

$$E = \varepsilon E_b = \varepsilon \sigma T_s^4 \tag{2.13}$$

where

E = Emissive power (W/m^2)
ε = Emissivity ($0 < \varepsilon < 1$)

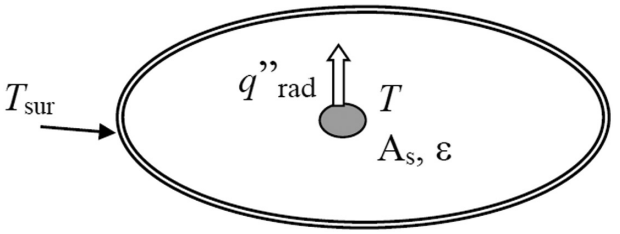


Figure 2.7 Radiation heat transfer.

We are often interested in the special case where the surface receives radiation from a large enclosure at temperature T_{sur}, as shown in Figure 2.7.

For heat exchange between an object and a relatively large enclosure, the net radiative flux is

$$q''_{rad} = \varepsilon\sigma\left(T^4 - T_{sur}^4\right) \tag{2.14}$$

2.4.4 Viscous Fluid Shear: Newton's Viscosity Law

Consider a fluid confined between solid surfaces with a relative velocity V, as shown in Figure 2.8.

The shear force in a moving fluid is postulated to be proportional to the velocity gradient in the form

$$\tau = \mu\frac{dV}{dy} \tag{2.15}$$

where

τ = shear stress (N/m^2)
μ = viscosity (N·m/s)
V = velocity (m/s)

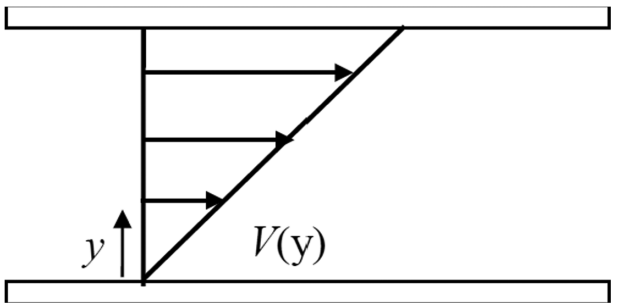


Figure 2.8 Fluid confined between solid surfaces with a relative velocity V.

2.4.5 Binary Mass Diffusion: Fick's Law

Now consider diffusion in a two-component mixture. The mass flux of component 1 through a mixture of components 1 and 2 is modeled as

$$\dot{m}_1'' = -\rho D_{12} \frac{d\omega_1}{dx} = -D_{12} \frac{d\rho_1}{dx} \tag{2.16}$$

where

$\dot{m}_1''$ = mass flux of component 1 (kg/m^2 s)
ω_1 = ρ_1/ρ
D_{12} = mass diffusivity of component 1 in component 2 (m^2/s)
ρ_1 = concentration of component 1 (kg/m^3)

2.4.6 Electrical Conduction: Ohm's Law

In a process very similar to the conduction of heat caused by a temperature gradient, electricity is conducted due to a potential or voltage gradient. Experiments indicate that the current flow is predictable by the following relation, known as Ohm's law:

$$J = -\sigma \frac{dV}{dx} \tag{2.17}$$

where

J = current density (A/m^2)
σ = electrical conductivity (A/m V)
V = voltage (V)

2.4.7 Stress-Strain: Hooke's Law

In an elastic material, the stress or force per area is observed to be proportional to the strain or relative displacement. The proportionality constant between stress and strain is the modulus of elastic or Young's modulus, resulting in the following relation, referred to as Hooke's law:

$$\sigma = E \cdot \varepsilon \tag{2.18}$$

where

σ = stress (N/m^2)
ε = strain (m/m)
E = modulus of elasticity (N/m^2)

2.5 DIFFUSION ANALOGIES

The transport or flow of a conserved substance from regions of high concentration to regions of low concentration is referred to as *diffusion*. All diffusion rate laws have the form

$$J = -D \frac{d\phi}{dx} \tag{2.19}$$

where

J = flux = conserved quantity/time/area
ϕ = concentration of a conserved quantity (amount per volume)
x = position (m)
D = diffusivity (m^2/s)

The application of this general concept to some fundamental processes is summarized in Table 2.1. These are all constitutive relations relating flows to potential gradients.

Table 2.1 Diffusion rate laws for some selected physical processes

Process	*Conserved Quantity*	*Rate Equation*	*Diffusivity (m^2/s)*
Heat conduction: Fourier's law	Thermal energy	$q''_{cond} = -k\dfrac{dT}{dx} = -\alpha\dfrac{d(\rho cT)}{dx}$	$\alpha = \dfrac{k}{\rho c}$
Viscous fluid shear: Newton's viscosity law	Momentum	$\tau = \mu\dfrac{dV}{dy} = \nu\dfrac{d(\rho V)}{dy}$	$\nu = \dfrac{\mu}{\rho}$
Binary mass diffusion: Fick's law	Mass of species-1	$\dot{m}''_1 = -\rho D_{12}\dfrac{d\omega_1}{dx} = -D_{12}\dfrac{d\rho_1}{dx}$	D_{12}
Electrical conduction: Ohm's law	Electrical charge	$J = -\sigma\dfrac{dV}{dx} = -\alpha_e\dfrac{d(C_eV)}{dx}$	$\alpha_e = \dfrac{\sigma}{C_e}$

Chapter 3

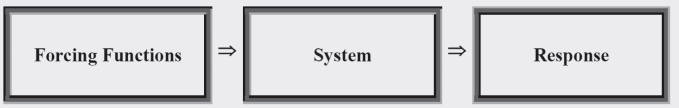
Modeling of Physical Processes

CHAPTER OBJECTIVES

This chapter describes the process of using the basic principles of physics, combined with appropriate boundary condition, initial conditions, and approximations, to form a complete and well-posed mathematical model of a system.

Specific objectives and topics covered are

- Cause and effect
- Mathematical modeling
- Complete mathematical models for classic vibrations and heat transfer problems
- Dimensionless formulations
- Inverse and parameter estimation problems
- Mathematical classification of physical problems



3.1 CAUSE AND EFFECT

A physical system is characterized by its material properties and geometry as well as the physical processes occurring within the system. Action or displacement from equilibrium is caused by external forces or stimuli acting on the system. The response to these external stimuli is in the form of flows and potentials, for example, heat flow and temperature change in a thermal system or motion for a mechanical system.

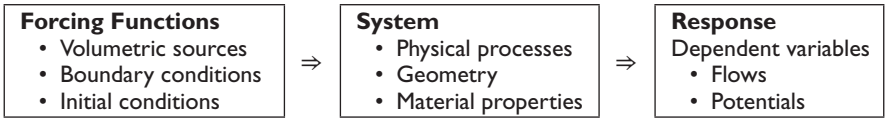
3.1.1 General Physical Process

In general, a mathematical solution is an equation or rule that is represented in the form

$$\text{Dependent variables} = \text{function}\left(\begin{matrix}\text{Independent} \\ \text{variables}\end{matrix}, \begin{matrix}\text{System} \\ \text{parameters}\end{matrix}, \begin{matrix}\text{Forcing} \\ \text{functions}\end{matrix}\right)$$

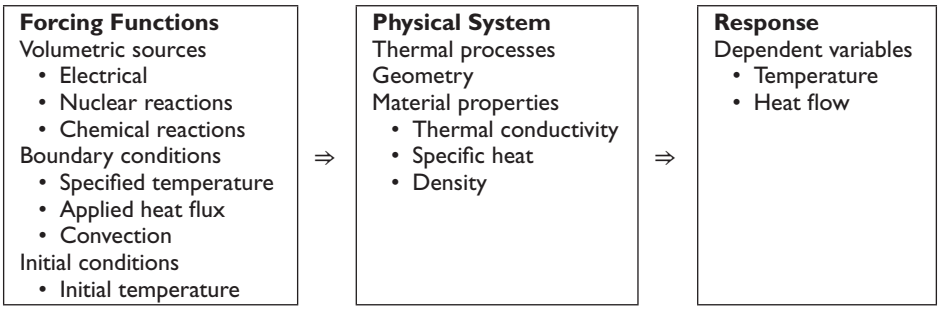
The *dependent variables* describe the response or state of the system. The *independent variables* are usually dimensions, such as location or time. The *system parameters* characterize the system's properties or composition. The *forcing functions* are external stimuli acting on the system. Together, these ingredients form a mathematical model.

This relationship between the components of the model is depicted in the following cause and effect diagram for an arbitrary physical system.

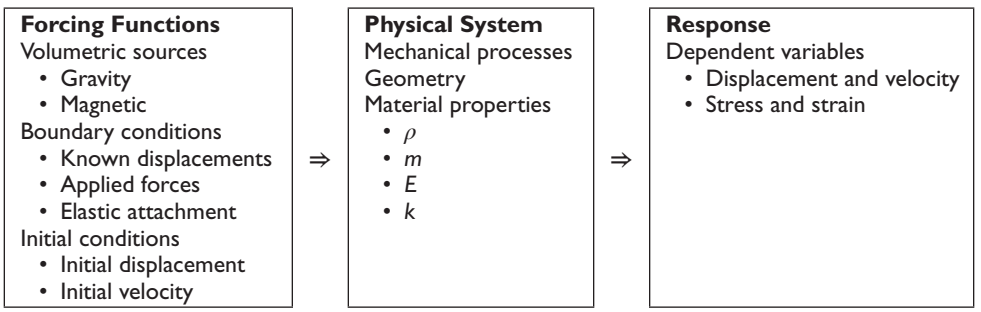


This is the fundamental relationship between *cause* (*forcing functions*) and *effect* (*response*). Forcing functions consist of three basic types: volumetric sources, boundary conditions, and initial conditions. All physical processes are caused by some combination of these forcing functions acting on the system. A system in equilibrium, with all forcing functions equal to zero, simply remains in equilibrium. Once stimulated by forcing functions, the system responds according to the physical laws of nature, as described in Chapter 2. The cause and effect schematics for thermal and mechanical systems are depicted in the following diagrams.

3.1.2 Thermal Processes



3.1.3 Mechanical Processes



3.2 MATHEMATICAL MODELING

For engineers and applied mathematicians, the ultimate purpose of mathematics is to describe and predict the behavior of physical systems. The basic steps involve the translation of the physical processes acting on a system into a mathematical model, followed by the solution of the model equations. This is represented schematically in the cause and effect diagrams in Section 3.1.

Formulating a mathematical model for a complex system can be challenging. The ingredients are shown in the schematic in Figure 3.1.

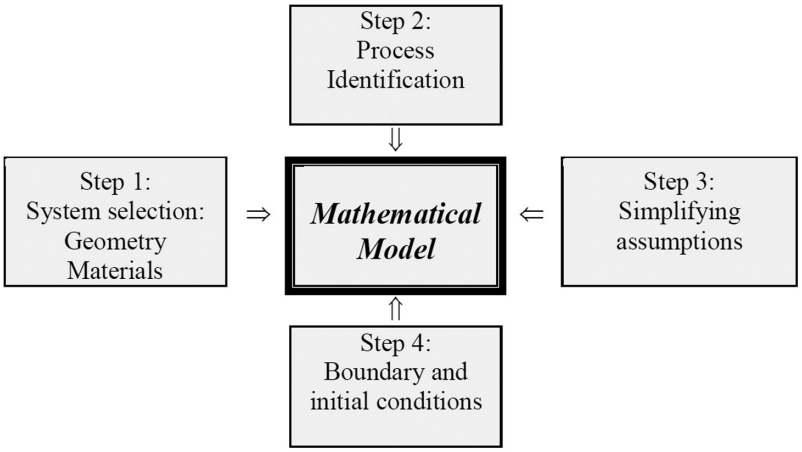


Figure 3.1 Overview of the mathematical modeling process.

The selection of a meaningful and useful system for mathematical analysis is dependent on the problem objectives. Any system, from the entire universe to the smallest differential control volume, could be selected. When analyzing a power plant, the entire plant, a boiler, a boiler tube, or any other subsystem can be selected.

No foolproof formula or procedure can be spelled out to select the best system to achieve the desired objectives. System selection is somewhat of an art form and requires experience. Most would agree that in addition to training in the standard surgical procedures, a skilled surgeon must have experience. Most of us, if it were our brain that needed the surgery, would not feel comfortable with an inexperienced surgeon fresh out of medical school.

Using a sound knowledge of the fundamental principles of physics, the processes occurring in the system of interest should be identified. Processes that are most important to the analysis and those that can be safely neglected should be identified. Knowing which processes are dominant requires a combination of experience and estimation. *Order-of-magnitude* or *dimensional analysis* can be used to perform this estimation. A classic example is *boundary layer theory* in fluid mechanics, where an order-of-magnitude analysis can be used to show that streamwise diffusion-type terms can be neglected compared with streamwise advection terms.

A major component of the *art of engineering* is the ability to reduce a problem to its simplest yet still accurate form. Simply following a prescribed procedure or formula every time will not be sufficient to analyze challenging problems of engineering and applied mathematics.

A variety of approximations can be made in order to simplify the mathematical representation of the system of interest. Some of these include

- Spatially uniform or spatially distributed (zero-, one-, two-, or three-dimensional)
- Spatially finite or infinite
- Steady or transient
- Continuous or discrete

- Linear or nonlinear
- Viewpoint: *stationary* (Eulerian) or *moving* (Lagrangian)
- Symmetry

One could always model every problem as generally as possible—as three-dimensional, transient, and nonlinear. However, this overkill approach would amount to a lot of unnecessary analysis and computation for most problems. More importantly, simply pounding out solutions with no insight or interpretation is a dangerous and scary way to approach engineering and design problems. The best engineer utilizes a good balance of physical insight and mathematical skills.

A complete mathematical description of problems involving spatially varying quantities requires the specification of appropriate conditions at the boundaries. These so-called *boundary conditions* are usually in the form of specified potentials, specified gradients, or a mixture of specified potential and gradients.

In thermal heat conduction problems where temperature is the unknown variable, the classical boundary conditions are

1. Specified temperature
2. Specified heat flux
3. Convection exchange with a fluid

In mechanical problems where displacement is the unknown variable, possible boundary conditions are

1. Specified displacement
2. Specified force
3. Elastic attachment

Problems with only time dependence and no spatial dependence are called *lumped system* or *lumped capacity problems*. Here, no boundary conditions are required.

For transient or time-dependent problems, the initial state of the system must be specified in order to give a unique or properly posed mathematical formulation. First-order differential equations in time require the initial position. Second-order differential equations, such as oscillators and wave equations, require the specification of both the initial position and the velocity.

3.3 COMPLETE MATHEMATICAL MODEL

For systems described by differential equations, a complete mathematical model must include the following:

- Governing differential or algebraic equations
- Boundary conditions (for spatially distributed systems)
- Initial conditions (for transient problems)

Problems with no spatial distribution, referred to as *lumped parameter systems*, have no mathematical boundary conditions. Any boundary effects will be included directly in the

differential equations. Likewise, steady-state problems do not require initial conditions, since all memory of past states is long forgotten.

The complete mathematical models of two classic problems, mechanical vibrations and heat diffusion, are described next.

3.3.1 Mechanical Vibrations

Consider a mass attached to a linear spring with spring constant k and a damper with damping coefficient c. The mass is acted on by an applied force. The initial displacement is x_0, and the initial velocity is v_0. The system and the free body diagram are displayed in Figure 3.2. The objective is to determine the displacement of the mass, $x(t)$. The displacement is measured from static equilibrium.

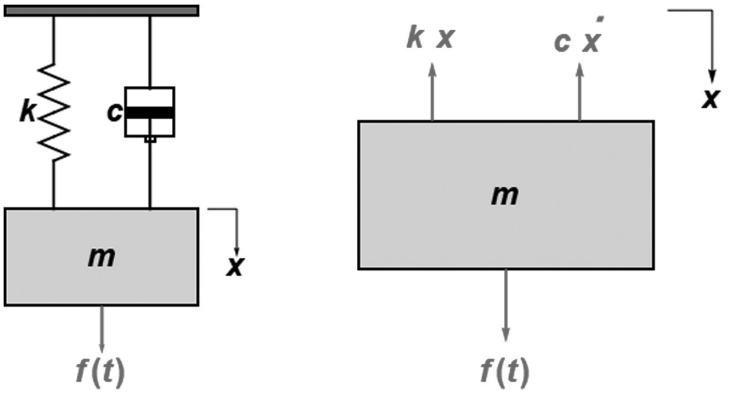


Figure 3.2 Mechanical oscillator.

The components are organized using a cause–effect diagram.

Forcing Functions $f(t), x_0, v_0$	⇒	**System** m, k, c	⇒	**Response** $x(t)$

The goal is to find a solution of the form

$$x = \text{function}\left(t, m, k, c, f(t), x_0, v_0\right)$$

In order to accomplish this goal, the equations governing the motion of the system must be formulated. Apply Newton's second law using the free body diagram of the mass shown on the right in Figure 3.2 to deduce

$$\text{Mass} \times \text{Acceleration} = \sum \text{Forces}$$

$$m \cdot a = m\frac{d^2x}{dt^2} = -c\frac{dx}{dt} - kx + f(t) \tag{3.1}$$

The complete mathematical model of the system consists of the previous equation of motion along with the necessary initial conditions. The complete mathematical model is

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = f\left(t\right) \tag{3.2}$$

$$\left.\begin{array}{l} x = x_0 \\ \dfrac{dx}{dt} = v_0 \end{array}\right\} \; t = 0 \tag{3.3}$$

Since the governing differential equation is a second-order initial value problem in time, specification of both the initial displacement and velocity is required. There are no mathematical boundary conditions required, since the system is spatially lumped. That is, every position in the mass shares the same displacement and velocity. The applied forces are physically applied at the surface of the mass but show up mathematically in the spatially lumped differential equation.

3.3.2 Heat Conduction

We next examine the heat conduction problem displayed in Figure 3.3.

The problem consists of transient heat transfer in a solid material under the following conditions and approximations:

- rectangular region of length L with one-dimensional heat flow, that is, temperature gradients only in one direction
- initially at uniform temperature T_0
- constant thermal properties
- volumetric heat source, $g(x,t)$
- applied heat flux, $q_s''(t)$, at $x = 0$
- constant temperature, T_L, at $x = L$

The objective is to derive the energy equation governing the transient temperature distribution and to formulate the complete mathematical model of the system including appropriate boundary and initial conditions.

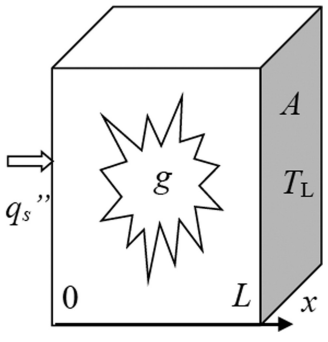


Figure 3.3 Heat conduction schematic.

In a manner similar to applying Newton's second law to a *free body diagram* in a mechanical system, we apply the conservation of energy principle to a *control volume* in a thermal system. Since we are interested in the spatial distribution in the x-direction but are neglecting gradients in the other directions, the control volume of size $\Delta x - by - A$ shown in Figure 3.4 is appropriate.

Conservation of energy applied to our differential control volume requires

$$\rho \frac{\partial e}{\partial t} A \cdot \Delta x = q_x - \left(q_x + \frac{\partial q_x}{\partial x} \Delta x \right) + g \cdot A \cdot \Delta x \tag{3.4}$$

where

e = specific internal energy (energy/mass ~J/kg)

$q_x = Aq_x''$ = heat flow rate by conduction (energy/time ~W)

Note that all terms have units of energy per time (W). Cancel the volume, $A \cdot \Delta x$, in Equation 3.4 and simplify to get

$$\rho \frac{\partial e}{\partial t} = -\frac{\partial q_x''}{\partial x} + g \tag{3.5}$$

We now have one equation in two unknown variables, e and q_x''. We would like to express this energy equation in terms of a single, measurable variable. Thus, both e and q_x'' are expressed in terms of temperature, T. Specific internal energy is related to temperature through the physical property specific heat as

$$de = c \cdot dT \tag{3.6}$$

where c = specific heat (energy/mass/temperature change ~ J/kg/K)

As described in Section 2.4.1, heat flux is related to temperature using Fourier's *rate law* via

$$q_x'' = -k \frac{\partial T}{\partial x} \tag{3.7}$$

Substitute Equations 3.6 and 3.7 into the energy conservation equation (Equation 3.5) to obtain the *heat diffusion equation*

$$\rho c \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + g \tag{3.8}$$

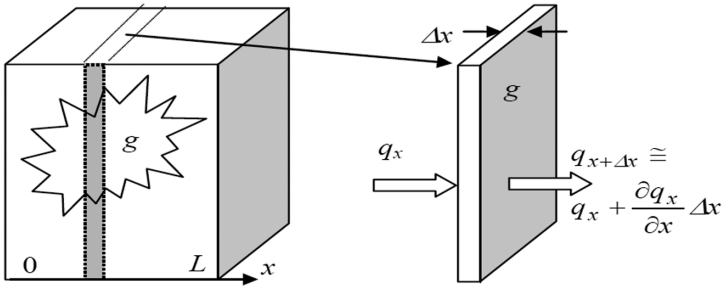


Figure 3.4 Control volume for the heat conduction equation.

For constant properties, the previous heat diffusion equation reduces to

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + \frac{g}{\rho c} \tag{3.9}$$

$$\alpha = \frac{k}{\rho c} = \text{thermal diffusivity } \left(\text{m}^2/\text{s}\right)$$

The heat diffusion equation (Equation 3.9) has a second-order spatial derivative and thus requires two boundary conditions, one on each boundary. It also has a first-order time derivative and thus requires one initial condition—the initial temperature distribution. Thus, the *complete mathematical model* consists of the following energy equation, boundary conditions, and initial condition.

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + \frac{g}{\rho c} \tag{3.10}$$

$$\begin{aligned} -k\frac{\partial T}{\partial x} &= q_s''(t), \quad x = 0 \\ T &= T_L, \quad x = L \end{aligned} \tag{3.11}$$

$$T = T_0(x), \quad t = 0 \tag{3.12}$$

Boundary and initial conditions are essential elements for a *complete* and *well-posed* mathematical model of a physical system. Without them, the problem is not uniquely specified. On the other hand, extra and unnecessary boundary or initial conditions make the problem over specified.

3.4 DIMENSIONLESS FORMULATION

3.4.1 General Procedure

A general procedure to convert mathematical models to dimensionless form is outlined.

***Step* 1:** List all variables and parameters along with their units.

	Symbol	*Units*
Dependent variables	θ	
	…	
Independent variables	x	
	t	
	…	
Parameters	p_1	
	p_2	
	…	

***Step* 2:** Count the number of independent dimensions (or units) in the problem. These are usually *m*, *s*, *J*, *kg*, etc.

***Step* 3:** Choose a number of independent reference parameters, $p_{\text{ref},1}$, $p_{\text{ref},2}$, $p_{\text{ref},3}$,…, equal to the number of dimensions.

***Step* 4:** Form dimensionless variables from the reference parameters.

$$\theta_{\text{ref}} = \left(p_{\text{ref},1}\right)^{e_1}\left(p_{\text{ref},2}\right)^{e_2}\cdots$$

$$x_{\text{ref}} = \left(p_{\text{ref},1}\right)^{e_1}\left(p_{\text{ref},2}\right)^{e_2}\cdots$$

Dimensionless dependent variables: $\theta^+ = \dfrac{\theta}{\theta_{\text{ref}}}$

Dimensionless independent variables: $x^+ = \dfrac{x}{x_{\text{ref}}}$

***Step* 5:** Substitute dimensionless variables to obtain the dimensionless formulation. Identify dimensionless parameter groups that naturally emerge. Specifically determine:

- Dimensionless differential equation
- Dimensionless boundary conditions
- Dimensionless initial condition
- Dimensionless parameters

3.4.2 Mechanical Vibrations

The general procedure outlined in the previous section is now applied to the classical mass-spring-damper problem. The mathematical model is given by Equations 3.2 and 3.3 with a constant applied force. The model equations are

$$M\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k\cdot x = f_0 \tag{3.13}$$

$$\left.\begin{aligned} x &= x_0 \\ \frac{dx}{dt} &= v_0 \end{aligned}\right\} \quad t = 0 \tag{3.14}$$

***Step* 1:** List all the variables and parameters along with their units.

		Symbol	*Units*
Dependent variable	Displacement	x	m
Independent variable	Time	t	s
Parameters	Mass	M	$\text{kg} = \dfrac{\text{N}}{\text{m/s}^2} = \dfrac{\text{N}\cdot\text{s}^2}{\text{m}}$
	Damping coefficient	c	$\dfrac{\text{N}}{\text{m/s}} = \dfrac{\text{N}\cdot\text{s}}{\text{m}}$
	Spring constant	k	N/m
	Applied force	f_0	N
	Initial displacement	x_0	m
	Initial velocity	v_0	m/s

***Step* 2:** List and count the number of independent dimensions in the problem.

Note that *kg* is related to *N*, *s*, and *m* and is not an independent dimension. We have three independent dimensions: *N*, *m*, and *s*.

***Step* 3:** Choose independent reference parameters equal to the number of dimensions.

We have numerous choices as long as all three dimensions (*N*, *m*, and *s*) are included. Thus, the combination k, f_0, and x_0 would not be valid, since the dimension (*s*) is not included. We have to choose three parameters from a list of six. This gives $6 \times 5 \times 4 = 120$ combinations. The combination k, f_0, and x_0 is invalid, leaving us with 119 possible ways to normalize this problem!

Let's choose M, k, and f_0.

***Step* 4**: Form the dimensionless variables.

$$x_{\text{ref}} = M^{e1} \cdot k^{e2} \cdot {f_0}^{e3}$$

$$m \sim (\text{kg})^{e1} \cdot \left(\frac{\text{N}}{\text{m}}\right)^{e2} \cdot (\text{N})^{e3} \tag{3.15}$$

Consistent dimensions require $e2 = -1$, $e3 = 1$, and $e1 = 0$. Thus,

$$x_{\text{ref}} = \frac{f_0}{k}$$

Similarly, we find

$$t_{\text{ref}} = \sqrt{M / k}$$

Now we form our dimensionless dependent variable: $x^+ = \dfrac{x}{x_{\text{ref}}} = x\dfrac{k}{f_0}$

and dimensionless independent variable: $t^+ = \dfrac{t}{t_{\text{ref}}} = t\sqrt{k/M}$

***Step* 5:** Substitute and identify dimensionless parameters.

$$M\frac{d^2\left(x_{\text{ref}} \cdot x^+\right)}{dt^{+2}}\frac{1}{{t_{\text{ref}}}^2} + c\frac{d\left(x_{\text{ref}} \cdot x^+\right)}{dt^+}\frac{1}{t_{\text{ref}}} + k \cdot \left(x_{\text{ref}} \cdot x^+\right) = f_0$$

The chain rule was used. Divide by $Mx_{\text{ref}}\dfrac{1}{{t_{\text{ref}}}^2}$.

$$\frac{d^2x^+}{dt^{+2}} + c\frac{t_{\text{ref}}}{M}\frac{dx^+}{dt^+} + k\frac{{t_{\text{ref}}}^2}{M} \cdot x^+ = f_0\frac{1}{Mx_{\text{ref}}}{t_{\text{ref}}}^2$$

Use the definitions of x_{ref} and t_{ref} and simplify.

$$\frac{d^2x^+}{dt^{+2}} + \frac{c}{\sqrt{Mk}}\frac{dx^+}{dt^+} + x^+ = 1$$

A dimensionless damping coefficient has emerged.

$$\zeta = \frac{c}{2\sqrt{Mk}}$$

The dimensionless formulation of Equations 3.13 and 3.14 is

$$\frac{d^2x^+}{dt^{+2}} + 2\zeta\frac{dx^+}{dt^+} + x^+ = 1 \tag{3.16}$$

$$\left.\begin{aligned} x^+ &= x_0{}^+ \\ \frac{dx^+}{dt^+} &= v_0{}^+ \end{aligned}\right\} \; t^+ = 0 \tag{3.17}$$

The dimensionless parameters are

$$\zeta = \frac{c}{2\sqrt{Mk}}, \quad x_0{}^+ = x_0\frac{k}{f_0}, \quad v_0{}^+ = v_0\frac{\sqrt{Mk}}{f_0} \tag{3.18}$$

Our choice of reference parameters did not include the damping coefficient. Thus, a dimensionless damping parameter wound up in the transformed equations. This allows a clear and meaningful study of the effects of damping.

3.4.3 Steady Heat Conduction

The general procedure is now applied to a basic steady-state heat transfer problem. The model equations are the following:

	Variable shift $\theta(x) = T(x) - T_L$
$k\dfrac{d^2T}{dx^2} + g = 0$	$k\dfrac{d^2\theta}{dx^2} + g = 0$
$-k\dfrac{dT}{dx} = q_0'', \quad x = 0$ $T = T_L, \quad x = L$	$-k\dfrac{d\theta}{dx} = q_0'', \quad x = 0$ $\theta = 0, \quad x = L$

Note that we have shifted or "grounded" the problem before starting any analysis by defining a new variable $\theta(x)$ as the temperature rise above the ground or sink temperature T_L.

***Step* 1:** List all the variables and parameters along with their units.

		Symbol	*Units*
Dependent variable	Temperature rise	θ	K
Independent variable	Position	x	m
Parameters	Region length	L	m
	Volumetric heat generation	g	W/m^3
	Surface heat flux	q_0''	W/m^2
	Thermal conductivity	k	W/(m K)

***Step* 2:** List and count the number of independent dimensions in the problem.

We have three independent dimensions: K, m, and W.

***Step* 3:** Choose independent reference parameters equal to the number of dimensions.

We must choose k, since it is the only parameter with K. Choosing k along with any two of the other three parameters gives us six combinations.

Let's choose: k, L, and q_0''.

***Step* 4**: Form the dimensionless variables.

$$\theta_{\text{ref}} = k^{e1} \cdot L^{e2} q_0''^{e3}$$

$$K \sim \left(\frac{\text{W}}{\text{m} \cdot \text{K}}\right)^{e1} \cdot (\text{m})^{e2} \cdot \left(\frac{\text{W}}{\text{m}^2}\right)^{e3}$$

Consistent dimensions require $e1 = -1$, $e3 = 1$, and $e1 = 1$. Thus,

$$\theta_{\text{ref}} = \frac{q_0'' L}{k}$$

Similarly, we find

$$x_{\text{ref}} = L$$

Now we form our dimensionless dependent variable: $\theta^+ = \dfrac{\theta}{\theta_{\text{ref}}} = \theta \dfrac{k}{q_0'' L}$

and dimensionless independent variable: $x^+ = \dfrac{x}{x_{\text{ref}}} = \dfrac{x}{L}$

***Step* 5:** Substitute and identify dimensionless parameters.

$$k \frac{d^2\left(\theta_{\text{ref}} \cdot \theta^+\right)}{dx^{+2}} \frac{1}{x_{\text{ref}}{}^2} + g = 0$$

The chain rule was used. Divide by $k\theta_{\text{ref}} \dfrac{1}{x_{\text{ref}}{}^2}$ and use the definitions of θ_{ref} and x_{ref}.

$$\frac{d^2\theta^+}{dx^{+2}} + g\frac{L}{q_0''} = 0$$

A dimensionless volumetric heat generation has emerged.

$$g^+ = g\frac{L}{q_0''}$$

The dimensionless formulation and dimensionless parameters are

$$\frac{d^2\theta^+}{dx^{+2}} + g^+ = 0 \tag{3.19}$$

$$\begin{aligned} -\frac{d\theta^+}{dx^+} &= 1, \quad & x^+ &= 0 \\ \theta^+ &= 0, \quad & x^+ &= 1 \end{aligned} \tag{3.20}$$

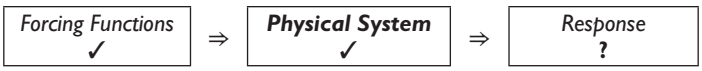
$$g^+ = g\frac{L}{q_0''} \tag{3.21}$$

Our choice of reference parameters did not include the heat generation. Thus, a dimensionless heat generation wound up in the transformed equations. This allows for a clear and meaningful study of the heat generation.

3.5 INVERSE AND PARAMETER ESTIMATION PROBLEMS

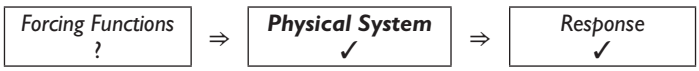
The difference between *direct*, *inverse*, and *parameter estimation* problems is shown in the following cause and effect schematics.

3.5.1 Direct Problem



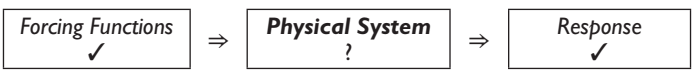
In this model, the system and all forcing functions are completely specified. The task is to determine the system response.

3.5.2 Inverse Problem



In this model, it is assumed that the physical system is specified. This time, however, the response is also known, or at least partially known in the form of measurements at discrete times and locations. The goal is to find some unknown forcing functions.

3.5.3 Parameter Estimation Problem



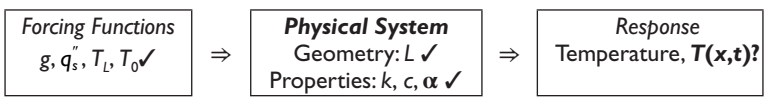
Here again, the system response is known, or at least partially known in the form of measurements at discrete times and locations. The goal this time is to estimate some unknown system parameters, such as material properties.

The inverse and parameter estimation problems are considerably more difficult to solve than the direct problems due to, among other things, stability problems and lack of uniqueness.

Example: Heat Conduction

Consider the heat conduction problem from Section 3.3.2. The mathematical model is given by Equations 3.10–3.12.

The direct problem is:



This is the typical problem encountered in engineering analysis. The forcing functions and system parameters are specified, and the goal is to solve for the temperature response.

The inverse problem is:

Forcing Functions	⇒	***Physical System***	⇒	*Response*
g, T_L, T_0 ✓		Geometry: L ✓		Temperature, $T(x,t)$
q_s'' ?		Properties: k, c, α ✓		✓

In this variation, the temperature response is known, usually from measurements at discrete times and locations. The goal might be to find some unknown forcing function such as the boundary heat flux, q_s''. This type of problem arises in the thermal analysis of spacecraft reentering the earth's atmosphere.

The parameter estimation problem is:

Forcing Functions	⇒	***Physical System***	⇒	***Response***
q_s'', g, T_L, T_0 ✓		Geometry: L✓		Temperature, $T(x,t)$ ✓
		Properties: k, c, α?		

Here, the temperature response, geometry, and forcing functions are known. The goal is to estimate the physical properties. This type of problem is used to estimate unknown properties of some material, such as a new composite.

3.6 MATHEMATICAL CLASSIFICATION OF PHYSICAL PROBLEMS

A dynamical view of the world is shown in Figure 3.5. The various dynamical systems are categorized by the number of variables on one axis and linear versus nonlinear on the other axis. The shaded areas in the lower right could be regarded as the frontier of current research.

PROBLEMS

Problem 3.1

Determine the solutions to the following problems.

a) $\dfrac{d\theta}{dt} = -\dfrac{1}{\tau}\theta$

$\theta = 0, t = 0$

b) $m\dfrac{d^2x}{dt^2} + c\dfrac{dx}{dt} + k \cdot x = 0$

$x = 0, \ \dfrac{dx}{dt} = 0, \ t = 0$

c) $\dfrac{\partial^2 T}{\partial x^2} + \dfrac{\partial^2 T}{\partial y^2} = 0$

$T = 0$ on all boundaries

d) $\dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2}$

$\dfrac{\partial T}{\partial x} = 0, \quad x = 0$

$T = 0, \quad x = L$

$T = 0, \ t = 0$

	Number of Variables				
	n = 1 **Growth and Decay**	**n = 2** **Oscillations**	**n ≥ 3**	**n >> 1** **Collective Phenomena**	**Continuum** **Waves and Patterns**
Linear	Exponential growth RC circuit Lumped thermal problems Radioactive decay	Linear oscillator Mass and spring RLC circuit 2-body problem (Kepler, Newton)	Civil engineering Electrical engineering Mechanical engineering	Coupled harmonic oscillators Solid state physics Molecular dynamics Equilibrium statistical mechanics	Elasticity Wave mechanics, Acoustics Electromagnetism (Maxwell) Quantum mechanics (Schrodinger, Heisenberg) Heat diffusion Viscous fluids
Non-Linear	Fixed points Bifurcations Overdamped systems, relaxation dynamics Logistic equation	Pendulum Anharmonic oscillators Limit cycles Biological oscillators (neurons, heart cells) Predator-prey cycles Nonlinear electronics	Chaos Strange attractors (Lorenz) 3-body problem (Poincare) Chemical kinetics Iterated maps (Feigenbaum) Fractals (Mandelbrot) Forced nonlinear oscillators (Levinson, Smale) Quantum chaos Practical uses of chaos	Coupled nonlinear oscillators Lasers, nonlinear optics Nonequilibrium statistical mechanics Nonlinear solid state physics (semiconductors), Josephson arrays Heart cell synchronization, Immune system Neural networks Ecosystems Economics	Nonlinear waves (shocks, solitions) Plasmas Earthquakes General relativity (Einstein) Quantum field theory Reaction-diffusion, biological, chemical waves Fibrillation, Epilepsy Turbulent fluids (Navier-Stokes) Life

Figure 3.5 A dynamical view of the world. (Reproduced from *Nonlinear Dynamics and Chaos*, 2nd edition (2014), by S.H. Strogatz. Courtesy of Taylor & Francis Books.)

Problem 3.2: Boundary and Initial Conditions

Consider the following systems of equations. In each case, identify whether the mathematical model is formulated properly in order to allow the possibility of a unique solution. If the system is not properly formulated, what additions or changes are necessary?

a) $$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S$$

b) $$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S$$
$$\theta = \theta_0, \quad t = 0$$
$$\theta = \theta_1, \quad t = t_1$$

c) $$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = f(t)$$
$$x = x_0, \quad t = 0$$

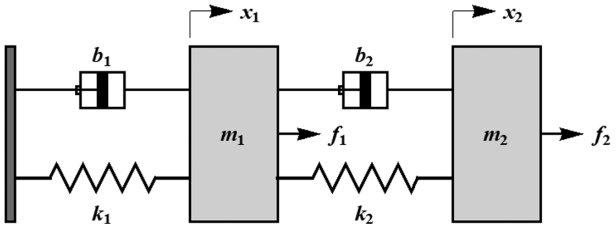
d) $$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = f(t)$$
$$x = x_0, \quad t = 0$$
$$x = x_1, \quad t = t_1$$

e) $$\frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2} + \frac{g(x,t)}{\rho c}, \quad 0 < x < L$$
$$-k\frac{\partial T}{\partial x} = q_0''(t), \quad x = 0$$

f) $$\frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2} + \frac{g(x,t)}{\rho c}, \quad 0 < x < L$$
$$-k\frac{\partial T}{\partial x} = q_0''(t), \quad x = 0$$
$$T = T_0(t), \quad x = 0$$
$$T = T_0, \; t = 0$$

Problem 3.3: Coupled Spring-Mass-Dampers

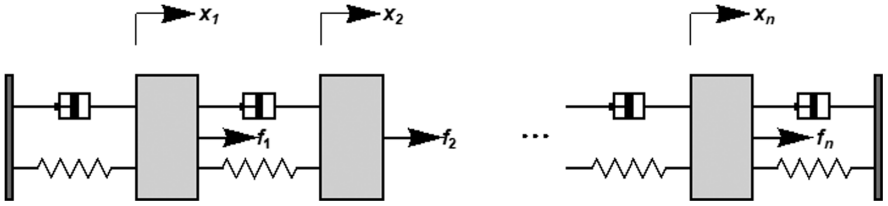
Consider the following system with two sets of mass-spring-dampers. A force f_1 is applied directly to mass-1, while a force f_2 is applied to mass-2. The spring constants are k_1 and k_2, and the damping coefficients are b_1 and b_2.The equilibrium positions with f_1=f_2=0 correspond to x_1=x_2=0.



Derive the complete mathematical model of this system.

Problem 3.4: A Series of Spring-Mass-Dampers

A series of coupled spring-mass-dampers consisting of n masses in series are connected by springs and dampers. The masses are identical, the springs are identical, and the dampers are identical. Thus, we only need single values of m, k, and b.



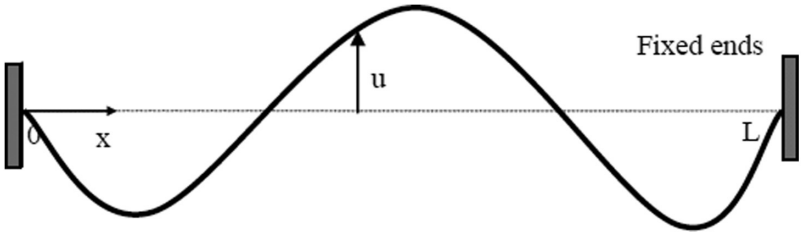
The initial conditions are

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_n(0) \end{bmatrix} = \begin{bmatrix} x_{1,0} \\ x_{2,0} \\ \vdots \\ x_{n,0} \end{bmatrix} \text{ and } \begin{bmatrix} \dot{x}_1(0) \\ \dot{x}_2(0) \\ \vdots \\ \dot{x}_n(0) \end{bmatrix} = \begin{bmatrix} v_{1,0} \\ v_{2,0} \\ \vdots \\ v_{n,0} \end{bmatrix}$$

Each mass has an individual external force applied, $f_i(t)$. The displacements are measured from static equilibrium with no applied forces. Derive the mathematical model of the system for the displacements x_i, for $i = 1$ to n.

Problem 3.5

Consider the motion of an elastic string.



$u(x,t)$ = transverse displacement (m)
T = tension (N)
ρ = density (kg/m^3)
A_c = cross-sectional area (m^2)
$F(x,t)$ = distributed load (N/m)

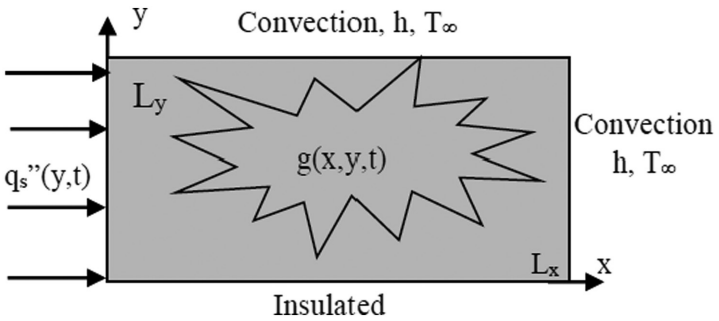
Assumptions

- The motion takes place entirely in one plane, and in this plane, each particle moves at right angles to the equilibrium position of the string.
- The deflection of the string during the motion is so small that the resulting change in length of the string has no effect on the tension T.
- The string is perfectly flexible; that is, it can transmit force only in the direction of its length.
- The slope of the deflection curve is small, so that $sin\theta$ can be replaced with $tan\theta$, where θ is the inclination angle of the tangent to the deflection curve.
- Derive the equation governing the transverse displacement of the string, $u(x,t)$.
- State the complete mathematical model necessary to determine the motion of the string.

- For the special case where the distributed load is zero, $F = 0$, and the string is initial in equilibrium (the initial displacement and velocity are zero), determine the solution $u(x,t)$ for the displacement of the string.
- What is the relationship between this mathematical model and the series of discrete spring-mass-dampers in Problem 3.4?

Problem 3.6

Consider transient heat transfer in a two-dimensional region (no temperature gradients in the z-direction), initially at temperature T_0, with the boundary conditions shown.



a) Formulate the mathematical model governing the transient temperature distribution, $T(x,y,t)$.
b) Assuming that temperature gradients are negligible in the x-direction, formulate the mathematical model governing the temperature distribution.
c) Assuming that temperature gradients are negligible in the y-direction, formulate the mathematical model governing the temperature distribution.
d) Assuming that temperature gradients are negligible in all directions (lumped capacity approximation), formulate the mathematical model governing the transient temperature distribution $T(t)$. For this lumped capacity model only, determine the steady state temperature if q_s'' and g are constant.

Chapter 4

Calculus

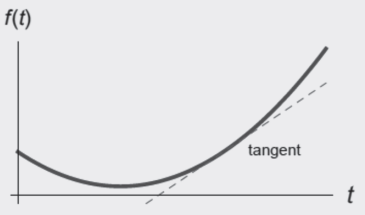
CHAPTER OBJECTIVES

First, the fundamental concept of a derivative is presented. This leads naturally to the chain rule, the product rule, and the concept of a partial derivative. The numerical evaluation of derivatives using Taylor series is presented.

Next, the fundamental concepts of integrals are presented. Also, the practical aspects of evaluating integrals using numerical integration are presented.

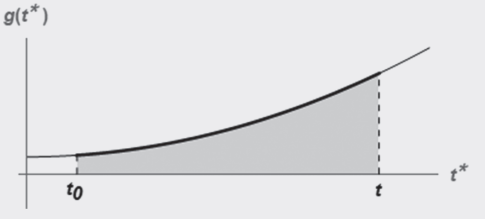
Specific objectives and topics covered are

- The basic concept of a derivative
- The chain rule
- Product rule
- Partial derivatives
- Numerical differentiation using the Taylor series expansion
- The basic concept of an integral
- The mean value theorem
- Integration by parts
- Leibniz rule—the derivative of an integral
- The step, impulse, and delta functions
- Numerical integration using the trapezoid rule, Simpson's rules, and Gauss quadrature
- Multiple integrals



The derivative $\frac{df}{dt}$

is the *slope* of the curve $f(t)$.



The integral $\int_{t^*=t_0}^{t} g(t^*)\,dt^*$

is the *area* under the curve $g(t^*)$.

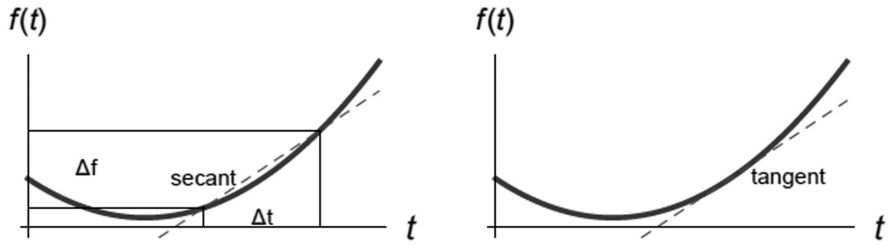


Figure 4.1 The geometric concept of a derivative.

4.1 DERIVATIVES

4.1.1 Basic Concept of a Derivative

In mathematics, the rate of change of a function is referred to as the *derivative*. For a function of a single variable, the derivative at a given point is the *slope of the tangent*, as visualized in Figure 4.1.

In order to estimate this slope, the function is evaluated at t and at a location Δt later. The secant between these points is then formed, as shown in the left portion of Figure 4.1. The slope of the secant is

$$\text{slope} \approx \frac{\Delta f}{\Delta t} = \frac{f(t+\Delta t) - f(t)}{\Delta t} \tag{4.1}$$

As Δt becomes small, the slope of the secant and the slope of the tangent at t become indistinguishable. The limit as Δt approaches zero produces an exact value for the rate of change and provides a formal definition for the derivative as

$$\frac{df}{dt} = \text{derivative of } f(t) \text{ with respect to } t = \lim_{\Delta t \to 0}\left(\frac{f(t+\Delta t) - f(t)}{\Delta t}\right) \tag{4.2}$$

This is called the *forward difference* form, since we are using a point t and a point $t+\Delta t$ in the forward direction. Since we are taking the limit as Δt approaches zero, we could use the equivalent *centered* or *backwards difference* forms given by

$$\begin{aligned} \frac{df}{dt} &= \lim_{\Delta t \to 0}\left(\frac{f(t+\Delta t/2) - f(t-\Delta t/2)}{\Delta t}\right) \\ &= \lim_{\Delta t \to 0}\left(\frac{f(t) - f(t-\Delta t)}{\Delta t}\right) \end{aligned} \tag{4.3}$$

4.1.2 Velocity from Displacement

Derivatives are used throughout applied mathematics to describe the rate of change of one quantity with respect to another. A classic example is velocity, defined by

$$v = \frac{dx}{dt} = \lim_{\Delta t \to 0}\left(\frac{x(t+\Delta t) - x(t)}{\Delta t}\right) \tag{4.4}$$

where

v is velocity
x is position
t is time

In a case where the function $x(t)$ is not known explicitly, but instead, numerical values of x are known at discrete times, the velocity can be estimated as

$$v = \frac{dx}{dt} \cong \frac{\Delta x}{\Delta t} = \frac{x\left(t+\Delta t\right) - x\left(t\right)}{\Delta t} \tag{4.5}$$

Note that this relation is approximate. If the x-values are sufficiently accurate, and the time interval Δt is sufficiently small, the previous approximation would be extremely accurate. Such an estimate is called a *finite difference* approximation. Various finite difference approximations of derivatives and their associated accuracy are derived in Section 4.5 using Taylor series.

4.1.3 Derivative of t^n

Consider the special case $f(t) = t^n$. In calculus class, we are presented with the formula

$$\frac{d\left(t^n\right)}{dt} = n \cdot t^{n-1} \tag{4.6}$$

Can we derive such a formula, or is it some sort of definition that must be memorized and accepted on blind faith? The most straightforward way to envision a derivative is with a finite difference approximation. Let's start with $n = 1$ and use the finite difference approximation

$$\frac{d\left(t^1\right)}{dt} \cong \frac{\left(t+\Delta t\right)^1 - t^1}{\Delta t} = 1 \tag{4.7}$$

We have obtained the expected result. Now, for $n = 2$,

$$\frac{d\left(t^2\right)}{dt} \cong \frac{\left(t+\Delta t\right)^2 - t^2}{\Delta t} = 2t + \Delta t \tag{4.8}$$

In the limit as $\Delta t \to 0$, we get

$$\frac{d\left(t^2\right)}{dt} = 2t \tag{4.9}$$

In a similar manner, the general case is

$$\frac{d\left(t^n\right)}{dt} \cong \frac{\left(t+\Delta t\right)^n - t^n}{\Delta t} = n \cdot t^{n-1} + \text{terms involving } \Delta t \tag{4.10}$$

In the limit as $\Delta t \to 0$, we get the well-known result

$$\frac{d\left(t^n\right)}{dt} = n \cdot t^{n-1} \tag{4.11}$$

Derivatives of other well-known functions can be derived in this way. All mathematical formulas come from a logical sequence or concept. Memorizing formulas with no idea of their origin or meaning should be avoided.

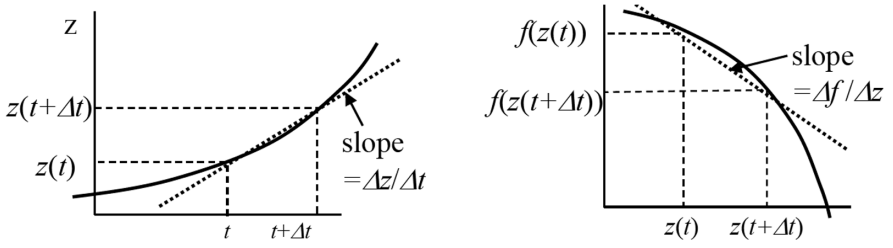


Figure 4.2 The geometric interpretation of the chain rule.

4.1.4 Chain Rule

There are many situations where we wish to find the derivative of a function with respect to t where the argument is a function of t. That is,

$$f = f\left(z(t)\right) \rightarrow \frac{df}{dt} = ? \tag{4.12}$$

This derivative is most easily envisioned by approximating it as a finite difference at discrete points in the form

$$\frac{d}{dt}\left(f\left(z(t)\right)\right) \cong \frac{\Delta f}{\Delta t} = \frac{\Delta f}{\Delta z} \cdot \frac{\Delta z}{\Delta t} \tag{4.13}$$

Now, in the limit as $\Delta t \rightarrow 0$, we obtain the chain rule

$$\frac{d}{dt}\left(f\left(z(t)\right)\right) = \frac{df}{dz} \cdot \frac{dz}{dt} \tag{4.14}$$

This logical extension of the basic concept of the derivative is known as the *chain rule*. The chain rule is visualized in Figure 4.2.

4.1.5 Product Rule

There are many applications where the derivative of the product of two functions is required. That is,

$$\frac{d(f \cdot g)}{dt} = ? \tag{4.15}$$

Sometimes, it is useful to expand this product. This procedure is easiest to understand by again using the fundamental notion of the finite difference approximation of the derivative to deduce

$$\frac{d(f \cdot g)}{dt} \cong \frac{f(t+\Delta t) \cdot g(t+\Delta t) - f(t) \cdot g(t)}{\Delta t} \tag{4.16}$$

Now, use the discrete approximation for the derivative of $f(t)$:

$$f(t+\Delta t) \cong f(t) + \frac{df}{dt}\Delta t \tag{4.17}$$

and a similar expression for $g(t)$ to deduce

$$\frac{d(f\cdot g)}{dt} \cong f\frac{dg}{dt} + g\frac{df}{dt} + \frac{df}{dt}\cdot\frac{dg}{dt}\Delta t \tag{4.18}$$

In the limit as $\Delta t \to 0$, the term involving Δt vanishes, and the following product rule is obtained.

$$\frac{d(f\cdot g)}{dt} = f\frac{dg}{dt} + g\frac{df}{dt} \tag{4.19}$$

4.1.6 Partial Derivatives

For functions of more than one independent variable, we often need the derivative with respect to only one of the dependent variables. This is called the *partial derivative*. For instance, consider a function of two independent variables

$$z = f(x,y) \tag{4.20}$$

In many physical problems, x and y would be position coordinates. They could also represent space and time coordinates. The partial derivative with respect to x is

$$\frac{\partial f}{\partial x} = \lim_{\Delta x\to 0}\left(\frac{f(x+\Delta x,y) - f(x,y)}{\Delta x}\right) \tag{4.21}$$

Note that the partial derivative with respect to x implies that the y value remains fixed. Also, the notation $\partial/\partial x$ is used to signify a partial derivative as opposed to d/dx for a total derivative. The partial derivative with respect to x at a given location (a,b) is shown geometrically in Figure 4.3.

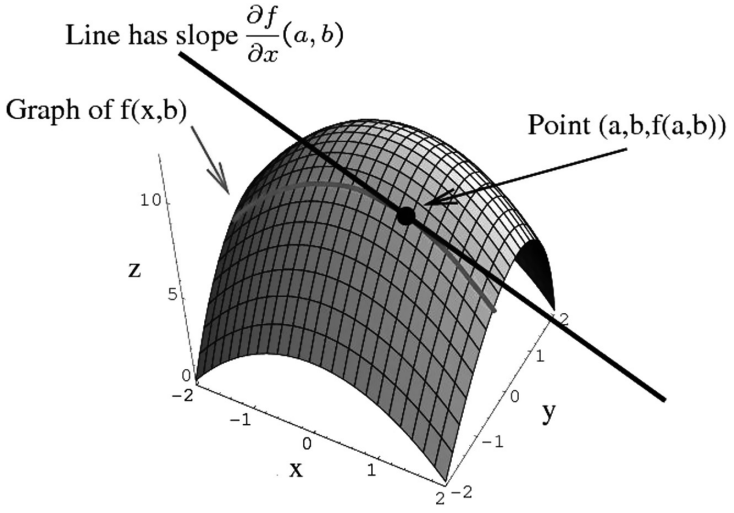


Figure 4.3 Partial derivatives.

In a similar fashion, the partial derivative with respect to y is

$$\frac{\partial f}{\partial y} = \lim_{\Delta y \to 0}\left(\frac{f(x, y+\Delta y) - f(x,y)}{\Delta y}\right) \tag{4.22}$$

4.2 NUMERICAL DIFFERENTIATION: TAYLOR SERIES

4.2.1 Taylor Series Expansion

The Taylor series expansion of the function $f(x)$ gives the value of a function at a point $x + \Delta x$ based on the function and its derivatives at point x. The general expression is

$$\begin{aligned} f(x+\Delta x) &= f(x) + \Delta x \cdot \frac{df}{dx} + \frac{\Delta x^2}{2!} \cdot \frac{d^2 f}{dx^2} + \cdots \\ &\quad + \frac{\Delta x^N}{N!} \cdot \frac{d^N f}{dx^N} + R_N \\ &= \sum_{n=0}^{\infty} \frac{\Delta x^n}{n!} \cdot \frac{d^n f}{dx^n} \end{aligned} \tag{4.23}$$

The remainder or truncation error for an N-term Taylor series is

$$R_N = \frac{\Delta x^{N+1}}{(N+1)!} \cdot \frac{d^{N+1} f}{dx^{N+1}}(x + \varepsilon \cdot \Delta x) \tag{4.24}$$

where $0 \le \varepsilon \le 1$. This means that the Taylor series is accurate to the order of magnitude of Δx^{N+1}, written as $O(\Delta x^{N+1})$.

Figure 4.4 shows the approximation of

$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2 \tag{4.25}$$

at $x = 0$ using a zero-order, first-order, and second-order Taylor series expansion. Clearly, the higher the order, the better the Taylor series approximates the given function. There is a tradeoff between accuracy and computation time, which becomes an issue when developing numerical approximations, such as numerical integration, discussed later in Section 4.6.

Let's say you are currently at a known location when $x = 0$. You wish to estimate where you will end up at $x + \Delta x$. If you only know the current location, the best you can do is use a zeroth-order approximation and remain at the current location. However, if you also know the first derivative, you can use a first-order approximation and get a better prediction. Continuing this line of reasoning, if you also know the second derivative, you can use a second-order approximation and improve your estimate further. Better and better estimates can be obtained at the expense of greater computation time.

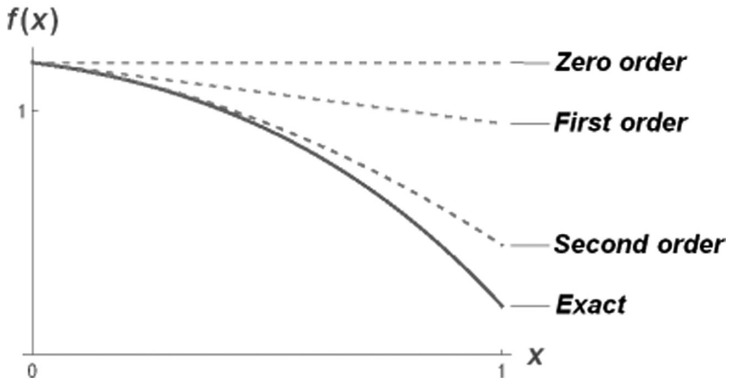


Figure 4.4 Approximations using zero-order, first-order, and second-order Taylor series.

4.2.2 First Derivatives Using Taylor Series

First derivatives can be derived using Taylor series for forward, backward, central differences, and three-point schemes as follows.

The *forward difference* scheme can be obtained from a Taylor series as

$$f(x+\Delta x) = f(x) + \Delta x \cdot \frac{df}{dx} + O(\Delta x^2)$$

$$\frac{df}{dx} = \frac{f(x+\Delta x) - f(x)}{\Delta x} + O(\Delta x) \tag{4.26}$$

Similarly, the *backwards difference* scheme is

$$f(x-\Delta x) = f(x) - \Delta x \cdot \frac{df}{dx} + O(\Delta x^2)$$

$$\frac{df}{dx} = \frac{f(x) - f(x-\Delta x)}{\Delta x} + O(\Delta x) \tag{4.27}$$

The *central difference* scheme is obtained using both a forward and a backward Taylor series up to the second derivative as follows.

$$f(x+\Delta x) = f(x) + \Delta x \cdot \frac{df}{dx} + \frac{\Delta x^2}{2!} \cdot \frac{d^2 f}{dx^2} + O(\Delta x^3)$$

$$f(x-\Delta x) = f(x) - \Delta x \cdot \frac{df}{dx} + \frac{\Delta x^2}{2!} \cdot \frac{d^2 f}{dx^2} + O(\Delta x^3) \tag{4.28}$$

Subtract these formulas and solve for *df*/*dx* to get

$$\frac{df}{dx} = \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x} + O(\Delta x^2) \tag{4.29}$$

Similarly, using algebraic combinations of Taylor series with additional points, other higher-order finite difference approximations can be derived. The three-point formulas are examples.

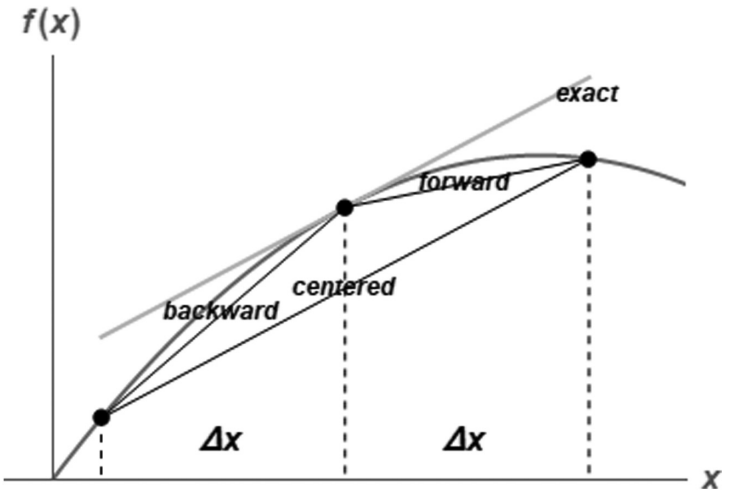


Figure 4.5 Graphical depiction of forward, backward, and centered finite difference approximations of a first derivative.

The three-point forward difference formula is

$$\frac{df}{dx} = \frac{-3f(x) + 4f(x+\Delta x) - f(x+2\Delta x)}{2\Delta x} + O\left(\Delta x^2\right) \tag{4.30}$$

Similarly, the three-point backward difference formula is

$$\frac{df}{dx} = \frac{f(x-2\Delta x) - 4f(x-\Delta x) + 3f(x)}{2\Delta x} + O\left(\Delta x^2\right) \tag{4.31}$$

Backward and forward differences have truncation error $O(\Delta x)$. Central differences and the three-point formulas have a truncation error $O(\Delta x^2)$. Figure 4.5 shows a graphical depiction of the forward, backward, and centered finite difference approximations of the first derivative. Usually, the centered difference gives greater accuracy.

4.2.3 Second Derivatives Using Taylor Series

Write a forward and backward Taylor series up to the third derivative:

$$\begin{aligned} f(x+\Delta x) &= f(x) + \Delta x \cdot \frac{df}{dx} + \frac{\Delta x^2}{2!} \cdot \frac{d^2f}{dx^2} + \frac{\Delta x^3}{3!} \cdot \frac{d^3f}{dx^3} + O\left(\Delta x^4\right) \\ f(x-\Delta x) &= f(x) - \Delta x \cdot \frac{df}{dx} + \frac{\Delta x^2}{2!} \cdot \frac{d^2f}{dx^2} - \frac{\Delta x^3}{3!} \cdot \frac{d^3f}{dx^3} + O\left(\Delta x^4\right) \end{aligned} \tag{4.32}$$

Add these and solve for d^2f/dx^2 to get

$$\frac{d^2f}{dx^2} = \frac{f(x-\Delta x) - 2f(x) + f(x+\Delta x)}{\Delta x^2} + O\left(\Delta x^2\right) \tag{4.33}$$

Second derivatives are naturally centered and have truncation error $O(\Delta x^2)$.

An alternate derivation uses the fundamental definition of a derivative to deduce

$$\frac{d^2f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right) = \frac{1}{\Delta x}\left(\left(\frac{df}{dx}\right)_{x+\frac{\Delta x}{2}} - \left(\frac{df}{dx}\right)_{x-\frac{\Delta x}{2}}\right)$$
$$= \frac{1}{\Delta x}\left(\frac{f(x+\Delta x)-f(x)}{\Delta x} - \frac{f(x)-f(x-\Delta x)}{\Delta x}\right) \tag{4.34}$$
$$= \frac{f(x-\Delta x)-2f(x)+f(x+\Delta x)}{\Delta x^2}$$

This is the same as Equation 4.33 derived using a Taylor series.

4.3 INTEGRALS

4.3.1 Basic Concept of an Integral

In Section 4.1, the concept of rate of change of a function, referred to as the *derivative*, was discussed. The derivative was logically derived as the limit of a finite difference. Now, suppose we know the derivative of a function, $df/dt = g(t)$, but want to determine the function itself. If the function $g(t)$ is known, how do we determine the function $f(t)$? We could refer to this process as finding the *anti-derivative*. The situation is depicted schematically in Figure 4.6.

In order to resolve this question, the derivative is approximated using a finite difference over the interval from a starting time t_0 to a final time t. The finite difference approximation for a derivative is

$$g(t) = \frac{df(t)}{dt} \cong \frac{f(t)-f(t_0)}{\Delta t}, \quad \Delta t = t - t_0 \tag{4.35}$$

If the function at the starting location $f(t_0)$ is known, the function at some time t is approximated as

$$f(t) - f(t_0) = g(t)\Delta t \tag{4.36}$$

Equation 4.35 uses a backward difference. A forward or centered difference could also be used. If the step size Δt is relatively small, this approximation will be accurate. However, for sufficient accuracy, it is usually necessary to advance the approximation in a series of relatively small steps, as shown in Figure 4.7.

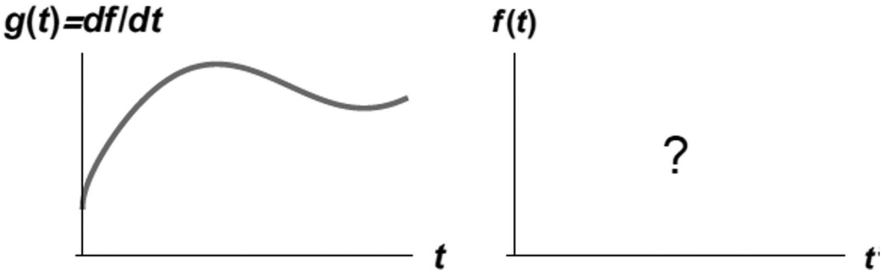


Figure 4.6 The objective of integration: determine the anti-derivative $f(t)$.

The solution is advanced from t_0 to $t = t_n$ in n discrete steps, each of duration $\Delta t = \dfrac{t_n - t_0}{n}$. The function $g(t)$ is evaluated at the corresponding discrete points $t_i = t_0 + i\Delta t$. The finite difference approximation applied over each of the n intervals between t_0 and t produces the following sequence:

$$\begin{aligned} f(t_1) - f(t_0) &\cong g(t_1)\Delta t \\ f(t_2) - f(t_1) &\cong g(t_2)\Delta t \\ &\vdots \\ f(t_i) - f(t_{i-1}) &\cong g(t_i)\Delta t \\ &\vdots \\ f(t_n) - f(t_{n-1}) &\cong g(t_n)\Delta t \end{aligned} \tag{4.37}$$

When these equations are added together, a telescoping cancellation of terms occurs, resulting in

$$f(t) - f(t_0) \cong \sum_{i=1}^{n} g(t_i)\Delta t \tag{4.38}$$

This is known as a Riemann sum. If Δt is sufficiently small, the summation will produce an accurate value for $f(t) - f(t_0)$. In the limit as Δt approaches zero, the summation involves an infinite number of terms, since n approaches infinity. However, the estimate for $f(t) - f(t_0)$ becomes perfect. This limiting process is referred to as the *integral* and is represented with the symbol $\int$ as follows:

$$f(t) - f(t_0) = \lim_{\Delta t \to 0}\left(\sum_{i=1}^{n} g(t_i)\Delta t\right) = \int_{t^* = t_0}^{t} g(t^*)\,dt^* \tag{4.39}$$

Note that the variable t^* is a dummy variable and could be replaced by any other symbol except those already used in the expression. In other words,

$$\int_{t^* = t_0}^{t} g(t^*)\,dt^* = \int_{\text{symbol}=t_0}^{t} g(\text{symbol})\,d(\text{symbol})$$

4.3.2 Geometric Interpretation of an Integral: Area Under a Curve

Referring to the image shown in Figure 4.7, each discrete interval represents a rectangle of area $g(t_i)\Delta t$. An integral is the sum of all these rectangles and thus can be geometrically envisioned as the *area under a curve*. Consider the graph of the function $g(t^*)$ in Figure 4.8. The shaded area under the curve is the *integral* between t_0 and t.

Note that we have deduced that the concept of an integral (or anti-derivative) can be *interpreted* as the area under a curve. We did not start with this notion of area under the curve as the definition of an integral.

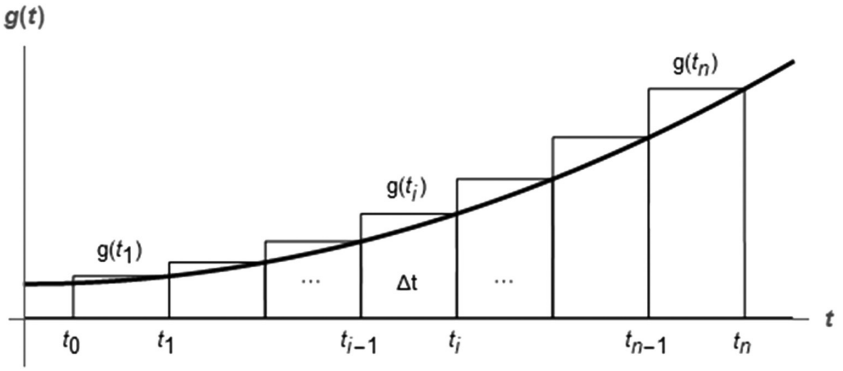


Figure 4.7 Integration in discrete steps.

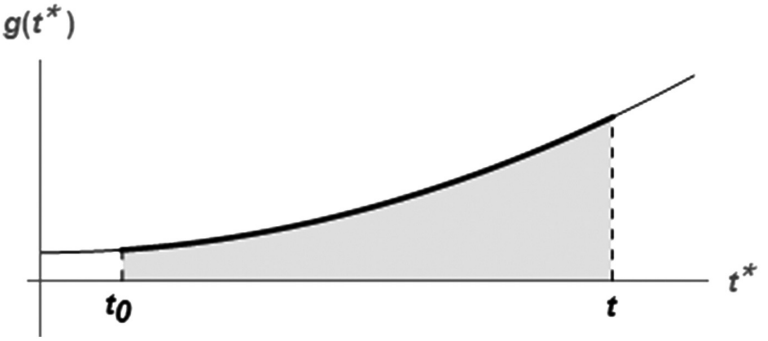


Figure 4.8 Geometric interpretation of an integral as the area under a curve.

Let's return to the foundational application of calculus—displacement and velocity. If velocity $dx/dt = v(t)$ is specified, we can calculate displacement $x(t)$ from a known starting position using integration. If our known starting position is x_0 at time t_0, we get

$$x(t) - x(t_0) = \int_{t^*=t_0}^{t} v(t^*)\,dt^* \cong \sum_{i=1}^{n} v(t_i)\Delta t \tag{4.40}$$

If the velocity is known only at discrete times, the second form in Equation 4.40 is used to approximate position.

4.3.3 Mean Value Theorem

The mean value theorem is stated as

$$\int_a^b g(t)\cdot dt = (b-a)\,g(\xi) \quad \text{where } a < \xi < b \tag{4.41}$$

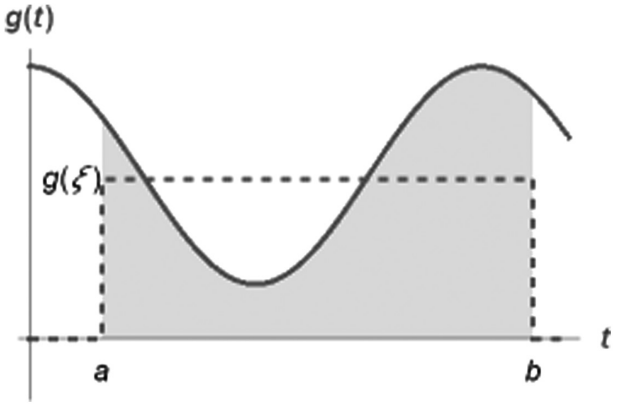


Figure 4.9 Geometric interpretation of the mean value theorem.

The geometric interpretation of the mean value theorem is depicted in Figure 4.9. If $g(\xi)$ is the mean value of g in the interval $a < t < b$, then the area of the rectangle $(b-a)g(\xi)$ exactly equals the value of the integral.

4.3.4 Integration by Parts

Using the product rule for derivatives given by Equation 4.19, the differential of the product of two functions can be expanded and rearranged as

$$d(f \cdot g) = f \cdot dg + g \cdot df \tag{4.42}$$

$$f \cdot dg = d(f \cdot g) - g \cdot df$$

Integrate both sides:

$$\int_a^b f \cdot dg = \int_a^b d(f \cdot g) - \int_a^b g \cdot df \tag{4.43}$$

The first term on the RHS can be directly evaluated to get

$$\int_a^b f \cdot dg = f(b)g(b) - f(a)g(a) - \int_a^b g \cdot df \tag{4.44}$$

This process is referred to as *integration by parts* and is useful if $\int_a^b g \cdot df$ is easier to evaluate than $\int_a^b f \cdot dg$.

4.3.5 Leibniz Rule: Derivatives of Integrals

Consider a function of t defined in terms of an integral in the form

$$I(t) = \int_{x=\alpha(t)}^{\beta(t)} f(x,t)\,dx \tag{4.45}$$

The independent variable t occurs in the integrand, $f(x,t)$, as well as in the limits of integration, $\alpha(t)$ and $\beta(t)$. The derivative of this function is

$$\frac{dI(t)}{dt} = \frac{d}{dt}\left(\int_{x=\alpha(t)}^{\beta(t)} f(x,t)\,dx \right) \tag{4.46}$$

There are many applications for a mathematical form of this type. Examples include phase change problems and boundary layer analysis in fluid mechanics. So, how do we evaluate such a function? To answer this question, the basic notion of a derivative is used in a visual manner.

We begin with the following simpler case with constant limits of integration a and b, so that the t dependence is only found in the integrand. Equation 4.46 simplifies to

$$\begin{aligned} \frac{dI(t)}{dt} &= \frac{d}{dt}\left(\int_{x=a}^{b} f(x,t)\,dx \right) \\ &\cong \frac{1}{\Delta t}\left(\int_{x=a}^{b} f(x,t+\Delta t)\,dx - \int_{x=a}^{b} f(x,t)\,dx \right) \\ &= \int_{x=a}^{b} \frac{f(x,t+\Delta t) - f(x,t)}{\Delta t}\,dx \end{aligned} \tag{4.47}$$

The derivative has been approximated with a finite difference. The difference of the two integrals in this approximation is the shaded area in Figure 4.10.

By taking the limit as $\Delta t \to 0$, the derivative in question becomes

$$\frac{dI(t)}{dt} = \lim_{\Delta t \to 0} \int_{x=a}^{b} \left(\frac{f(x,t+\Delta t) - f(x,t)}{\Delta t} \right) dx = \int_{x=a}^{b} \frac{\partial f(x,t)}{\partial t}\,dx \tag{4.48}$$

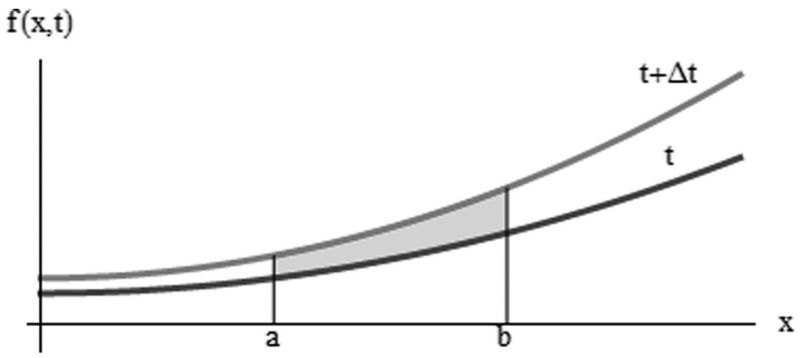


Figure 4.10 The difference of two integrals from Equation 4.47.

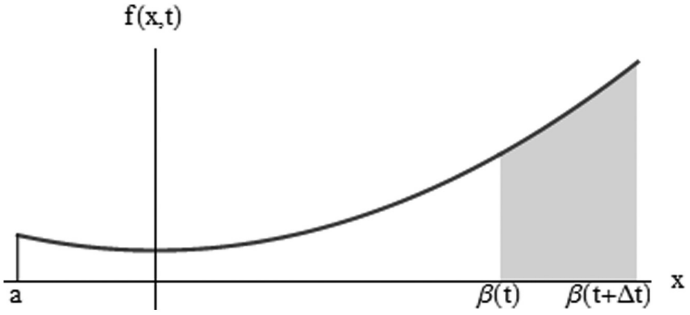


Figure 4.11 The difference of two integrals from Equation 4.49.

Next, we now examine the case where only the upper limit of integration varies with t, while the integrand and the lower limit of integration have no t dependence. Equation 4.46 reduces to

$$\begin{aligned}\frac{dI(t)}{dt} &= \frac{d}{dt}\left(\int_{x=a}^{\beta(t)} f(x)\,dx\right) \\ &\cong \frac{1}{\Delta t}\left(\int_{x=a}^{\beta(t+\Delta t)} f(x)\,dx - \int_{x=a}^{\beta(t)} f(x)\,dx\right) = \frac{1}{\Delta t}\int_{x=\beta(t)}^{\beta(t+\Delta t)} f(x)\,dx\end{aligned} \tag{4.49}$$

This integral is shown as the shaded area in Figure 4.11.

By taking the limit as $\Delta t \to 0$ in Equation 4.49, we conclude that

$$\frac{d}{dt}\left(\int_{x=a}^{\beta(t)} f(x)\,dx\right) = \lim_{\Delta t\to 0}\left(f(\beta(t))\frac{\beta(t+\Delta t)-\beta(t)}{\Delta t}\right) = f(\beta(t))\frac{d\beta}{dt} \tag{4.50}$$

A similar conclusion can be derived when the lower limit of integration is a function of time. Substituting Equations 4.48 and 4.50 into Equation 4.46 leads to the general expression known as the Leibniz rule.

$$\begin{aligned}\frac{dI(t)}{dt} &= \frac{d}{dt}\left(\int_{x=\alpha(t)}^{\beta(t)} f(x,t)\,dx\right) \\ &= \int_{x=\alpha(t)}^{\beta(t)} \frac{\partial f(x,t)}{\partial t}\,dx + f(\beta(t),t)\frac{d\beta}{dt} - f(\alpha(t),t)\frac{d\alpha}{dt}\end{aligned} \tag{4.51}$$

Note that the derivation and visualization of this expression was facilitated by thinking in terms of finite difference approximation of derivatives.

4.4 SUMMARY OF DERIVATIVES AND INTEGRALS

Starting with the fundamental concept of rate of change

$$g(t) = \frac{df}{dt} \tag{4.52}$$

the mathematical concepts of derivatives and integrals can be summarized and visualized as shown in Figure 4.12.

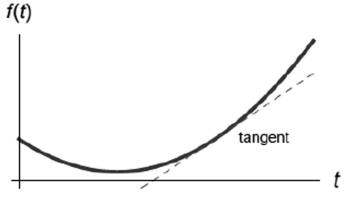
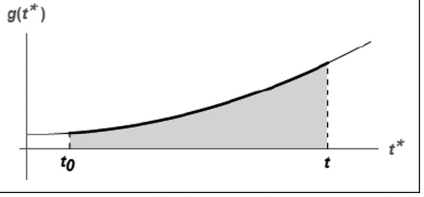
	Derivatives	Integrals
Concept	*Differentiation* is the process of finding $g(t)$ from $f(t)$. $\overbrace{g(t)}^{?} = \frac{d\overbrace{f}^{\surd}}{dt}$	*Integration* is the process of finding the anti-derivative $f(t)$ from $g(t)$. $\overbrace{g(t)}^{\surd} = \frac{d\overbrace{f}^{?}}{dt}$
Numerical Approximation	$\frac{df}{dt} \cong \frac{f(t+\Delta t) - f(t)}{\Delta t}$	$f(t) - f(t_1) \cong \sum_{i=1}^{n} g(t_i)\Delta t$ $t_i = t_1 + i\Delta t$
Mathematical Definition	$g(t) = \frac{df}{dt}$ $= \lim_{\Delta t \to 0}\left(\frac{f(t+\Delta t) - f(t)}{\Delta t}\right)$	$f(t) - f(t_0) = \int_{t^*=t_0}^{t} g(t^*)dt^*$ $= \lim_{\Delta t \to 0}\left(\sum_{i=1}^{n} g(t_i)\Delta t\right)$
Geometric Interpretation	$\frac{df}{dt}$ is the *slope* of the $f(t)$ curve. 	$\int_{t^*=t_0}^{t} g(t^*)dt^*$ is the *area* under the $g(t)$ curve. 

Figure 4.12 A summary of calculus.

4.5 THE STEP, PULSE, AND DELTA FUNCTIONS

4.5.1 The Step Function

An important function for modeling forcing functions in physical problems is the *unit step function*, shown graphically in Figure 4.13.

The step function is defined by the following logical expression:

$$H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases} \tag{4.53}$$

The step function is zero when its argument is negative and one when its argument is positive or zero. The step function models phenomena that suddenly change in magnitude, such as a light switch that suddenly turns on.

4.5.2 The Unit Pulse Function

Another important forcing function for modeling inputs that are pulsed in time or localized in space is the unit pulse function, shown in Figure 4.14. It has a magnitude of $1/\Delta t$ over an interval from 0 to Δt and is zero everywhere else. The mathematical expression of this function is

$$I(t, \Delta t) = \frac{1}{\Delta t}\left(H(t) - H(t - \Delta t)\right) = \begin{cases} 0, & t < 0 \\ 1/\Delta t, & 0 \leq t < \Delta t \\ 0, & t \geq \Delta t \end{cases} \tag{4.54}$$

The pulse function can be expressed as a difference in step functions or equivalently, as a multiple-part function. The representation in terms of a difference in step functions is shown graphically in Figure 4.15.

The physical response to a pulsed forcing function usually involves various combinations of the integral of the pulse function. The integral of the pulse function, activated at t_o with a duration of Δt, is

$$\theta(t) = \int_{t^*=0}^{t} I(t^* - t_o, \Delta t)\, dt^* = \begin{cases} 0, & t < t_o \\ \dfrac{t - t_o}{\Delta t}, & t_o \leq t < t_o + \Delta t \\ 1, & t \geq t_o + \Delta t \end{cases} \tag{4.55}$$

$$= \frac{t - t_o}{\Delta t}\left(H(t - t_o) - H(t - t_o - \Delta t)\right) + H(t - t_o - \Delta t)$$

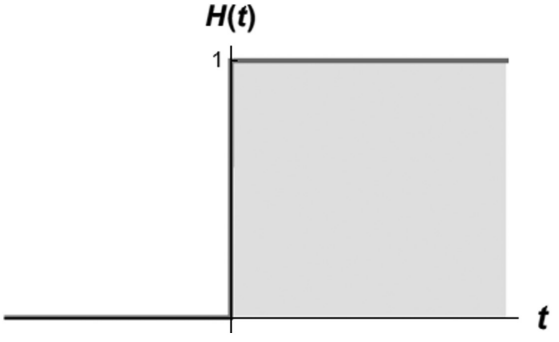


Figure 4.13 The unit step function.

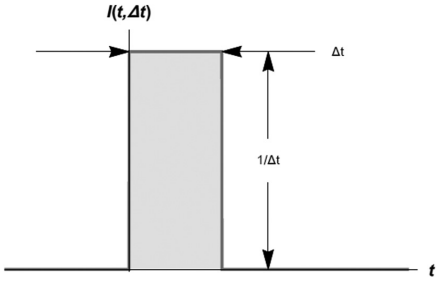


Figure 4.14 The unit pulse function.

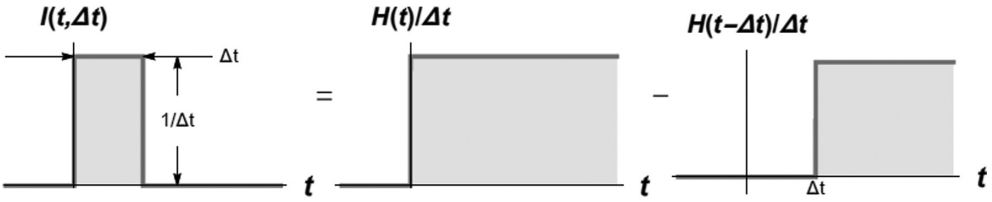


Figure 4.15 Unit pulse function represented as a difference in step functions.

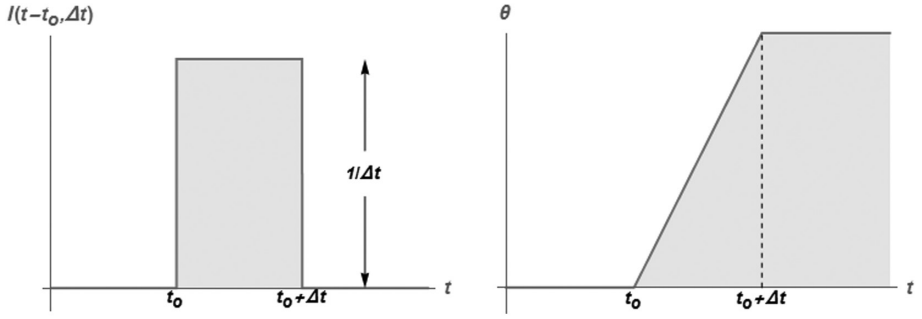


Figure 4.16 The unit pulse function and the integral of the impulse.

The integrand I and integral θ are displayed in Figure 4.16.

Note that t^* is the dummy variable of integration, while t_o is some particular value. This function can be expressed as a multiple-part function or equivalently, in terms of step functions. The integral can be evaluated by considering the upper limit of integration in each of the three possible regions.

Another important integral is that of the impulse function times a smoothly varying function:

$$\int_{t^*=0}^{t} f\left(t^*\right)\cdot I\left(t^*-t_o,\Delta t\right)dt^* \cong f\left(t_o\right)\int_{t^*=0}^{t} I\left(t^*-t_o,\Delta t\right)dt^* = f\left(t_o\right)\cdot\theta\left(t\right) \tag{4.56}$$

This approximation is justified if Δt is relatively small, such that $f(t) \cong f(t_o)$ over the entire pulse, as is the case in Figure 4.17.

4.5.3 The Delta Function

The delta function is the limit of the pulse function as $\Delta t \to 0$. It is written in the form

$$\delta(t) = \lim_{\Delta t \to 0} \left(I(t, \Delta t) \right) = I(t, 0) \tag{4.57}$$

As such, the delta is not a well-behaved function in the normal sense—it can only be defined as the limit of a well-behaved function. However, it is useful in modeling forcing functions or physical inputs that are very short in duration compared with the system's time constant.

Since the delta function has zero thickness and infinite height, it is impossible to plot. However, the graph is often represented as shown in Figure 4.18.

Taking the limit of the pulse function produces the following important properties peculiar to the delta function:

$$\int_{t^*=0}^{t} \delta\left(t^* - t_o\right) dt^* = \lim_{\Delta t \to 0} \left(\int_{t^*=0}^{t} I\left(t^* - t_o, \Delta t\right) dt^* \right) = H\left(t - t_o\right) \tag{4.58}$$

$$\int_{t^*=0}^{t} f\left(t^*\right) \cdot \delta\left(t^* - t_o\right) dt^* = \lim_{\Delta t \to 0} \left(\int_{t^*=0}^{t} f\left(t^*\right) \cdot I\left(t^* - t_o, \Delta t\right) dt^* \right) = f\left(t_o\right) \cdot H\left(t - t_o\right) \tag{4.59}$$

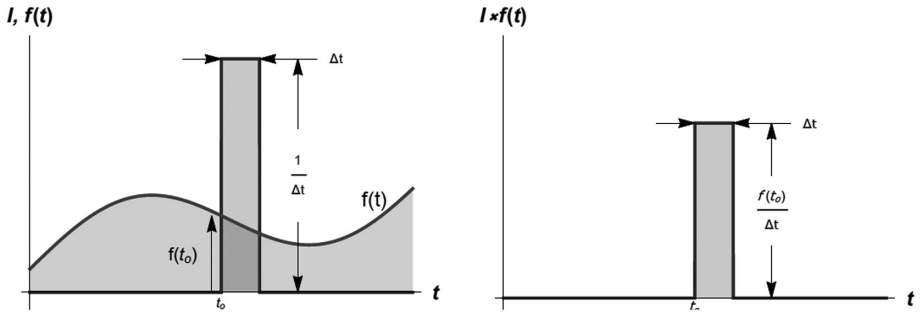


Figure 4.17 The impulse function for relatively small Δt such that $f(t) \cong f(t_o)$.

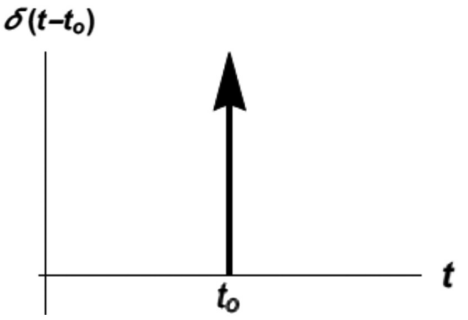


Figure 4.18 The delta function.

Thus, the integral of the delta function sifts particular values of $f(t)$.

Over any interval of integration, Equations 4.58 and 4.59 can be generalized as

$$\int_{x^*=a}^{b} \delta\left(x^* - x_o\right) dx^* = \begin{cases} 0, & x_o < a \\ 1, & a \le x_o \le b \\ 0, & x_o > b \end{cases} \tag{4.60}$$

$$\int_{x^*=a}^{b} f\left(x^*\right)\delta\left(x^* - x_o\right) dx^* = \begin{cases} 0, & x_o < a \\ f\left(x_o\right), & a \le x_o \le b \\ 0, & x_o > b \end{cases} \tag{4.61}$$

Thus, it is shown that when integrated, the delta function sifts out particular values of a function. All of this can be extremely difficult to understand unless one thinks in terms of the limit of the unit pulse function, which is completely understandable and straightforward.

4.6 NUMERICAL INTEGRATION

Numerical integration is the process of approximating the integral:

$$I = \int_{x=a}^{b} f\left(x\right) dx \tag{4.62}$$

The integral can be visualized as the area under the curve. This area can be approximated using a simple function to represent the integrand. The Newton–Cotes integration rules are obtained by approximating the integrand with a polynomial that interpolates $f(x)$ at equally spaced points. Several possibilities are shown in Figure 4.19. The higher-order polynomials follow the curvature of the integrand more closely and give a more accurate estimate of the integral.

The area estimate can also be improved by dividing the area into multiple segments. For instance, Figure 4.20 shows the approximate integral using three segments with the trapezoid and Simpson's rule. These are called the *composite rules*.

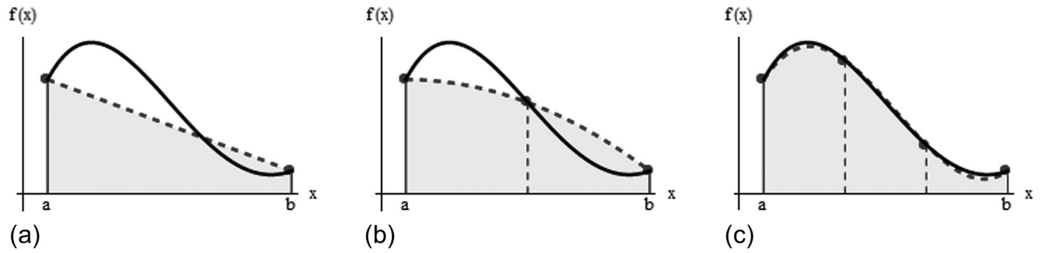


Figure 4.19 (a) The trapezoid rule: linear interpolation, (b) Simpson's rule: quadratic interpolation, (c) Simpson's 3/8 rule: cubic interpolation.

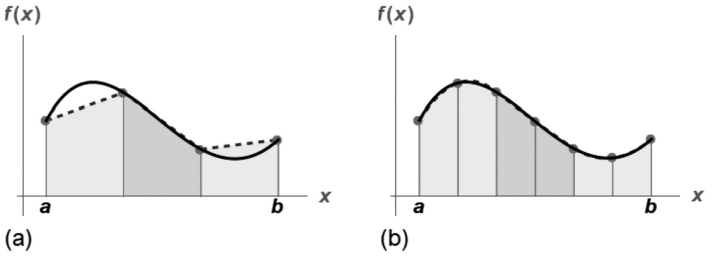


Figure 4.20 (a) Composite trapezoid rule; (b) composite Simpson's rule.

A few selected examples of how integration is used to evaluate areas in engineering and scientific applications are:

a) A surveyor might need to know the area of a field bounded by a meandering stream and two roads.
b) A hydrologist might need to know the cross-sectional area of a river.
c) A structural engineer might need to determine the net force due to a nonuniform wind blowing against the side of a skyscraper.

4.6.1 Trapezoid Rule

The *trapezoid rule* uses a linear interpolation to approximate $f(x)$. An integral using the trapezoid rule approximation is shown in Figure 4.21.

Replacing the function $f(x)$ by a linear approximation gives

$$I = \int_a^b f(x)\,dx \cong \int_a^b \left(f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \right) dx \tag{4.63}$$

Although the exact integral may be difficult to evaluate, the trapezoidal approximation is easy to evaluate. The result of carrying out the integration in Equation 4.63 is the trapezoid rule:

$$I \cong (b - a)\left(\frac{f(a) + f(b)}{2} \right) \tag{4.64}$$

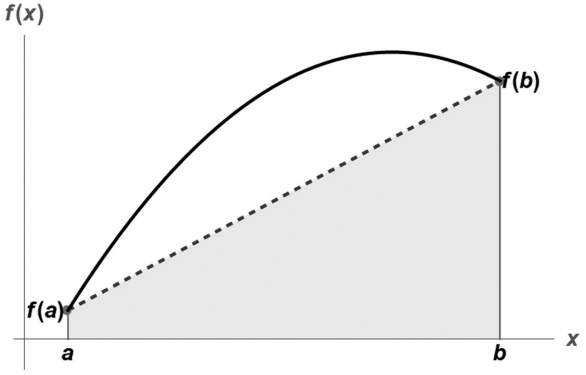


Figure 4.21 The trapezoid approximation of an integral.

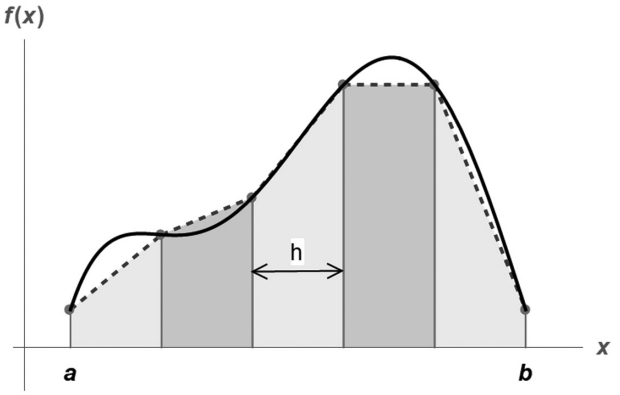


Figure 4.22 The composite trapezoid rule.

Geometrically, this formula is interpreted as

$$I = \text{width} \times \text{average height} \tag{4.65}$$

In fact, due to the mean value concept of basic calculus, all numerical approximations to an integral can be interpreted as *width* × *average height*.

One way to improve the accuracy of the trapezoid rule is to divide the interval into a number of segments and apply the method to each segment. This is known as the *composite trapezoid rule*. The idea is shown in Figure 4.22.

If the entire region is divided into n segments, the width of each segment is $h = (b-a)/n$. The total integral is

$$\begin{aligned} I &= \int_{x=a}^{b} f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,dx \\ &\cong h\frac{f(x_0)+f(x_1)}{2} + h\frac{f(x_1)+f(x_2)}{2} + \cdots + h\frac{f(x_{n-1})+f(x_n)}{2} \end{aligned} \tag{4.66}$$

where $x_i = a + i \cdot h,\ \ i = 0,\ldots n$. Combining terms in Equation 4.66 produces the *composite trapezoid rule*:

$$I \cong \frac{h}{2}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) \tag{4.67}$$

This formula can also be interpreted as $I = width \times average$ height by writing Equation 4.67 as

$$I = \underbrace{(b-a)}_{\text{width}} \underbrace{\frac{1}{2n}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n) \right)}_{\text{average height}} \tag{4.68}$$

We now turn to the accuracy of the trapezoid rule. When we approximate the area under a curve as the area under a simple straight line segment, there is some penalty in the form of error. Using Taylor series, an estimate for the error using a single trapezoid is

$$E = -\frac{1}{12}(b-a)^3 f''(\xi) \tag{4.69}$$

where $f''(\xi)$ is the second derivative and ξ lies somewhere in the interval a to b. Apply this error estimate to the composite trapezoid rule to find

$$E = -\frac{1}{12}h^3 \sum_{i=1}^{n} f''(\xi_i) \tag{4.70}$$

The mean value theorem of calculus implies that

$$\overline{f''} = \frac{1}{n}\sum_{i=1}^{n} f''(\xi_i) \tag{4.71}$$

$$\sum_{i=1}^{n} f''(\xi_i) = n\overline{f''} = \frac{b-a}{h}\overline{f''}$$

Thus,

$$E = -\frac{1}{12}(b-a)h^2\overline{f''} \tag{4.72}$$

The conclusion is that the error is of the order h^2. The implication is that if we cut the mesh size by 2, we should expect the error to decrease by a factor of about 4.

Example

As an example, consider the integral

$$I = \int_{x=a}^{b} f(x)dx = \int_{x=a}^{b} xe^{-x}dx \tag{4.73}$$

In this case, the exact integral can be determined as

$$I_{\text{exact}} = \left(-xe^{-x} - e^{-x}\right)_{x=a}^{x=b} = \left(-be^{-b} - e^{-b}\right) - \left(-ae^{-a} - e^{-a}\right) \tag{4.74}$$

Choosing $a = 0$ and $b = 4$ gives $I_{\text{exact}} = 0.9084$.

Numerical approximations using one, two, four, and eight trapezoids are shown graphically in Figure 4.23. It can be seen that a single trapezoid significantly underestimates the integral in this case. Using more trapezoids gives continually better estimates. One can always obtain greater accuracy at the expense of greater computational effort. Also, the error decreases by a factor of approximately h^2, as anticipated from the error analysis given by Equation 4.72.

4.6.2 Trapezoid Rule for Unequal Segments

When integrating a function for which we have a formula $y = f(x)$, we can use the formula to determine $f(x)$ at any x we wish. We can thus perform trapezoidal rule integration with any

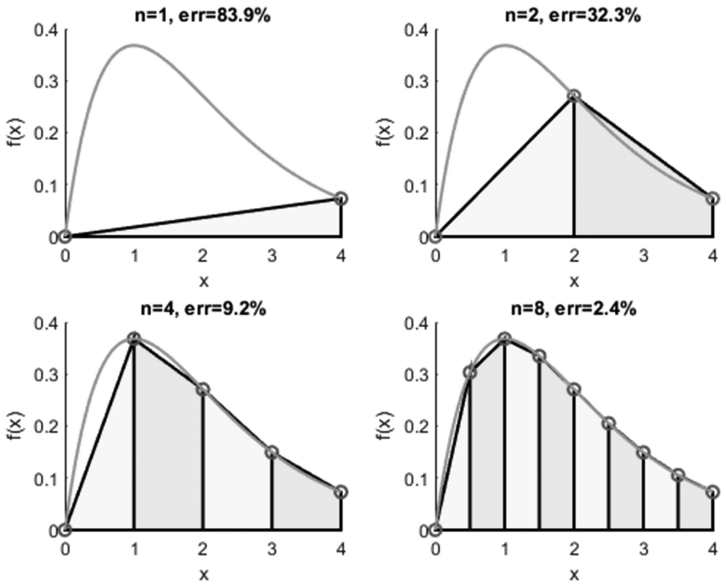


Figure 4.23 The trapezoid rule for one, two, four, and eight trapezoids.

step size. If we want to integrate discrete data, however, the values of x for which we have values of f are determined by the experimental procedure. Furthermore, these values of x may not even be spaced evenly, as is the case for the data shown in Figure 4.24.

The trapezoidal rule can still be used in this case. The integral is still a sum of trapezoid areas, except now the trapezoids are not all of the same width. For $n+1$ pairs of x-f data points, the numerical approximation of the integral is obtained by simply applying the trapezoid rule over each segment and adding the results:

$$I = h_1 \frac{f(x_0)+f(x_1)}{2} + h_2 \frac{f(x_1)+f(x_2)}{2} + \cdots + h_n \frac{f(x_{n-1})+f(x_n)}{2}$$
$$= \sum_{i=1}^{n} h_i \frac{f(x_{i-1})+f(x_i)}{2} \tag{4.75}$$

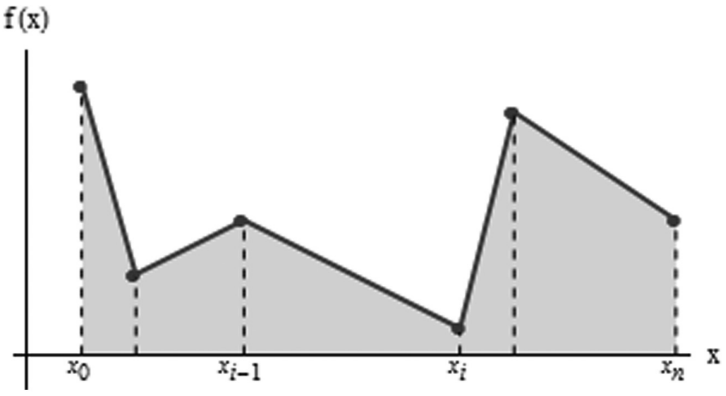


Figure 4.24 The trapezoid rule for uneven segments.

Here $h_i = x_i - x_{i-1}$ is the width of segment i. For constant h_i, this reduces to the normal composite trapezoid rule, Equation 4.67.

4.6.3 Simpson's Rule

Instead of a linear interpolation as used with the trapezoid rule, Simpson's rules use higher-order polynomials. The quadratic and cubic interpolating functions are shown in the following Figure 4.25.

Simpson's rule uses a quadratic polynomial to approximate the integrand. Using a single panel, the integral from $a = x_0$ to $b = x_2$ is approximated using a Lagrange interpolating function:

$$\begin{aligned} I &= \int_{x_0}^{x_2} f(x)\,dx \\ &\cong \int_{x_0}^{x_2} \left(\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) \right. \\ &\qquad \left. + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \right) dx \end{aligned} \tag{4.76}$$

The result of the integration is

$$I \cong \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right) \tag{4.77}$$

where $h = (b-a)/2 = (x_2 - x_0)/2$. This is known as Simpson's 1/3 rule, since h is multiplied by one-third.

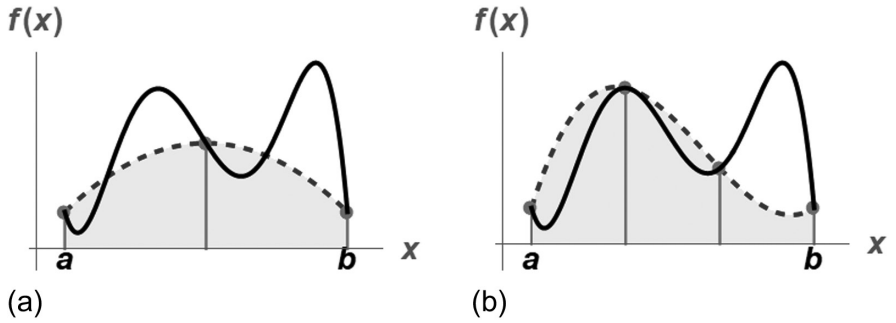


Figure 4.25 (a) Simpson's rule: area under a parabola connecting three points. (b) Simpson's 3/8 rule: area under a cubic equation connecting four points.

Just like the trapezoid rule, Simpson's rule can be improved by dividing the interval into n segments of equal width $h = (b-a)/n$. The total integral is

$$\begin{aligned} I &= \int_{x=a}^{b} f(x)\,dx = \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \cdots + \int_{x_{n-2}}^{x_n} f(x)\,dx \\ &\cong \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right) + \frac{h}{3}\left(f(x_2) + 4f(x_3) + f(x_4)\right) + \ldots \\ &\quad + \frac{h}{3}\left(f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right) \end{aligned} \tag{4.78}$$

Grouping terms produces the composite Simpson's 1/3 rule.

$$I = \frac{h}{3}\left(f(x_0) + 4\sum_{i=1,3,\ldots}^{n-1} f(x_i) + 2\sum_{i=2,4,\ldots}^{n-2} f(x_i) + f(x_n)\right) \tag{4.79}$$

Note that the number of segments n must be a multiple of 2 in order to apply this rule.

An error estimate can be obtained by a procedure similar to that used for the trapezoid rule:

$$\begin{aligned} &\text{Single segment:} \quad E = -\frac{1}{90}h^5 f^{(4)}(\xi), \quad h = \frac{b-a}{2} \\ &\text{Composite } \frac{1}{3} \text{ rule:} \quad E = -\frac{1}{180}(b-a)h^4\overline{f^{(4)}}, \quad h = \frac{b-a}{n} \end{aligned} \tag{4.80}$$

The conclusion is that the error E is $O(h^4)$. This implies that by cutting h in half, the expected error decreases by a factor of 16. Also, the error is proportional to the fourth derivative; thus, the composite Simpson's 1/3 rule gives an exact answer for polynomials of order three or lower. Comparing with the error analysis of the trapezoid rule, we would have expected the error to be proportional to only the third derivative.

A numerical example is shown in Figure 4.26 for the case used previously for the trapezoid rule displayed in Figure 4.23.

$$I = \int_{x=a}^{b} f(x)\,dx = \int_{x=0}^{4} xe^{-x}dx = 0.9084 \tag{4.81}$$

4.6.4 Simpson's 3/8 Rule

In a manner similar to the derivation of Simpson's rule using Equation 4.76, a third-order Lagrange interpolating function can be used to obtain

$$I = \frac{3h}{8}\left(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right) \tag{4.82}$$

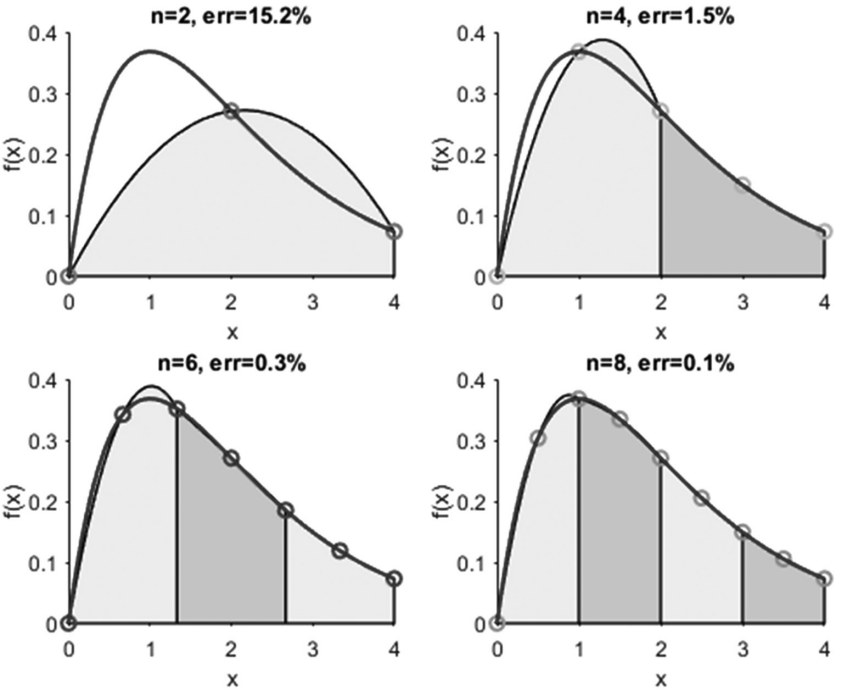


Figure 4.26 Accuracy of Simpson's rule.

$$h = \frac{b-a}{3} = \frac{x_3 - x_0}{3}$$

The error with this approximation is

$$\text{Single segment:}\quad E = -\frac{3}{80}h^5 f^{(4)}(\xi) \tag{4.83}$$

4.6.5 Gauss Quadrature

The Newton–Cotes rules, such as the trapezoid and Simpson's rules, use equally spaced function values. For example, as depicted in the left panel of Figure 4.27, the trapezoid rule uses the area under a straight line and uses the endpoints of the interval—resulting in a rather large error for the case shown. Now, suppose that the constraint of fixed base points is removed, and any points could be chosen to form a straight line, as shown in the right panel of Figure 4.27. By positioning these points wisely, we could get a straight line that could balance the positive and negative errors. This strategy is called *Gauss quadrature*.

First we consider a two-point Gauss–Legendre formula. The idea is to approximate an integral using a formula of the type

$$I = c_0 f(x_0) + c_1 f(x_1) \tag{4.84}$$

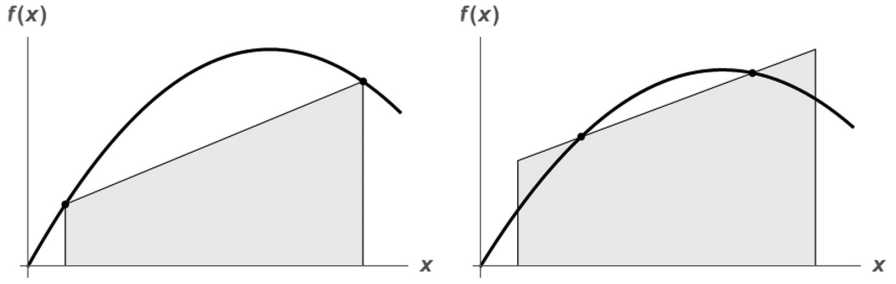


Figure 4.27 Comparison of the trapezoid rule and two-point Gauss quadrature.

where both the coefficients, c_0 and c_1, as well as the function evaluation points, x_0 and x_1, are unknown. After scaling the limits of integration to the range [-1, 1], it can be shown that

$$c_0 = c_1 = 1 \tag{4.85}$$

$$x_0 = -\frac{1}{\sqrt{3}}, \quad x_1 = \frac{1}{\sqrt{3}}$$

A graphical representation of the two-point Gauss–Legendre formula is depicted in Figure 4.28.

In a similar manner, more accurate, higher-order approximations can be derived. For an n-point Gauss–Legendre formula, the numerical approximation for an integral is

$$I = \sum_{i=0}^{n-1} c_n f\left(x_n\right) \tag{4.86}$$

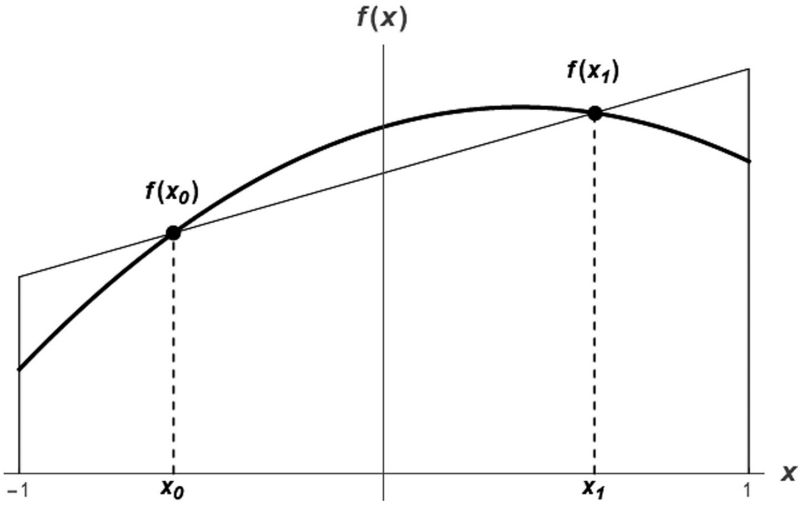


Figure 4.28 The two-point Gauss quadrature.

Points	Weighting Factors	Function Arguments	Truncation Error
1	$c_0 = 2$	$x_0 = 0$	$f^{(2)}(\xi)$
2	$c_0 = 1$ $c_1 = 1$	$x_0 = -1/\sqrt{3}$ $x_1 = 1/\sqrt{3}$	$f^{(4)}(\xi)$
3	$c_0 = 5/9$ $c_1 = 8/9$ $c_2 = 5/9$	$x_0 = -\sqrt{3/5}$ $x_1 = 0$ $x_2 = \sqrt{3/5}$	$f^{(6)}(\xi)$
4	$c_0 = (18 - \sqrt{30})/36$ $c_1 = (18 + \sqrt{30})/36$ $c_2 = (18 + \sqrt{30})/36$ $c_3 = (18 - \sqrt{30})/36$	$x_0 = -\sqrt{525 + 70\sqrt{30}/35}$ $x_1 = -\sqrt{525 - 70\sqrt{30}/35}$ $x_2 = \sqrt{525 - 70\sqrt{30}/35}$ $x_3 = \sqrt{525 + 70\sqrt{30}/35}$	$f^{(8)}(\xi)$
5	$c_0 = (322 - 13\sqrt{70})/900$ $c_1 = (322 + 13\sqrt{70})/900$ $c_2 = 128/225$ $c_3 = (322 + 13\sqrt{70})/900$ $c_4 = (322 - 13\sqrt{70})/900$	$x_0 = -\sqrt{245 + 14\sqrt{70}/21}$ $x_1 = -\sqrt{245 - 14\sqrt{70}/21}$ $x_2 = 0$ $x_3 = \sqrt{245 - 14\sqrt{70}/21}$ $x_4 = \sqrt{245 + 14\sqrt{70}/21}$	$f^{(10)}(\xi)$

Figure 4.29 Weighting factors and function arguments for Gauss–Legendre formulas.

The results using 1 through 5 points are summarized in Figure 4.29. The accuracy or truncation error improves significantly with the higher-order approximations.

4.7 MULTIPLE INTEGRALS

A double integral can be written as

$$I = \int_{y=c}^{d} \left(\int_{x=a}^{b} f(x,y)\,dx \right) dy = \int_{x=a}^{b} \left(\int_{y=c}^{d} f(x,y)\,dy \right) dx \tag{4.87}$$

This integral is visualized as the volume under a two-dimensional surface, as shown in Figure 4.30.

Multiple integrals can be computed numerically by extending the methods for a function of one variable. Methods such as the trapezoid or Simpson's rule can readily be applied. First, a rule is applied in one dimension with each value of the second dimension held constant. Then, the rule is applied in the second dimension to obtain a numerical integration of a double integral.

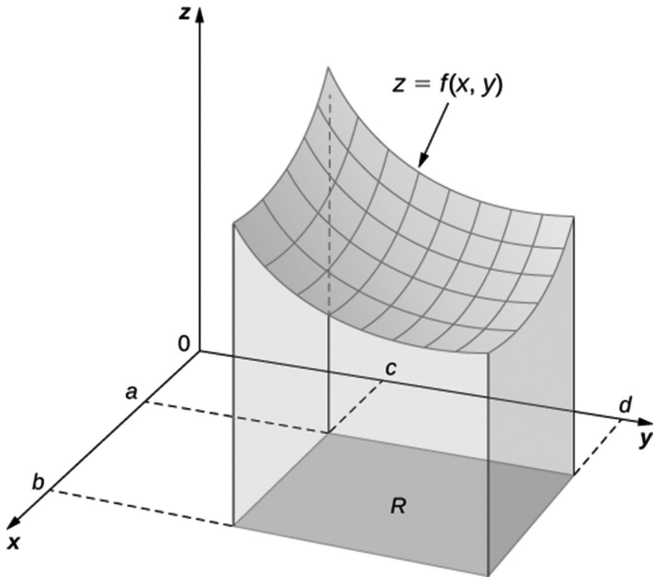


Figure 4.30 A double integral as the volume under a surface.

PROBLEMS

Problem 4.1

Determine and plot the derivative of the step function, $H(t)$.

$$H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases}$$

Problem 4.2

Determine and plot the derivative of the ramp function, $Ramp(t)$.

$$\text{Ramp}(t) = t \cdot H(t) = \begin{cases} 0, & t < 0 \\ t, & t \geq 0 \end{cases}$$

Problem 4.3

Determine and plot the derivative of the pulse function.

$$\text{Pulse}(t) = H(t) - H(t - \Delta t) = \begin{cases} 0, & t < 0 \\ 1, & 0 \leq t \leq \Delta t \\ 0, & t > \Delta t \end{cases}$$

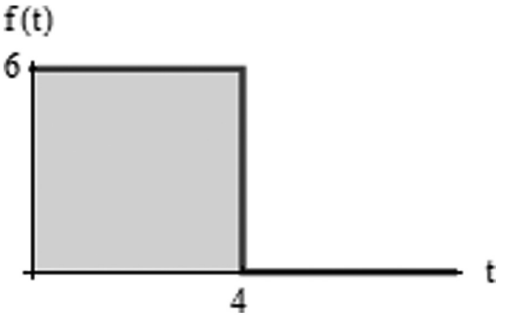
Problem 4.4

Determine and plot the derivative of the triangle function.

$$\text{Triangle}(t) = \begin{cases} 0, & t < -1 \\ t+1, & -1 \le t < 0 \\ 1-t, & 0 \le t < 1 \\ 0, & t \ge 1 \end{cases}$$

Problem 4.5

Consider the function $f(t)$ shown.



Evaluate and sketch the following integrals, $I(t)$.

a) $$I(t) = \int_{t_o=0}^{3} f(t_o)\,dt_o$$

b) $$I(t) = \int_{t_o=0}^{8} f(t_o)\,dt_o$$

c) $$I(t) = \int_{t_o=0}^{t} f(t_o)\,dt_o$$

Problem 4.6

Consider the pulse function $f(t)$ shown.

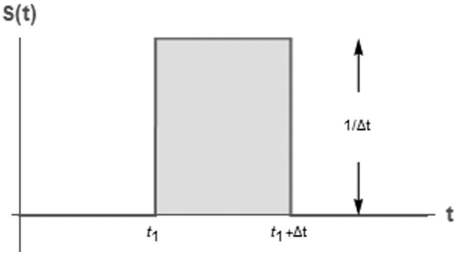
Evaluate and sketch the following integrals. Note that the independent variable t can be less than t_1 or greater than t_1. Since the integrand $f(t)$ is a two-part function, the integral $I(t)$ is best evaluated as a two-part function.

a) $$I(t) = \int_{t_o=0}^{t} f(t_o)\,dt_o$$

b) $$I(t) = \int_{t_o=0}^{t} f(t_o)\,e^{-(t-t_o)/\tau}\,dt_o$$

Problem 4.7: Pulse Integral

Consider the impulse function $S(t)$.



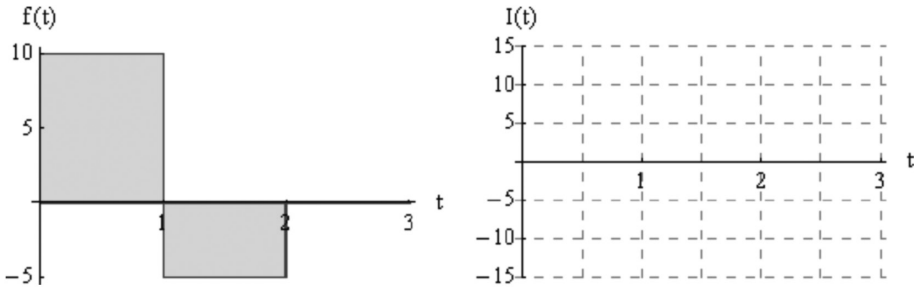
a) Determine the symbolic solution and sketch the function $F_1(t) = \int_{t_o=0}^{t} S(t_o)\,dt_o$. Examine the limit as $\Delta t \to 0$.

b) Determine the symbolic solution and sketch the function $F_2(t) = \int_{t_o=0}^{t} S(t_o)\,e^{-(t-t_o)/\tau}\,dt_o$. Examine the limit as $\Delta t \to 0$.

c) Evaluate the function F_2 from part (b) using numerical integration. Use $t_1 = 2$, $\Delta t = 1$, and $\tau = 1$. Plot both the analytical solution and the numerical solution on the same graph.

Problem 4.8: Pulses

Consider the function $f(t)$ shown.



Determine an expression for the integral $I(t) = \int_{t_o=0}^{t} f(t_o)\,dt_o$ over the entire range $0 \le t \le 3$. Also, sketch $I(t)$.

Problem 4.9

What are the units of the following expressions? In all cases, t is in seconds(s) and x is in meters(m).

a) $\delta(t)$
b) $\delta(x)\delta(t)$
c) $H(t)$ = Step function
d) $Ramp(x)$ = Ramp function

Problem 4.10

Evaluate the following integrals. H is the step function and $\delta(x)$ is the delta function.

a) $\int_{x=0}^{3} H(x-1)\,dx$

b) $\int_{x=2}^{3} H(x-1)\,dx$

c) $\int_{x=0}^{3} \delta(x-2)\,dx$

d) $\int_{x=0}^{3} \delta(x-5)\,dx$

e) $\int_{x=0}^{3} f(x)\delta(x-2)\,dx$

f) $\int_{x=0}^{3} f(x)\delta(x-5)\,dx$

Problem 4.11

Evaluate and sketch the following integrals.

a) $f(x) = \int_{x_o=a}^{b} \delta(x-x_o)\,dx_o$

b) $f(x) = \int_{x_o=a}^{b} H(x-x_o)\,dx_o$

c) $f(x) = \int_{x_o=a}^{b} (x - x_o)\, dx_o$

d) $f_1(t, t_1) = \int_{t_o=0}^{t} \delta(t_o - t_1)\, dt_o$

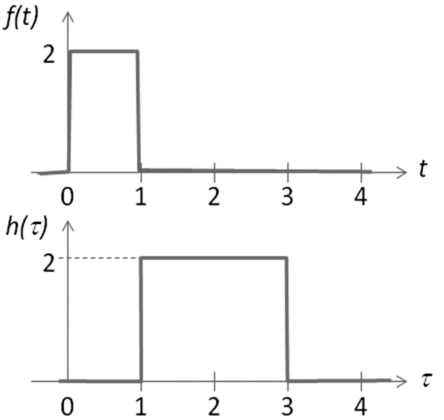
e) $f_2(t, t_1) = \int_{t_o=0}^{t} f_1(t_o, t_1)\, dt_o$, where f_1 is the function determined in part (d).

Problem 4.12

The error function is $erf(z) = \frac{2}{\sqrt{\pi}} \int_{x=0}^{z} e^{-x^2}\, dx$. Determine $\frac{d}{dz} erf(z)$ and $\frac{d}{dz} erf(z^2)$.

Problem 4.13: Convolution Integral

Two functions, $f(t)$ and $h(t)$, are shown in the figure.



A third function $g(t)$ is related to the two functions plotted in the figure using the following integral:

$$g(t) = \int_{0}^{t} h(\tau) f(t - \tau)\, d\tau$$

This integral is the well-known convolution integral, used to determine how a signal $f(t)$ will be filtered by a system's impulse response $h(\tau)$.

a) Determine the function $g(t)$for all values of t.
b) Plot the function $g(t)$.

Hint: This integral is best undertaken by breaking the functions up into sections.

Problem 4.14

Why do we even need numerical integration? Why not simply evaluate all integrals in an exact, symbolic form?

Problem 4.15: Numerical Integration Basic

Consider the integral $I = \int_{x=0}^{4} \left(1 - e^{-2x}\right) dx$. Evaluate this integral using the methods listed.

Rule	*Numerical Approximation*	*% Error*
Exact		0
Trapezoid, $n = 1$		
Trapezoid, $n = 2$		
Trapezoid, $n = 4$		
Simpson's 1/3, $n = 2$		
Simpson's 1/3, $n = 4$		
Simpson's 3/8, $n = 3$		

Problem 4.16: *erf*

a) Use the trapezoid rule to evaluate $erf(1)$ for a sequence of increasing n (decreasing h): $n = 2^p$ for $p = 1$ to 10.
b) Determine the n required to get an absolute error less than 10^{-7}.

Problem 4.17: Distance from Velocity

The following velocity versus time data is available:

t (s)	0	3	4	6	8	12
v (m/s)	0	9	16	36	33	25

Estimate the distance traveled using the trapezoid rule with

a) $n = 1$ (one trapezoid)
b) $n = 2$ (two trapezoids)
c) all the available data

Make sketches of the data and the various approximations.

Problem 4.18: Ellipse

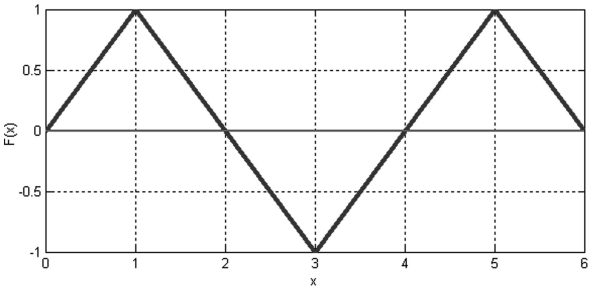
The perimeter P of an ellipse is given by $P = 4a\int_0^{\pi/2} \sqrt{1 - k^2 \sin^2\theta}\, d\theta$
where
a = major axis
b = minor axis
$k = \sqrt{a^2 + b^2}/a$

Write a function that calculates the perimeter of an ellipse. The input arguments should be a and b, and the output argument should be P. Use your function to calculate the perimeter of the following ellipses:

a) $\frac{x^2}{5^2} + \frac{y^2}{2^2} = 1$, b) $\frac{x^2}{4^2} + \frac{y^2}{7^2} = 1$, c) $\frac{x^2}{100^2} + \frac{y^2}{2^2} = 1$

Problem 4.19

Consider the function $F(x)$ shown.

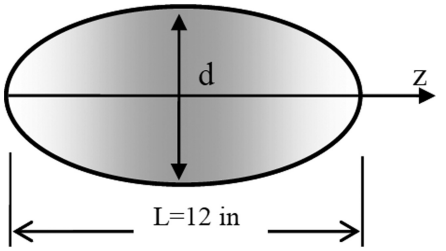


We wish to compute the integral $I(t) = \int_0^6 F(x)\,dx$

a) Compute the exact value of the integral.
b) Approximate the integral with the trapezoid rule using 1, 2, 3, 6, and 10^6 trapezoids.
c) Approximate the integral with the Simpson's 1/3 rule using one and two panels.

Problem 4.20: Football

To estimate the surface area of a football, the diameter of the ball is measured at different points along the ball.



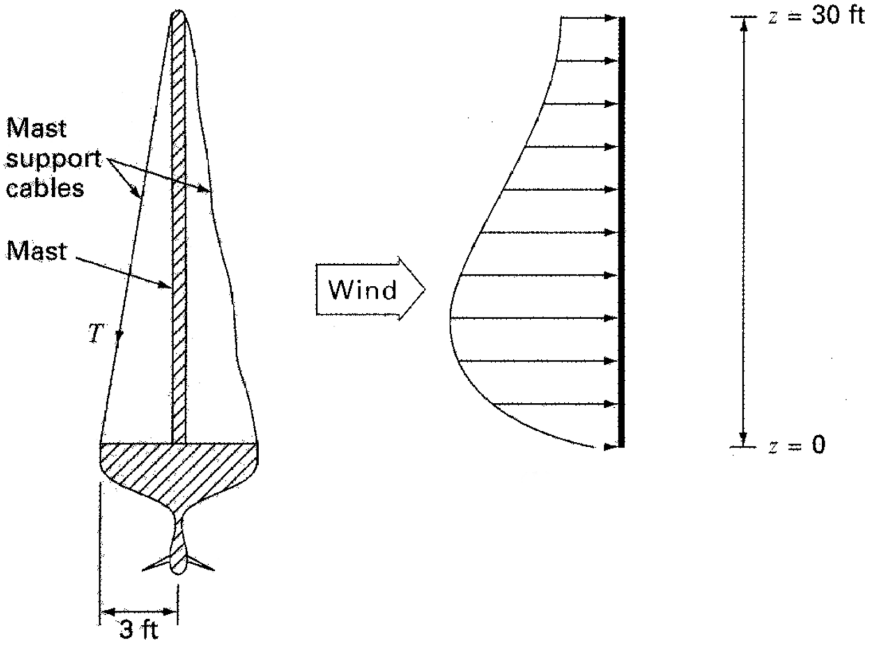
z (in)	0	1.5	3	4.5	6
d (in)	0.0	2.9	4.8	5.8	6.2

The surface area can be determined from $S = \pi \int_0^L d \cdot dz$

a) Use the composite trapezoid rule with the given data to estimate the surface area of the ball.
b) Use the composite Simpson's rule with the given data to estimate the surface area of the ball.

Problem 4.21: Sailboat

A cross section of a racing sailboat is shown.



H is the height of the mast and W is the distance to support cable attachment. Wind forces exerted per length of mast, $f(z)$ (lb/ft), vary as a function of distance above the deck, z (ft). A good approximation of the wind force per length of mast is

$$f(z) = 200\left(\frac{z}{5+z}\right)e^{-\frac{2z}{H}} \text{ (lb/ft)}$$

Assume that the right support cable is completely slack and the mast joins the deck in a manner that transmits horizontal and vertical forces but no moments. The total force, F (lb), exerted by the wind on the mast is

$$F = \int_0^H f(z) \cdot dz$$

Write a function that computes the total force, F (lb), and the tension in the left mast support cable, T (lb), as a function of H (ft) and W (ft). To compute T, sum the moments about the base of the mast. The moment caused by the variable wind force is $M = \int_0^H z \cdot f(z) \cdot dz$. Next, write a separate file to explore the system parameters. For $H = 30$ ft and $W = 3$ ft, compute F and T. Also, plot F and T versus H in the range 10 ft $< H <$ 40 ft while keeping $W = 3$ ft.

Problem 4.22: Force on a Dam

Water exerts pressure on the upstream face of a dam. The pressure increases linearly with depth and can be characterized by $p(z) = \rho g(D - z)$

where

- $p(z)$ is pressure in N/m^2 exerted at an elevation z meters above the reservoir bottom
- ρ is density of water, which for this problem is assumed to be a constant 10^3 kg/m^3
- g is acceleration due to gravity (9.81 m/s^2)
- D is elevation (in m) of the water surface above the reservoir bottom

Omitting atmospheric pressure (because it works against both sides of the dam face and essentially cancels out), the total force f_t can be determined by multiplying pressure times the area of the dam face. Because both pressure and area vary with elevation, the total force is obtained by evaluating

$$f_t = \int_0^D \rho g \cdot (D - z) w(z) \cdot dz$$

where $w(z)$ is width of the dam face (m) at elevation z. The line of action can also be obtained by evaluating

$$d = \frac{\int_0^D z \rho g \cdot (D - z) w(z) \cdot dz}{\int_0^D \rho g \cdot (D - z) w(z) \cdot dz}$$

The following table contains data of stream width at various elevations.

Elevation z (m)	0	10	20	30	40	50	60
Width *w* (m)	122	130	135	160	175	190	200

Write a function that accepts vectors containing the z locations and the corresponding w values on the dam face. The function should perform numerical integration using the trapezoid rule for discrete data and return the values f_t and d. Test your function using the specific data given.

Problem 4.23: Wind Force

A wind force distributed against the side of a skyscraper is measured as

Height z (m)	0	30	60	90	120	150	180	210	240
Force *F* (N/m)	0	340	1200	1600	2700	3100	3200	3500	3800

Compute the net force and line of action due to this distributed wind. The total force F exerted on the mast is

$$F_t = \int_0^H F(z) dz$$

The line of action can also be determined by integration:

$$d = \frac{\int_0^H z F(z) dz}{\int_0^H F(z) dz}$$

Problem 4.24: Blackbody Radiation Function

The thermal radiation emitted from an object is a function of its absolute temperature. A blackbody emitter is an ideal surface that emits radiation uniformly in all directions and absorbs all radiation that is incident on its surface. The variation of the emissive power, $E_{\lambda,b}$, as a function of wavelength for a blackbody emitter is described by the Planck distribution

$$E_{\lambda,b} = \frac{c_1}{\lambda^5\left(\exp\left(\frac{c_2}{\lambda T}\right)-1\right)}\left(\frac{\mathrm{W}\cdot\mathrm{m}^2}{\mu\mathrm{m}}\right)$$

where

λ = wavelength (μm)
T = absolute temperature (K)
c_1 = 3.7418×10^8 W*μm^4/m^2
c_2 = 1.4388×10^4 μm*K

The fraction of energy emitted in the wavelength band $0 \le \lambda \le \lambda^*$ is the *blackbody radiation function*,

$$F_{0\to\lambda^*}\left(\lambda^* T\right) = \int_0^{\lambda^* T} \frac{E_{\lambda,b}}{\sigma T^5}\, d\left(\lambda T\right)$$

where σ = Stefan–Boltzmann constant = 5.6696×10^{-8} W/m^2 K^4

The integrand that defines $F_{0\to\lambda^*}$ depends on the product T, not on and T separately.

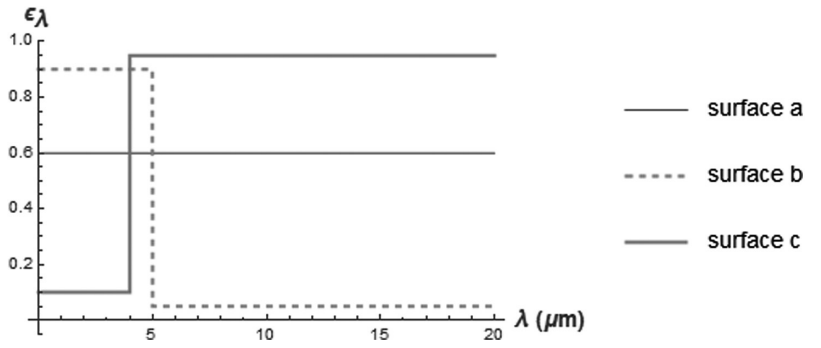
a) Plot $E_{\lambda,b}$ versus λ both on a linear scale and on the classical log-log scale. On each graph, plot various values of T from 50 K to 5800 K (surface of the sun).
b) Create a function to evaluate $F_{0\to\lambda^*}$ as a function of the product λT. Please use this function throughout the remainder of this problem.
c) Plot $F_{0\to\lambda^*}$ in the range $\lambda T = 0$ to 20,000 μm K.
d) The range of visible light for a typical human is 0.4 to 0.7 μm. Determine the fraction of solar radiation in the visible range, assuming the sun as a blackbody at 5800 K.
e) Emissivity is defined as the actual amount of radiation emitted by a real surface divided by the radiation that a blackbody would emit at the same temperature.

$$\text{Spectral emissivity} = \varepsilon_\lambda = \frac{E_\lambda}{E_{\lambda,b}}$$

$$\text{Total emissivity} = \varepsilon = \frac{E}{E_b} = \frac{E}{\sigma T^4}$$

$$E = \int_{\lambda=0}^{\infty} \varepsilon_\lambda E_{\lambda,b}\, d\lambda$$

f) Consider surfaces with the following spectral variations of emissivity. For each case, plot the total emissivity as a function of temperature in the range $T = 0$ to 5000 K.



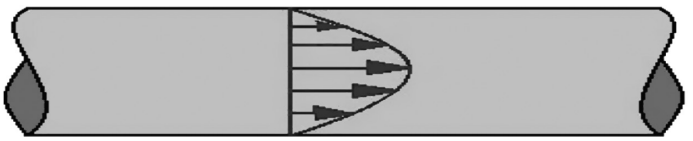
Problem 4.25: Volume Flow Rate

The volume flow rate of a fluid flowing in a round pipe of radius r_0 is

$$Q = \int_0^{r_0} v(r) 2\pi r \, dr$$

The velocity can be approximated by

$$v(r) = 2\left(1 - \frac{r}{r_0}\right)^{1/6}$$



a) Create a function to compute Q as a function of r_0. Evaluate the integral numerically.
b) Use your function to plot Q versus r_0 in the range $0 < r_0 < 6$ cm.

Chapter 5

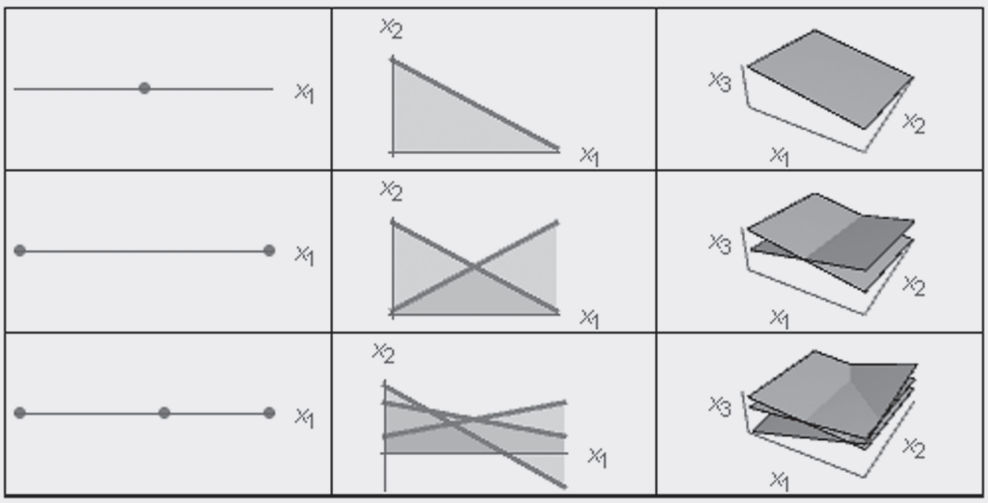
Linear Algebra

CHAPTER OBJECTIVES

This chapter consists of the fundamental concepts and characteristics of simultaneous systems of linear algebraic equations. Numerical methods to evaluate linear systems are presented.

Specific objectives and topics covered are

- Cause and effect
- Selected applications
- Geometric interpretations
- Possibility of solutions
- Characteristics of square, overdetermined, and underdetermined systems
- Row operations
- Determinants and Cramer's rule
- Gaussian elimination and LU factorization
- Gauss–Seidel iteration
- Matrix inversion
- Least squares regression



5.1 INTRODUCTION

We examine the general case of m simultaneous algebraic equations involving n unknown quantities, $x_1, x_2, \cdots, x_n$,

$$\begin{aligned} f_1(x_1, x_2, \cdots, x_n) &= 0 \\ f_2(x_1, x_2, \cdots, x_n) &= 0 \\ &\vdots \\ f_m(x_1, x_2, \cdots, x_n) &= 0 \end{aligned} \tag{5.1}$$

This system of equations can be either linear or nonlinear. The general form of m simultaneous *linear* equations involving n unknown quantities is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{5.2}$$

These simultaneous equations are usually written in the matrix form

$$\boldsymbol{A} * \boldsymbol{x} = \boldsymbol{b} \tag{5.3}$$

where

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \ddots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} \\ x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{5.4}$$

There are numerous general procedures to solve systems of equations, such as Gaussian elimination and LU decomposition. There are numerous other procedures designed to handle special types of systems, such as the Thomas algorithm for tridiagonal systems and conjugate-gradient methods for sparse systems. A variety of iterative procedures, such as the Gauss–Seidel method, are also available to obtain approximate solutions.

5.2 CAUSE AND EFFECT

Systems of coupled, linear algebraic equations can be found in all branches of engineering and physics. Applications are numerous and of great importance. The general linear system has the form $\mathbf{A}*\mathbf{x}=\mathbf{b}$. Usually, the matrix $\mathbf{A}$ contains the *system parameters* that represent how parts of the system interact with other parts. The vector $\mathbf{b}$ contains the *forcing functions* or *external stimuli* acting on the system. The vector $\mathbf{x}$ represents the resulting unknown *response* or *state* of the system, which we are trying to determine. Systems of linear equations can thus be cast as a *cause and effect* relationship, as depicted in Figure 5.1.

This *cause–effect* or *stimuli–response* relationship could also be expressed as

$$[\text{System parameters}]\{\text{Response}\} = \{\text{Stimuli}\}$$

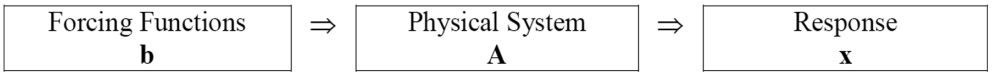


Figure 5.1 Cause and effect relationship between variables in a system of algebraic equations.

5.3 APPLICATIONS

5.3.1 Networks

Many systems of mechanical, electrical, and thermal systems can be modeled as a network of components. For example, consider the cooling of an integrated circuit (IC) package shown in Figure 5.2, where heat is generated in the electronic device at a rate Q_c and dissipated by convection with the surrounding air and conduction through the wall.

The objective here is to find the unknown response variables as a function of the system parameters and forcing functions. The equations required to determine the response are obtained by applying (1) energy conservation to each junction and (2) $Q = \Delta T/R$ in each leg of the thermal circuit. These equations are summarized in Table 5.1.

The energy balance at node T_w can be used to equate $Q_2 = Q_5$ and immediately eliminate Q_5 as an unknown. Thus, the seven remaining unknowns can be put in a matrix by isolating the unknowns to get

$$\begin{bmatrix} R_1 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & R_2 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & R_3 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & R_4 & 0 & -1 & 0 \\ 0 & R_5 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ T_c \\ T_p \\ T_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -T_a \\ -T_a \\ -T_a \\ Q_c \\ 0 \end{bmatrix}$$

Solving these equations gives the unknown heat flows and temperatures.

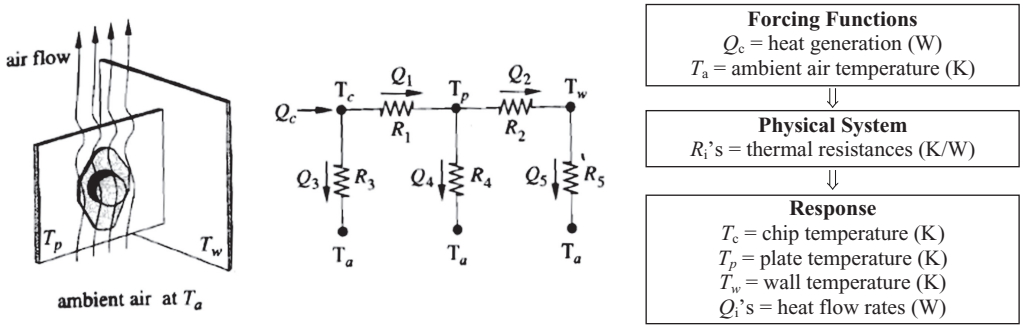


Figure 5.2 Thermal circuit modeling electronic cooling. The variables are defined and categorized in the cause–effect diagram.

Table 5.1 Equations for the thermal circuit

Flow Rates	*Energy Balances*
R_1: $Q_1 = (T_c - T_p)/R_1$	Node T_c: $Q_c = Q_1 + Q_3$
R_2: $Q_2 = (T_p - T_w)/R_2$	Node T_p: $Q_1 = Q_2 + Q_4$
R_3: $Q_3 = (T_c - T_a)/R_3$	Node T_w: $Q_2 = Q_5$
R_4: $Q_4 = (T_p - T_a)/R_4$	
R_5: $Q_5 = (T_w - T_a)/R_5$	

5.3.2 Finite Difference Equations

Another major application of linear algebra occurs in the numerical solution of differential equations. Consider steady one-dimensional heat conduction in a region of length L with internal heat generation $g(x)$. The ends are maintained at fixed temperatures. The model is displayed in Figure 5.3.

The variables are categorized and defined in the following cause–effect diagram.

Forcing Functions	⇒	**System**	⇒	**Response**
g = heat generation rate (W/m³) T_0 = surface temp at $x = 0$ (K) T_L = surface temp at $x = L$ (K)		L = region length (m) k = thermal conductivity (W/m·K)		$T(x)$ = temp distribution (K)

The mathematical formulation of this problem consists of the steady-state energy equation with both boundary temperatures specified. These conditions are expressed with the following mathematical model:

$$\frac{d^2T}{dx^2} + \frac{g}{k} = 0 \tag{5.5}$$

$$\begin{aligned} T &= T_0, \quad x = 0 \\ T &= T_L, \quad x = L \end{aligned} \tag{5.6}$$

The heat conduction equation is approximated at discrete points, $x_i = (i - 1)\Delta x$ (Figure 5.4).

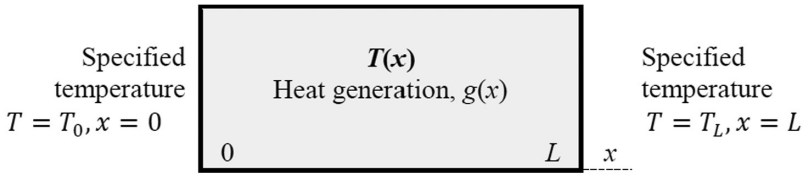


Figure 5.3 Steady-state, one-dimensional heat conduction problem.

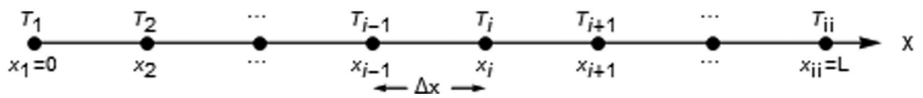


Figure 5.4 Finite difference grid for steady-state, one-dimensional heat conduction.

The finite difference approximation of Equation 5.5 at a typical interior node i is

$$k\left(\frac{T_{i-1}-2T_i+T_{i+1}}{\Delta x^2}\right)+g=0,\ \text{ valid for } i=2,3,\cdots,ii-1 \tag{5.7}$$

This produces the following set of simultaneous linear algebraic equations for the unknown temperatures T_i, $i = 2, 3, \ldots, ii - 1$:

$$\begin{bmatrix} 2 & -1 & 0 & & \cdots & & 0 \\ -1 & 2 & -1 & 0 & \cdots & & 0 \\ 0 & & \ddots & & & & \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ \vdots & & & & \ddots & & \\ & & & & -1 & 2 & -1 \\ 0 & & \cdots & & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} T_2 \\ T_3 \\ \vdots \\ T_i \\ \vdots \\ T_{ii-2} \\ T_{ii-1} \end{bmatrix} = \begin{bmatrix} S+T_0 \\ S \\ \vdots \\ S \\ \vdots \\ S \\ S+T_L \end{bmatrix} \tag{5.8}$$

where $S = g \cdot \Delta x^2 / k$. The temperatures at both surfaces are known, so no finite difference equation is need for nodes 1 and ii. This set of equations has a tridiagonal structure that can be used to solve the system of equations in an extremely efficient manner. This simple example is reminiscent of the more general strategy where differential equations are discretized into a set of simultaneous linear algebraic equations suitable for computation on a computer.

5.4 GEOMETRIC INTERPRETATIONS

Simultaneous linear equations have direct geometric interpretations. For the purpose of visualization, we will consider two simultaneous equations in the form

$$\begin{aligned} &\text{Equation } (1):\ a_{11}x_1 + a_{12}x_2 = b_1 \\ &\text{Equation } (2):\ a_{21}x_1 + a_{22}x_2 = b_2 \end{aligned} \tag{5.9}$$

These can be expressed in matrix form as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{5.10}$$

These can be visualized using either a *row* or a *column* interpretation.

5.4.1 Row Interpretation

The simultaneous solution of these equations for x_1 and x_2 can be visualized graphically as the intersection of the two straight lines defined by Equations 5.9. The intersection of the two equations is visualized in Figure 5.5.

5.4.2 Column Interpretation

Rather than each row being interpreted as an equation, each column can be interpreted as a vector. Specifically, Equation 5.10 is rewritten as

$$\mathbf{C}_1 x_1 + \mathbf{C}_2 x_2 = \mathbf{b}$$

$$\mathbf{C}_1 = \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix},\quad \mathbf{C}_2 = \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix},\quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

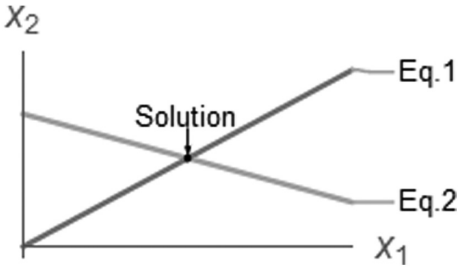


Figure 5.5 Row interpretation of two linear equations.

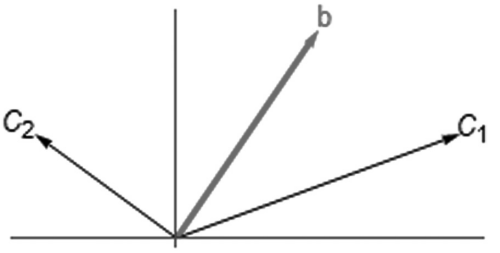


Figure 5.6 Column interpretation of two linear equations.

The graphical interpretation is shown in Figure 5.6.

The solution for x_1 and x_2 must be the values that form the vector **b** as a linear combination of the column vectors $\mathbf{C}_1$ and $\mathbf{C}_2$. Although equations with more than three unknowns ($n > 3$) cannot be visualized graphically, the solutions can still be viewed with either the row or the column interpretation.

5.5 POSSIBILITY OF SOLUTIONS

- **Linear Independence:** An n-by-n system of linear equations has a unique solution only if the columns of the coefficient matrix **A** are *linearly independent*. That is, no column can be obtained as a linear combination of other columns.
- **Rank:** The *rank* of a matrix is the number of linearly independent columns of the coefficient matrix **A**.
- **Consistency:** A system of equations is *consistent* only if solutions exist. **A** and **b** are consistent if and only if **A** and the augmented matrix, $\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{b}]$, have the same rank.

5.6 CHARACTERISTICS OF SQUARE MATRICES

In engineering and applied physics, the most commonly occurring linear systems are those where the number of equations equals the number of unknowns ($m = n$). These are referred to as *square systems*. In order to graphically visualize and understand the range of behaviors that these square systems can exhibit, we again turn to a 2-by-2 system, given by Equation 5.10. Although the following graphical depictions are for this 2-by-2 system, the same types of behavior are found in all n-by-n systems.

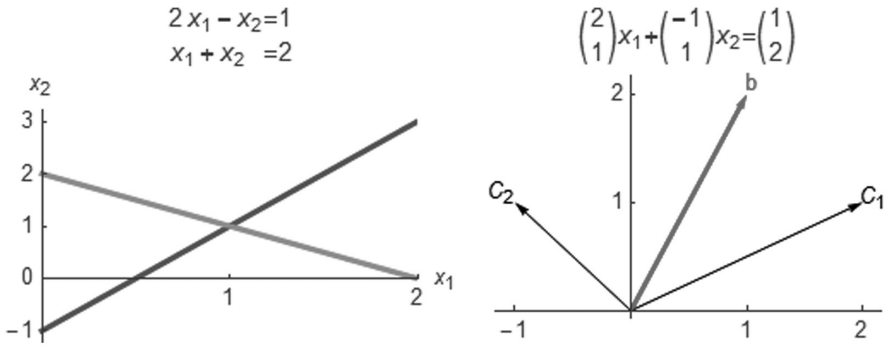


Figure 5.7 Typical behavior of a well-behaved set of simultaneous equations with a unique solution.

(a) **Unique solution**

When the rows and columns of the coefficient matrix **A** are linearly independent, then Rank(**A**) = n, and a unique solution exists. An example is displayed graphically for this case, showing both the row and the column view (Figure 5.7).

The row view shows that the two equations have a distinct intersection corresponding to the unique solution at $(x_1, x_2) = (1,1)$. The columns of the coefficient matrix reach out in different directions and can be added in some combination to reach any point **b** in the vector space.

(b) **No solutions: Singular systems**

If the rows and columns are linearly dependent, then Rank(**A**) < n. In addition, if Rank($\tilde{\mathbf{A}}$) = Rank(**A**) + 1, there can be no solutions. An example of this type of system is shown (Figure 5.8).

The row view shows two parallel lines that never meet, and no combination of (x_1, x_2) can ever satisfy both equations at the same time. The column view shows that the columns are along the same line, while the **b** vector lies along a different line; thus, no combination of $\mathbf{C}_1$ and $\mathbf{C}_2$ can ever reach **b**. This is known as a *singular system*.

(c) **Infinite number of solutions**

This case occurs when the rows and columns are linearly dependent, so that Rank(**A**) < n. However, if also Rank($\tilde{\mathbf{A}}$) = Rank(**A**), the system is consistent, and an infinite number of solutions are possible. An example of this type of system is shown (Figure 5.9).

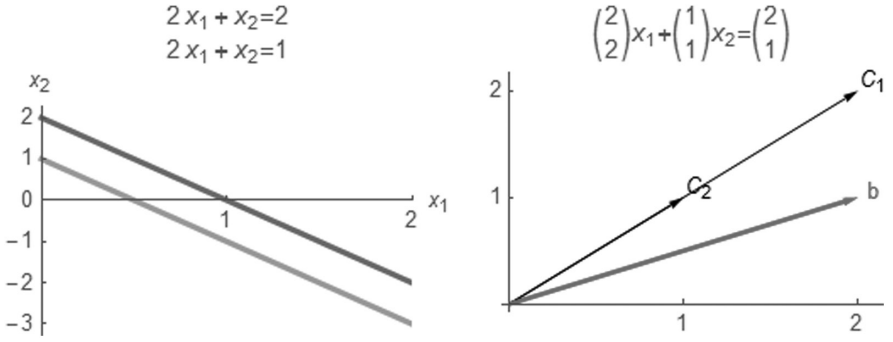


Figure 5.8 A set of simultaneous equations with no solutions.

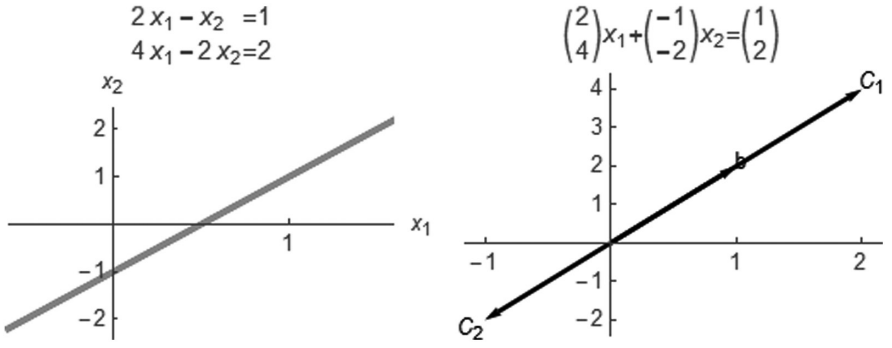


Figure 5.9 A set of simultaneous equations with an infinite number of solutions.

The row view shows that since the equations are identical, any of the infinity of points along the single line is an allowable solution. The column view shows that since $\mathbf{C}_1$, $\mathbf{C}_2$, and **b** all lie in the same direction, an infinite number of combinations of $\mathbf{C}_1$ and $\mathbf{C}_2$ can reach **b**.

(d) **Ill-conditioned systems**

Consider a system of equations that is nearly singular, such as

$$\begin{bmatrix} 2 & 1+\varepsilon \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \tag{5.11}$$

for $\varepsilon << 1$. Results for $\varepsilon = 0.1$ are displayed. The solution is $(x_1, x_2) = (-4.5, 10)$ (Figure 5.10).

The row view shows that the slopes are so close that the point of intersection is difficult to detect visually. The two columns point in slightly different directions and can, in principle, be added to reach any point **b**. However, the solution is numerically sensitive to small changes and round-off error. For instance, consider the following slightly different case with $\varepsilon = -0.1$. The solution has changed drastically to $(x_1, x_2) = (5.5, -10)$. The main characteristic is that a small change in the coefficient matrix has produced a significant change in the solution (Figure 5.11).

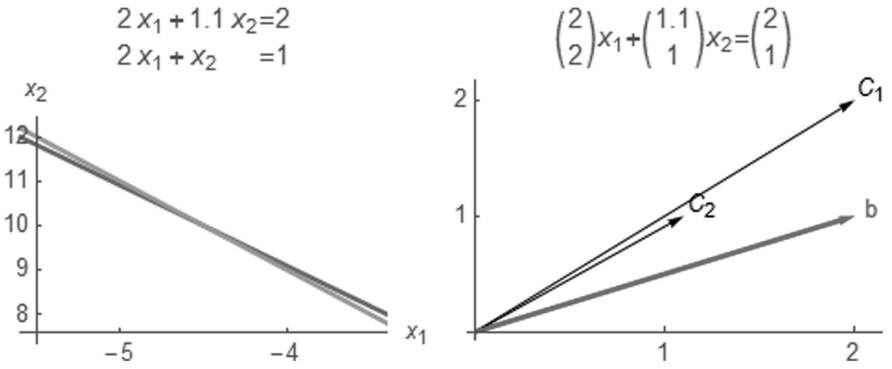


Figure 5.10 Simultaneous ill-conditioned equations.

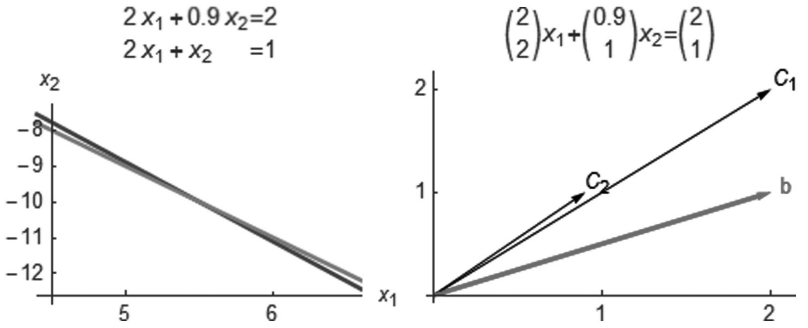


Figure 5.11 Simultaneous ill-conditioned equations. The coefficients are close to those in the previous figure, but the solution is drastically different.

5.7 SQUARE, OVERDETERMINED, AND UNDERDETERMINED SYSTEMS

In addition to square systems of equations containing the same number of equations as unknowns ($m = n$), it is possible to have overdetermined systems containing more equations than unknowns ($m > n$) and underdetermined systems with more unknowns than equations ($n > m$). Examples of these are shown in the panels of Figure 5.12.

5.7.1 Overdetermined Systems

When there are more equations than unknowns, generally no solution can satisfy all the equations simultaneously. These cases are shown in the lower diagonal panels in Figure 5.12. For instance, the lower left panel shows the case of a single variable x_1 taking on the impossible task of satisfying three equations at the same time ($m = 3$, $n = 1$). Similarly, the lower center panel shows two variables attempting to satisfy three equations at the same time. Except for special cases, the three lines representing these three equations will not pass through a single point.

An important application involving overdetermined systems is curve fitting of experimental data. For instance, finding the best polynomial fit to a large data set (regression analysis) involves a large number of equations ($m >> 1$) with only a small number of unknowns corresponding to the coefficients of the polynomial ($n = 2$ for a linear fit, $n = 3$ for a quadratic fit, and so on). In such cases, a least squares optimization technique is often used to determine the best fit.

5.7.2 Underdetermined Systems

When there are more unknowns than equations, there are generally an infinite number of solutions that satisfy the system. These cases are shown in the upper diagonal panels in Figure 5.12. In the upper right panel, a single equation with three variables is shown, and any point on the plane defined by the single equation is a valid solution. Similarly, the middle right panel shows that for two equations with three unknowns, any point along the line of intersection of the two planes is a valid solution. For such cases, some optimization procedure is used to determine the best combination of unknowns.

5.7.3 Square Systems

Systems with the same number of equations as unknowns ($m = n$) generally lead to a unique solution, except for some special cases shown in the previous section. The diagonal panels of Figure 5.12 show such cases. A single equation defines a unique point. Two equations define two lines, and their intersection is the unique solution. Similarly, three equations define

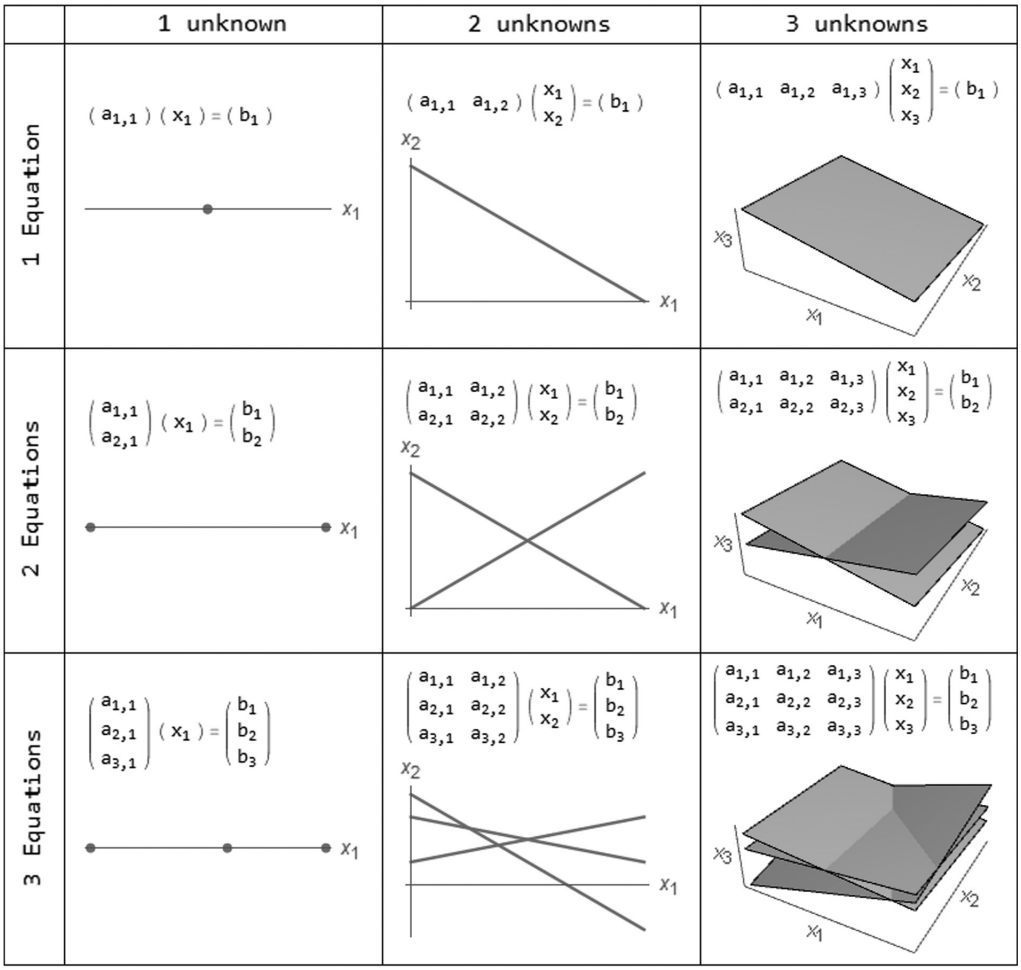


Figure 5.12 Square, overdetermined, and underdetermined systems of equations.

three planes, and their intersection is the unique point satisfying all three. Characteristics of square systems are explored in more detail in the following section.

5.8 ROW OPERATIONS

Solutions of systems of simultaneous linear algebraic equations are obtained by manipulation of matrices using *row operations*. Row operations consist of the following:

- Multiplying a row by a constant
- Adding or subtracting rows
- Exchanging rows

Clearly, any of these operations is mathematically legitimate.

As described in the following sections, direct matrix solution methods such as *Gaussian elimination* and *LU decomposition* use row operations to manipulate the matrices in order to solve for the unknown variables. Other techniques, such as the Gauss–Seidel method, involve iteration.

5.9 THE DETERMINANT AND CRAMER'S RULE

Cramer's rule is a solution technique that is best suited to small numbers of equations. Consider a 2-by-2 matrix

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{5.12}$$

By using row operations, it can be shown that each unknown in a system of equations can be expressed as the ratio of two determinants.

$$\begin{aligned} x_1 &= \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\text{Det}} = \frac{b_1 a_{22} - b_2 a_{12}}{a_{11} a_{22} - a_{21} a_{12}} \\ x_2 &= \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\text{Det}} = \frac{b_2 a_{11} - b_1 a_{21}}{a_{11} a_{22} - a_{21} a_{12}} \end{aligned} \tag{5.13}$$

The determinant Det is a single number composed of the elements of the coefficient matrix **A**, defined as

$$\text{Det} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11} a_{22} - a_{21} a_{12} \tag{5.14}$$

Solutions for higher-order systems have solutions following this same pattern; however, the expressions quickly become unmanageable for large systems of equations, and other methods such as Gauss elimination are used. The determinant of a 3-by-3 matrix is

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{5.15}$$

Determinants of higher-order matrices can all be expressed in terms of the determinants of lower-order systems.

5.10 GAUSSIAN ELIMINATION

5.10.1 Naïve Gaussian Elimination

The Gaussian elimination algorithm consists of two basic steps: (1) eliminate the elements below the diagonal and (2) back substitute to get the solution. The technique will be demonstrated for the 3-by-3 matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{5.16}$$

(1) Forward Elimination of Unknowns

Using row operations, we can eliminate the elements below the diagonal. Start by multiplying row 1 of Equation 5.16 by the factor $f_{21} = a_{21}/a_{11}$ and subtracting the result from row 2 to get

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b'_2 \\ b_3 \end{bmatrix} \tag{5.17}$$

The first element of row 2 is now zero. The other elements in row 2 have also changed and are designated with a prime.

Next, multiply row 1 of Equation 5.17 by a_{31}/a_{11} and subtract the result from row 3. Continue in this manner until all the elements below the diagonal are zero. The result is the following new but equivalent matrix with an upper triangular structure:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & 0 & a''_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b'_2 \\ b''_3 \end{bmatrix} \tag{5.18}$$

(2) Back Substitution

We now can start with the last row of Equation 5.18 to directly solve for x_3 and back substitute to get

$$\begin{aligned} x_3 &= \frac{b''_3}{a''_{33}} \\ x_2 &= \frac{b'_2 - a'_{23}x_3}{a'_{22}} \\ x_1 &= \frac{b_1 - a_{12}x_2 - a_{13}x_3}{a_{11}} \end{aligned} \tag{5.19}$$

This procedure can be readily generalized to any *n*-by-*n* system of linear equations.

5.10.2 Pivoting

The previous technique is called *naïve* because during the elimination and back substitution processes, it is possible that a division by zero can occur. For example, consider the system

$$\begin{bmatrix} 0 & 1 & 2 \\ 5 & 3 & 2 \\ 2 & -1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \\ 4 \end{bmatrix} \tag{5.20}$$

Since the pivot element $a_{11} = 0$, the naïve Gaussian elimination algorithm results in division by zero. The way to avoid these difficulties is to switch the rows so that the coefficient with the largest absolute value is the pivot element. In Equation 5.20, the first two rows can be switched to get an equivalent but well-behaved system.

$$\begin{bmatrix} 5 & 3 & 2 \\ 0 & 1 & 2 \\ 2 & -1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 \\ 5 \\ 4 \end{bmatrix} \tag{5.21}$$

This is known as *partial pivoting*. If the columns are also switched to find the largest element, the procedure is known as *complete pivoting*. Complete pivoting is rarely used, since it changes the order of the unknown x_is; thus, partial pivoting only is used.

Numerical problems with round-off errors can also occur if the magnitude of a pivot element is much smaller than the other elements in the matrix. For instance, switching rows would make the following system less prone to round-off errors:

$$\begin{bmatrix} 0.0001 & 2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 1 \\ 0.0001 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{5.22}$$

5.10.3 Tridiagonal Systems

A special form of the coefficient matrix **A**, often encountered in finite difference and finite element solutions of differential equations, has a banded structure with nonzero elements only on the diagonal, lower diagonal, and upper diagonal. The system **A*****x**=**b** with this form is given by

$$\begin{bmatrix} d_1 & u_1 & 0 & \cdots & 0 \\ l_1 & d_2 & u_2 & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & l_{n-2} & d_{n-1} & u_{n-1} \\ 0 & \cdots & 0 & l_{n-1} & d_n \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix} \tag{5.23}$$

A special version of the Gaussian elimination algorithm can easily be devised to solve Equation 5.23 using the minimum storage and number of arithmetic steps possible. Only the nonzero elements are stored, and unnecessary arithmetic with all the known zeros is not performed. Using row operations, the lower diagonal can be eliminated. Back substituting is then used to find the solution.

5.11 LU FACTORIZATION

LU decomposition or factorization is advantageous for solving systems that have the same coefficient matrices **A** but multiple right-hand-side vectors **b**. LU decomposition is the process of separating the time-consuming elimination part of the Gauss elimination method from the back substitution manipulations of the right-hand-side vector **b**.

Any square matrix can be decomposed or factored into the product of a lower and upper diagonal matrix.

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \tag{5.24}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},\ \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix},\ \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

From the Gauss elimination process, it can be shown that the upper and lower diagonal matrices are

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ f_{21} & 1 & 0 \\ f_{31} & f_{32} & 1 \end{bmatrix},\quad \mathbf{U} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & 0 & a''_{33} \end{bmatrix} \tag{5.25}$$

where the f_{ij}s are the factors used in Gauss elimination to convert **A** to an upper triangular structure (i.e., $f_{21}=a_{21}/a_{11}$, …). **U** is the final upper triangular matrix obtained in the

Gaussian elimination process. Using $\mathbf{A}=\mathbf{LU}$, the system of linear equations $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$ can be written as

$$\mathbf{L}\cdot\underbrace{\mathbf{U}\cdot\mathbf{x}}_{\mathbf{d}}=\mathbf{b}$$

$$\mathbf{L}\cdot\mathbf{d}=\mathbf{b} \tag{5.26}$$

The LU decomposition algorithm is

- Decompose or factor **A** into **LU**.
- Use forward substitution to solve $\mathbf{L}\cdot\mathbf{d}=\mathbf{b}$ for **d**.
- Use back substitution to solve $\mathbf{U}\cdot\mathbf{x}=\mathbf{d}$ for **x**.

5.12 GAUSS–SEIDEL ITERATION

A whole class of solutions is based on iteration rather than direct matrix solutions. A popular method is the Gauss–Seidel iteration. For the 3-by-3 system given by Equation 5.15, the equations are rearranged in the form

$$\begin{aligned} x_1^i &= \frac{b_1 - a_{12}x_2^{i-1} - a_{13}x_3^{i-1}}{a_{11}} \\ x_2^i &= \frac{b_2 - a_{21}x_1^{i-1} - a_{23}x_3^{i-1}}{a_{22}} \\ x_3^i &= \frac{b_3 - a_{31}x_1^{i-1} - a_{32}x_2^{i-1}}{a_{33}} \end{aligned} \tag{5.27}$$

where i is the iteration counter. Starting from an initial guess $\left(x_1^0, x_2^0, x_3^0\right)$ for the solution at $i=0$, we use Equation 5.27 to continue to improve our solution until answers change by less than a specified accuracy.

5.13 MATRIX INVERSION

If a matrix **A** is square, there is a matrix $\mathbf{A}^{-1}$ called the *inverse* with the property that

$$\mathbf{A}\times\mathbf{A}^{-1}=\mathbf{A}^{-1}\times\mathbf{A}=\mathbf{I} \tag{5.28}$$

where **I** is the identity matrix. Once this inverse is known, the solution of a linear system $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$ is

$$\mathbf{x}=\mathbf{A}^{-1}\cdot\mathbf{b} \tag{5.29}$$

Computing the solution in this way using a full matrix inversion is more computationally intensive than other methods such as Gaussian elimination and **LU** factorization, discussed in previous sections. However, the inverse still has uses. For instance, the elements of the inverse matrix represent the response of a single part of a system to a unit stimulus in another part of the system.

The inverse provides a means to discern whether a system is ill-conditioned. Define the matrix norm as

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} \tag{5.30}$$

Then, the matrix condition number is

$$\text{Cond}[\mathbf{A}] = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \tag{5.31}$$

This number will be greater than or equal to 1. The larger the condition number, the more ill-conditioned the matrix is.

5.14 LEAST SQUARES REGRESSION

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response or dependent variable and one or more independent variables. Regression is a method for curve fitting data. Linear regression is the best straight line or linear curve fit to some given data. That is, linear regression is the process of finding the coefficients a_0 and a_1 of the function

$$y(x) = a_0 + a_1 x \tag{5.32}$$

that best fit some given data.

Consider the general case of a data set containing M data points, shown in Table 5.2. The objective is to find the coefficients a_0 and a_1 of the linear function that best fit the given data. We will define a "best fit" in a least squares sense; that is, we will minimize the sum of the squares of the differences between this linear function and the data. The differences between the data and the linear curve fit are shown in Figure 5.13.

The ordinary least squares norm or objective function is defined as

$$S = \sum_{m=1}^{M} \left(Y_m - y(x_m)\right)^2 = \sum_{i=1}^{N} \left(Y_m - (a_0 + a_1 x_m)\right)^2 \tag{5.33}$$

The goal is to minimize this objective function. To find the values of a_0 and a_1 that minimize S, we set the derivatives equal to zero.

$$\frac{\partial S}{\partial a_0} = -2\sum_{m=1}^{M} \left(Y_m - (a_0 + a_1 x_m)\right) = -2\left(\sum_{m=1}^{M} Y_m - a_0 M - a_1 \sum_{m=1}^{M} x_m\right) = 0 \tag{5.34}$$

Table 5.2 Typical data

m	x_m	Y_m
1	x_1	y_1
2	x_2	y_2
…	…	…
M	x_M	Y_M

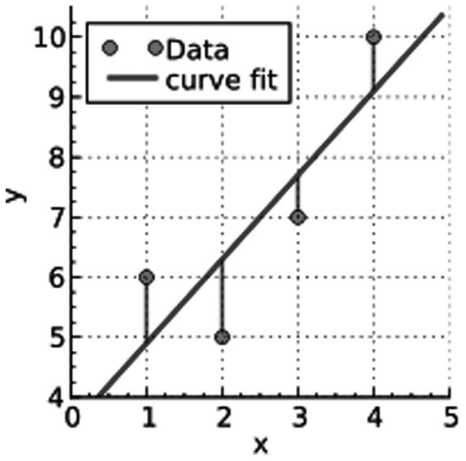


Figure 5.13 Differences between data and a linear curve fit.

$$\frac{\partial S}{\partial a_1} = -2\sum_{m=1}^{M}\left(Y_m - \left(a_0 + a_1 x_m\right)\right)x_m = -2\left(\sum_{m=1}^{M} x_m Y_m - a_0\sum_{m=1}^{M} x_m - a_1\sum_{m=1}^{M} x_m^2\right) = 0$$

Equations 5.34 provide two equations for the two unknowns a_0 and a_1, which are expressed in matrix form as

$$\begin{bmatrix} M & \sum_{m=1}^{M} x_m \\ \sum_{m=1}^{M} x_m & \sum_{m=1}^{M} x_m^2 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{m=1}^{M} Y_m \\ \sum_{m=1}^{M} x_m Y_m \end{bmatrix} \tag{5.35}$$

The solution of these two simultaneous equations gives the coefficients a_0 and a_1. They represent the best straight line curve fit to the data. Using Equations 5.13 for Cramer's rule, we find

$$\begin{aligned} a_0 &= \frac{1}{J}\left(\sum_{m=1}^{M} y_m \sum_{m=1}^{M} x_m^2 - \sum_{m=1}^{M} x_m y_m \sum_{m=1}^{M} x_m\right) \\ a_1 &= \frac{1}{J}\left(M\sum_{m=1}^{M} x_m y_m - \sum_{m=1}^{M} x_m \sum_{m=1}^{M} y_m\right) \\ J &= M\sum_{m=1}^{M} x_m^2 - \left(\sum_{m=1}^{M} x_m\right)^2 \end{aligned} \tag{5.36}$$

The previous concept of a least squares curve fit could be applied to polynomials of any order. For instance, a quadratic curve fit would involve finding the coefficients a_0, a_1, and a_2 of the quadratic polynomial

$$y(x) = a_0 + a_1 x + a_2 x^2 \tag{5.37}$$

The best fit minimizes the sum of the squares of the differences between this quadratic function and the data. The objective function is

$$S = \sum_{m=1}^{M} \left(Y_m - y\left(x_m\right)\right)^2 = \sum_{i=1}^{N} \left(Y_m - \left(a_0 + a_1 x_m + a_2 x_m^2\right)\right)^2 \tag{5.38}$$

The previous steps can be repeated to obtain three simultaneous linear equations for a_0, a_1, and a_2.

PROBLEMS

Problem 5.1

Consider the special cases of the systems of equations **Ax=b** listed.

Case	*A*	*b*
a	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 1 \end{bmatrix}$
b	$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 1 \end{bmatrix}$
c	$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$
d	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$
e	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$
f	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$
g	$\begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 1 & 3 & 4 \\ 1 & 0 & 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$
h	$\begin{bmatrix} 1 & 2 & -3 & 1 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 1 & -1 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$

For all cases:

a) Determine the rank of the coefficient matrix, **A**.
b) Determine the rank of the augmented matrix, $\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{b}]$.
c) Determine whether the system is consistent.
d) For each consistent system, give the solution.

For cases (a)–(f), sketch the row and column interpretations. Summarize these sketches in a table.

Problem 5.2: Equation of a Plane

Consider the equation of a plane written as $c_1x + c_2y + c_3 = z$.

a) Given three known points (x_1, y_1, z_1), (x_2, y_2, z_2), and (x_3, y_3, z_3), derive the system of equations that determine c_1, c_2, and c_3.
b) Write a function "`eq_plane`" that determines the vector of coefficients $\mathbf{c} = [c_1, c_2, c_3]$ given any three points.
c) Test your function by finding the equation passing through the points (1,0,0), (0,1,0), (0,0,1). What are the z values at $(x,y) = (0, 0.5)$, $(0.5, 0)$, $(0.25, 0.25)$, and $(0.5, 0.5)$? Plot this plane.

Problem 5.3: Gravel Pits

A civil engineer involved in a construction process requires a volume V_s of sand, a volume V_{fg} of fine gravel, and a volume V_{cg} of course gravel. There are three pits from which these materials can be obtained. The composition of these pits is

	Sand Fraction	*Fine Gravel Fraction*	*Course Gravel Fraction*
Pit 1	S1	FG1	CG1
Pit 2	S2	FG2	CG2
Pit 3	S3	FG3	CG3

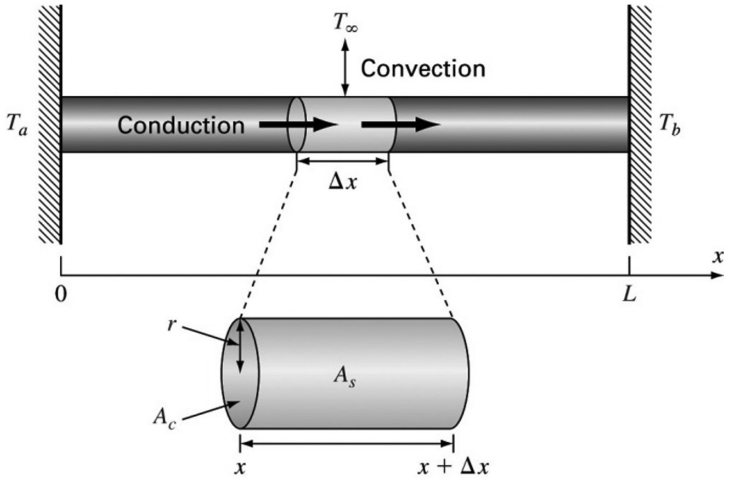
Develop the equations needed to determine the volumes V1, V2, and V3 that must be hauled from pits 1, 2, and 3 to exactly meet the construction needs. Put in matrix form. For the following special case, solve for the volumes V1, V2, and V3.

	Sand Fraction	*Fine Gravel Fraction*	*Course Gravel Fraction*
Pit 1	1	0	0.25
Pit 2	0	1	0.25
Pit 3	0	0	0.5

$V_s = 5000\ m^3$, $V_{fg} = 5000\ m^3$, $V_{cg} = 10{,}000\ m^3$

Problem 5.4: Heated Rod

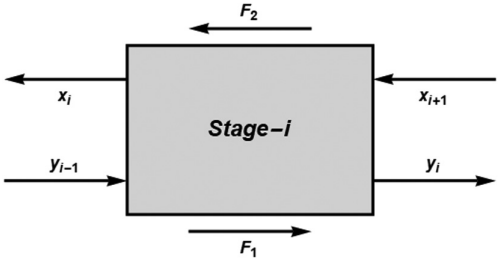
Consider a heated rod with convection from the sides.



a) Write a function to solve for the temperature along the rod for n equally spaced nodes. The function should accept n, L, T_0, T_L, h', and T_a as inputs and output the vectors of the x locations and the computed temperatures. Here, L is the rod length and T_L is the temperature at $x = L$ (T_5=200 in the example).
b) Write a separate code to compute and plot the following cases:
   a. $n = 6$, $L = 1$, $T_0 = 40$, $T_L = 100$, $h' = 0.01$, $T_a = 20$
   b. $n = 10$, $L = 1$, $T_0 = 100$, $T_L = 0$, $h' = 0$, $T_a = 100$
   c. $n = 100$, $L = 1$, $T_0 = 100$, $T_L = 100$, $h' = 0$, 10, and 100, $T_a = 10$ (single graph with three curves for the three h′ values)

Problem 5.5: Stage Extraction Process

A stage extraction process is depicted. In such systems, a stream containing a weight fraction y_{in} of a chemical enters at a mass flow rate F_1. Simultaneously, a solvent carrying a weight fraction x_{in} of the same chemical enters from the other side at a flow rate F_2.



A mass balance at a typical interior stage can be represented as

$$F_1 y_{i-1} + F_2 x_{i+1} = F_1 y_i + F_{21} x_i, \quad i = 2, 3, \ldots, n-1$$

This mass balance must be modified at the first and last stages. At each stage, equilibrium is assumed to be established between x_i and y_i as $K = x_i/y_i$, where K is a distribution coefficient.

a) Draw a cause–effect diagram.
b) Write a function to solve for the concentrations along the stage extractor with n stages.
c) Write a separate function to compute and plot the concentrations and determine y_{out} and x_{out} for the case $F_1 = 500$ kg/h, $y_{in} = 0.1$, $F_2 = 1000$ kg/h, $x_{in} = 0$, and $K = 4$.

Problem 5.6: Pentadiagonal Solver

A pentadiagonal system of equations is a special system with a bandwidth of 5. An n-by-n pentadiagonal system has the following form:

$$\begin{bmatrix} f_1 & g_1 & h_1 & & & \cdots & & & & 0 \\ e_2 & f_2 & g_2 & h_2 & & & & & & \\ d_3 & e_3 & f_3 & g_3 & h_3 & & & & & \\ & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \cdots \\ & & d_i & e_i & f_i & g_i & h_i & & & \\ \cdots & & & d_{i+1} & e_{i+1} & f_{i+1} & g_{i+1} & h_{i+1} & \cdots & \\ & & & & d_{i+2} & e_{i+2} & f_{i+2} & g_{i+2} & h_{i+2} & \\ & & & & & \cdots & \cdots & \cdots & \cdots & \\ & & & & & & d_{n-1} & e_{n-1} & f_{n-1} & g_{n-1} \\ 0 & & & & & \cdots & & d_n & e_n & f_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_i \\ x_{i+1} \\ x_{i+2} \\ \cdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \cdots \\ r_i \\ r_{i+1} \\ r_{i+2} \\ \cdots \\ r_{n-1} \\ r_n \end{bmatrix}$$

a) Write a flowchart or pseudocode describing the logic required to solve this special system of equations. Only the diagonals and RHS should be input. Do not form a full matrix or perform a full matrix solution.
b) Write a function to solve this pentadiagonal system.
c) Test your function with the following special cases.
Case 1:

$$\begin{bmatrix} 8 & -2 & -1 & 0 & 0 \\ -2 & 9 & -4 & -1 & 0 \\ -1 & -3 & 7 & -1 & -2 \\ 0 & -4 & -2 & 12 & -5 \\ 0 & 0 & -7 & -3 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 1 \\ 1 \\ 5 \end{bmatrix}$$

Case 2: d_i's = –1, e_i's = –3, f_i's = 8, g_i's = –3, h_i's = –1, r_i's = 1. Try n = 10,000 and 100,000. Please plot your solutions rather than printing out all these numbers. For comparison, you should try to solve this problem using the standard full matrix solver: x = A\b.

Problem 5.7: Cooling of an IC Package

Consider the cooling of an IC package from Section 5.3.1. The purpose of this exercise is to write a function to solve for the unknown heat flows and temperatures and test this function. You are requested to:

a) Write a function called "`CoolingIC`" that computes the heat flow rates and temperatures in the IC package. The inputs should be: **R** = vector containing the five thermal resistance values, T_a = ambient temperature, and Q_c = power dissipated. The function returns two vectors:
T = a vector containing the three unknown temperatures
Q = a vector containing the four unknown heat flow rates

b) Test your "CoolingIC" function. This function assigns the input values, calls on "CoolingIC" to compute the **T** and **Q** vectors, and displays the output. Consider the following test cases.

Case	R_1 (°C/W)	R_2	R_3	R_4	R_5	Q_c (W)	T_a (°C)
1	2	2	2	2	2	0	25
2	2	2	2	2	2	10	25
3	2	0.5	35	0.7	1	10	25
4	1000	0.5	35	0.7	1	10	25

Problem 5.8: Cooling of an IC Package, Parameter Study

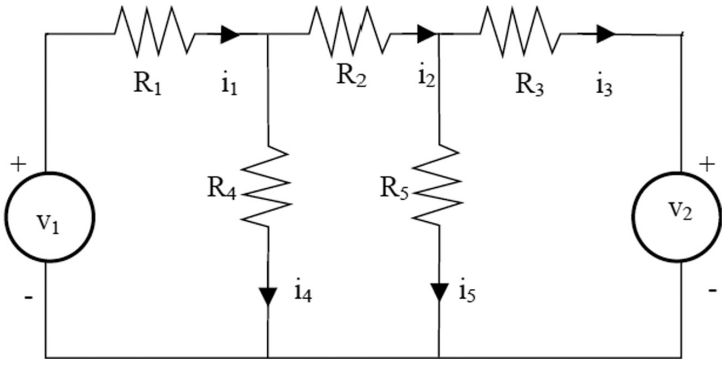
We will now use the function developed in the previous problem to perform a parameter study on the cooling arrangement of the IC system. The most important engineering result of this analysis is the chip temperature, T_c. This is particularly crucial, since electronic chips can fail if overheated. The cooling arrangement must maintain the chip temperature below a critical failure temperature, T_{crit}. As electrical circuits become smaller and draw more power, overheating becomes a limiting constraint in their design. The purpose of this exercise is to perform the following parameter studies.

a) Study of chip temperature as a function of power. Create a plot of T_c versus Q_c with Q_c ranging from 0 to 100 W. On a single graph, put curves for $R_4 = R_5 = 0$, 2, 4, and 6 °C/W. Take other parameter values from Case 3 in the previous problem. What can you conclude about T_c versus Q_c? If $T_{crit} = 150$ °C, what is the maximum allowable chip power, Q_c, for each case?
b) Study of chip temperature as a function of air flow (R_4 and R_5). The resistances R_4 and R_5 represent the effect of the air cooling; high resistance corresponds to low air velocity, and low resistance corresponds to high air velocity. Assume $R_4 = R_5$ and create a plot of T_c versus R_4 with $R_4 = R_5$ ranging from 0 (hurricane) to 100 °C/W (stagnant air). On a single graph, put curves for $Q_c = 0$, 5, 10, and 15 W. Take other parameter values from Case 3 in the previous problem. If $T_{crit} = 150$ °C, what is the allowable range of resistance values for $Q_c = 10$ W?

You should write a function to perform these studies. This function should call on the previously developed function from Problem 5.8.

Problem 5.9: Voltage Range

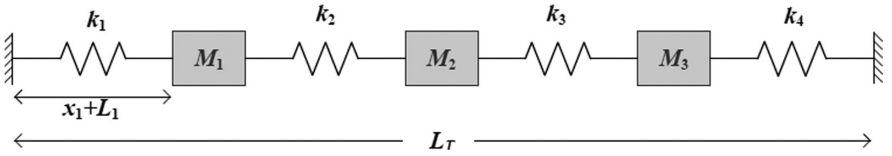
Consider the circuit shown in the figure.



a) If all resistance and voltage values are considered known, derive the equations necessary to determine the five unknown currents.
b) Compute the values of the currents given the following values of the resistances and the voltages:
$R_1 = 5\ \text{k}\Omega$, $R_2 = 100\ \text{k}\Omega$, $R_3 = 200\ \text{k}\Omega$
$R_4 = 150\ \text{k}\Omega$, $R_5 = 250\ \text{k}\Omega$
$v_1 = 100$ volts, $v_2 = 100$ volts
c) Suppose that each resistor is rated to carry a current of no more than 1 milliampere (=0.001 amperes). Determine the allowable range of positive values for the voltage v_2. Use the resistances and v_1 values from part (b).
d) Suppose we want to investigate how the resistance R_3 limits the allowable range for v_2. Obtain a plot of the allowable limits on v_2 as a function of R_3 for $150 \leq R_3 \leq 250\ \text{k}\Omega$.

Problem 5.10: Equilibrium Position of a System of Linear Springs and Masses

Consider the system of masses and linear springs. The total length of each spring is equal to the upstretched length (L_i) plus the stretched length (x_i). The width of each block is W, and the total length of the system is L_T. For a linear spring, the force is directly proportional to elongation, $F_{\text{spring}} = kx$.



a) Considering all the k_i, L_i, and W as known, derive the mathematical model for the stretched lengths x_i.
b) Write a function to solve the system of equations in part (a).
c) Use the function developed in part (b) to determine the x_is for the following cases.

Case 1	*Case 2*
$W = 0.2$ m	$W = 0.2$ m
$L_T = 8$ m	$L_T = 8$ m
$L_i = 1$ m	$L_i = 1$ m
$k_i = 2$ N/m	$k_1 = 1, k_2 = 2, k_3 = 3, k_4 = 4$ N/m

Chapter 6

Nonlinear Algebra

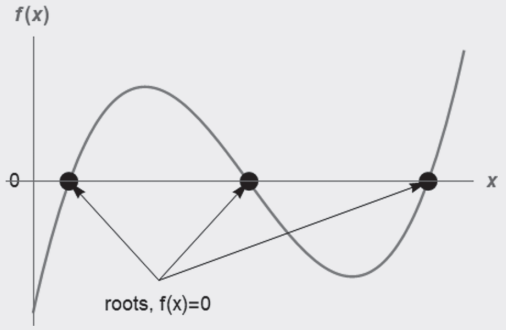
Root Finding

CHAPTER OBJECTIVES

The primary objective of this chapter is to learn how to solve nonlinear systems of algebraic equations. Numerical methods are presented to find the solution of a single nonlinear equation, referred to as *root finding*. Numerical methods to solve simultaneous nonlinear equations are also presented.

Specific objectives and topics covered are

- Introduction and selected applications
- Graphical method
- Bisection method
- False position method
- Newton–Raphson method
- Secant method
- Roots of simultaneous nonlinear systems



6.1 INTRODUCTION

Sometimes, we can solve algebraic equations directly using the rules of algebra. For instance, consider equations such as

$$ax + b = 0 \tag{6.1}$$

$$ax^2 + bx + c = 0 \tag{6.2}$$

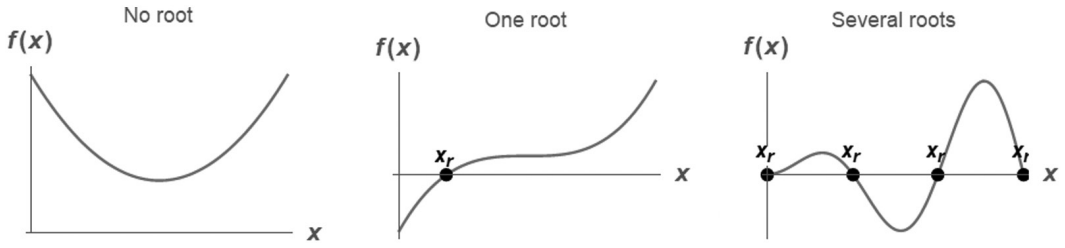


Figure 6.1 Possible roots for various functions **f(x)**.

These can be solved directly for the variable x. However, many times, we encounter a nonlinear algebraic equation of the form

$$f(x) = 0 \tag{6.3}$$

Although a few nonlinear equations can be solved exactly, such as the quadratic equation, a direct expression for x is usually impossible to find. For the function $f(x)$, we want to know the value x_r for which $f(x_r) = 0$. The value x_r is called the *root* or *zero* of the function $f(x)$. The task of estimating x_r numerically is called *root finding*.

Graphs of $f(x)$ versus x for several different nonlinear versions of Equation 6.3 are shown in Figure 6.1. The function $f(x)$ might have no root, one root, or several roots. Root finding can be challenging for complicated nonlinear equations, since we are not sure even how many roots are possible.

A few applications are described in the following section. These are but a small sample of nonlinear algebraic equations from engineering and applied physics.

6.2 APPLICATIONS

6.2.1 Simple Interest

Consider the simple interest formula

$$A = P\frac{i(1+i)^n}{(1+i)^n - 1} \tag{6.4}$$

where

P = present worth
A = annual payments
n = number of years
i = interest rate

Computing A or P knowing the other parameters is easy. However, it is impossible to directly solve for i or n using the basic rules of algebra. To solve for i, for instance, a new function is defined as

$$f(i) = P\frac{i(1+i)^n}{(1+i)^n - 1} - A \tag{6.5}$$

The solution is the value of i for which $f(i) = 0$.

6.2.2 Thermodynamic Equations of State

The equation of state for an ideal gas is

$$Pv = RT \tag{6.6}$$

where

P = absolute pressure (Pa = N·m^2)
v = specific volume (m^3/kmole/K)
R = universal gas constant (kJ/kmol/K)
T = absolute temperature (K)

This ideal gas law is accurate for relatively low pressures and high temperatures. It is easy to find v, P, or T once the other values are specified, since this equation of state is linear in the variables v, P, and T.

An equation of state that is more accurate over a larger pressure and temperature range is the van der Waal equation

$$\left(P + \frac{a}{v^2}\right)(v - b) = RT \tag{6.7}$$

where a and b are empirical constants. Now, given P and T, it is impossible to find v explicitly, since the equation is nonlinear in the variable v. To find v, we define the function

$$f(v) = \left(P + \frac{a}{v^2}\right)(v - b) - RT \tag{6.8}$$

The value of v such that $f(v) = 0$ is the solution we are looking for. The best we can do is to estimate v numerically.

6.2.3 Heat Transfer: Thermal Radiation

Consider a surface exposed to the sun with an insulated (no heat flow) bottom surface as shown in Figure 6.2.

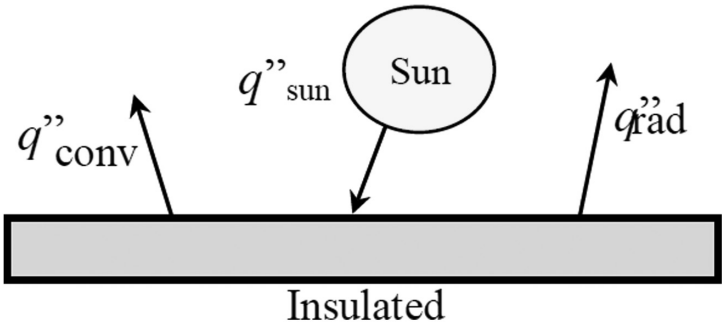
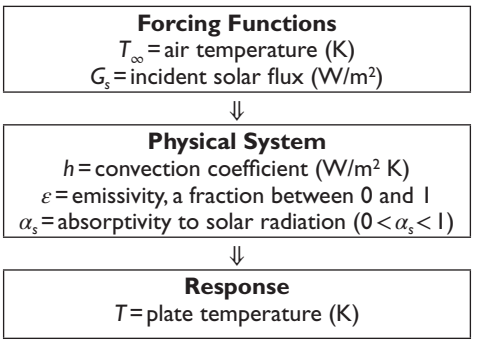


Figure 6.2 Control volume showing thermal processes for the plate in the sun.

The sun heats the surface, supplying an amount of heat per surface area G_s. Heat is lost from the plate to the environment by convection with air at temperature T_∞ and by thermal radiation. These terms are

$$\begin{aligned} q''_{\text{sun}} &= \alpha_s G_s = \text{heat flux absorbed from the sun } \left(\text{W/m}^2\right) \\ q''_{\text{conv}} &= h\left(T - T_\infty\right) = \text{heat flux by convection } \left(\text{W/m}^2\right) \\ q''_{\text{rad}} &= \varepsilon\sigma T^4 = \text{heat flux by radiation } \left(\text{W/m}^2\right) \\ \sigma &= 5.67 \times 10^{-8}\ \text{W/m}^2\text{K}^4 = \text{Stefan} - \text{Boltzmann constant} \end{aligned} \tag{6.9}$$

The cause and effect relationship for this physical problem follows.



One of the most important principles in the thermal sciences is conservation of energy. As described in Section 2.3.3, a general statement of energy conservation is

$$\frac{dE}{dt} = \dot{E}_{\text{in}} - \dot{E}_{\text{out}} + \dot{E}_g \tag{6.10}$$

Apply this principle to the plate for steady-state conditions with no internal heat generation to get

$$0 = q''_{\text{sun}} - q''_{\text{conv}} - q''_{\text{rad}} \tag{6.11}$$

Substitute the expressions in Equation (6.9) to get

$$0 = \alpha_s \cdot G_s - h\left(T - T_\infty\right) - \varepsilon \cdot \sigma \cdot T^4 \tag{6.12}$$

The plate temperature T is impossible to find explicitly, since we have a nonlinear, fourth-order algebraic equation for T. We need to use some numerical root finding technique to estimate T numerically.

6.2.4 Design of an Electric Circuit

An important problem in electrical engineering involves the transient behavior of electric circuits. A typical LRC circuit is shown in Figure 6.3.

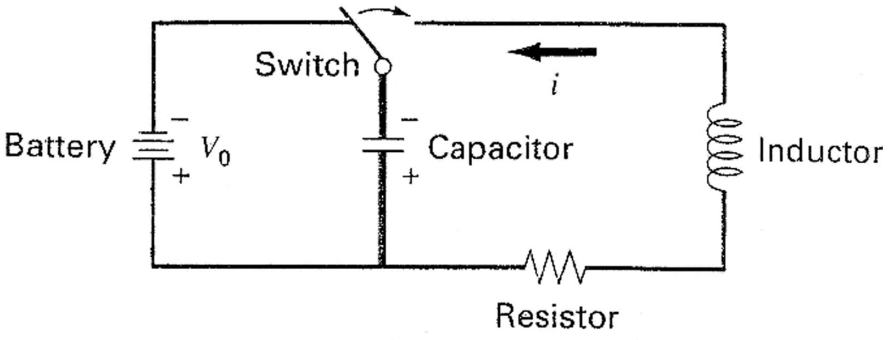


Figure 6.3 An LRC circuit.

When the switch is closed, the current will undergo a series of oscillations until a new steady state is reached. The voltage drops across the basic electric components are:

$$\text{Resistor: } V_R = i \cdot R$$

$$\text{Inductor: } V_L = L\frac{di}{dt} \tag{6.13}$$

$$\text{Capacitor: } V_C = \frac{q}{C}$$

where

i = current (A)
q = charge (coulombs)
V = voltage (V)
R = resistance (Ohm)
L = inductance (H)
C = capacitance (F)

Kirchhoff's voltage law states that the sum of the voltage drops around a closed loop circuit is zero. After the switch is closed, we have

$$L\frac{di}{dt} + R \cdot i + \frac{q}{C} = 0 \tag{6.14}$$

Since current is charge flow rate ($i = dq/dt$), the previous voltage drop equation can be expressed solely in terms of charge as

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = 0 \tag{6.15}$$

The mathematical solution to this equation subject to the initial condition $q = q_0 = CV_0 \text{ at } t = 0$ is

$$q(t) = q_0 \exp\left(-\frac{Rt}{2L}\right)\cos\left(t\sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}\right) \tag{6.16}$$

A typical electrical design problem might require the determination of the proper size resistor to dissipate energy at a specific rate such that q/q_0 is below a specified value in a specified

time. In such a case, we would need to determine the value of R such that the function $f(R) = 0$ where

$$f(R) = \frac{q}{q_0} - \exp\left(-\frac{Rt}{2L}\right)\cos\left(t\sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}\right) = 0 \tag{6.17}$$

All parameters except R would have specified numerical values. This is in the form of a typical root finding problem.

6.3 ROOT FINDING METHODS

A nonlinear algebraic equation is typically impossible to solve explicitly. A few special cases, such as the quadratic equation, can be solved exactly, but these cases are the exception. We could, of course, use the *graphical method* to plot the function and visually estimate the solution. It is always good to visualize mathematical procedures; however, a more systematic procedure using the computer is needed.

There are many possible algorithms for root finding. In general, some initial guess is required to get started, and an iterative procedure must be implemented to estimate the solution numerically. The two major classes are *bracketing methods* and *open methods*, which are distinguished by the type of initial guess.

The most widely used *bracketing methods* are the *bisection method* and the *false position method*. These are based on two initial guesses that bracket or surround the root. The bracket is then iteratively refined until a satisfactory approximation is obtained. These methods always converge to a root but are slow to converge to an accurate solution.

The most widely used *open methods* include *fixed-point iteration*, the *Newton–Raphson method*, and the *secant method*. They require one or more initial guesses, but they do not have to bracket the root. An iterative formula is then applied until a root is found to within an error tolerance. These methods do not always converge to a root, but when they work, they converge rapidly on the root.

6.4 GRAPHICAL METHOD

A simple method to estimate the root of the equation $f(x) = 0$ is to make a plot of the function and observe where it crosses the x-axis. For instance, the function

$$f(x) = x - e^{-x} = 0 \tag{6.18}$$

is shown in Figure 6.4.

The first graph shows that the root lies somewhere between 0.5 and 0.6. The second figure is limited to the range 0.5 and 0.6 and shows a refined estimate for the root between 0.56 and 0.57.

This visual process of graphical refinement could be continued indefinitely. However, the process needs to be automated for a digital computer and is generally used to get a rough estimate or to get a starting guess for numerical methods.

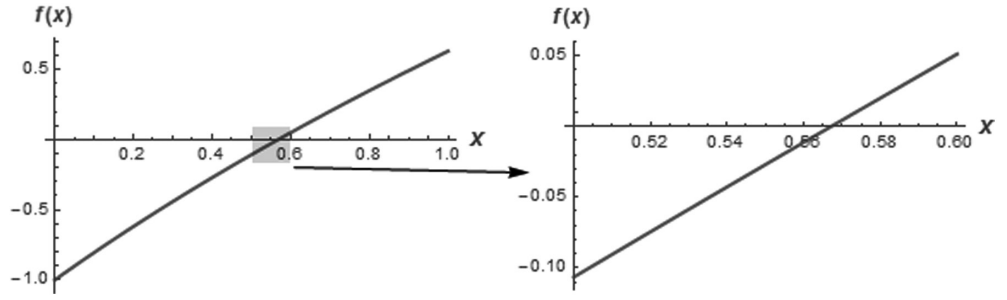


Figure 6.4 Visualization of the root of $\mathbf{f(x) = x - e^{-x}}$ using successively finer intervals.

6.5 BISECTION METHOD

The bisection method is one type of incremental search methods where the interval containing the root is refined by dividing into halves and retaining the subinterval containing the root. This process is repeated until some desired accuracy criterion is met. In order to start the search, an interval that brackets or surrounds the root must be located. In general, if $f(x)$ is real and continuous in the interval from x_l to x_u, and $f(x_l)$ and $f(x_u)$ have opposite signs (i.e., $f(x_l)f(x_u) < 0$), then there exists at least one real root between x_l and x_u. The method is depicted graphically in Figure 6.5.

There are a number of different possible stopping criteria. One is based on the relative change between the most recent iterations of the root. Once a tolerance is specified, the iteration stops once the following criterion is met.

$$\text{Relative error: } \varepsilon_a = \left| \frac{x_r^{\text{new}} - x_r^{\text{old}}}{x_r^{\text{new}}} \right| < \text{tol} \tag{6.19}$$

A second criterion is based on the absolute value of the function at the root.

$$\text{Tolerance in } f(x): \left| f\left(x_r^{\text{new}}\right) \right| < \text{tol} \tag{6.20}$$

In addition, it is good practice to terminate the search after a maximum number of iterations in order to avoid excessive computation times or infinite loops.

6.6 FALSE POSITION METHOD

False position is a variation of the bisection method. Rather than bisecting the interval, an improved estimate is located by joining a straight line or chord between $f(x_l)$ and $f(x_u)$. The improved estimate is the intersection of the straight line with the x-axis, as shown in Figure 6.6.

Using similar triangles implies

$$\frac{x_u - x_r}{f(x_u) - 0} = \frac{x_u - x_l}{f(x_u) - f(x_l)} \tag{6.21}$$

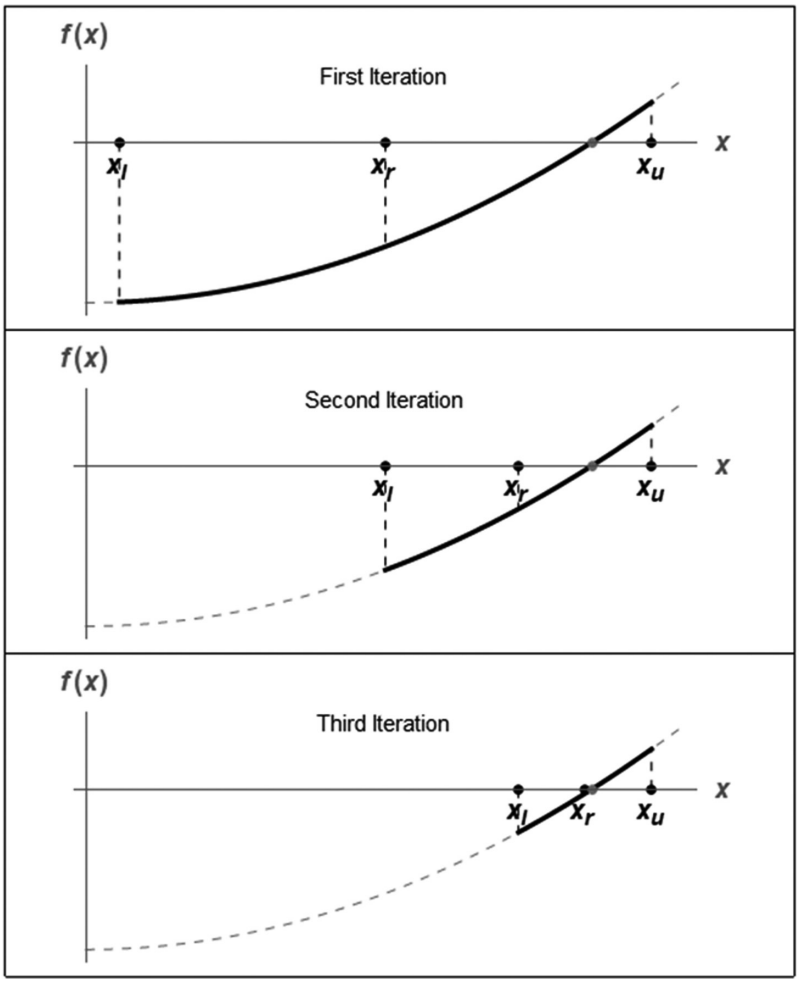


Figure 6.5 Visualization of the bisection method.

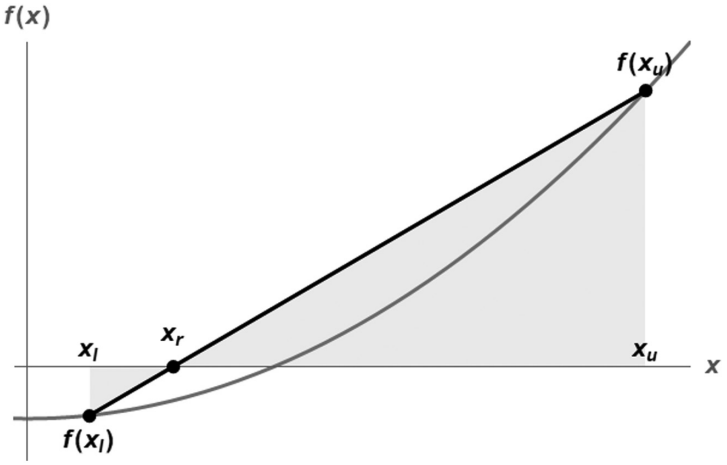


Figure 6.6 Visualization of the false position method.

Thus, starting with x_l and x_u, the rule for an improved estimate is obtained by solving for x_r to get

$$x_r = x_u - f(x_u)\left(\frac{x_u - x_l}{f(x_u) - f(x_l)}\right) \tag{6.22}$$

The bracket containing the root is retained. That is, if $f(x_l)f(x_r) < 0$, then the new bracket is $[x_l, x_r]$; otherwise, the new bracket is $[x_r, x_u]$. The procedure is then repeated until a specified accuracy has been obtained or an iteration limit is reached. As with the previous bisection method, stopping criteria are given by Equations 6.19 and 6.20.

6.7 NEWTON–RAPHSON METHOD

A commonly used root finding method is the *Newton–Raphson method*. This method for solving $f(x) = 0$ uses the tangent to the graph of $f(x)$ at any point and determines where the tangent intersects the x-axis. This intersection is usually an improved estimate of the root. The process is continued until some stopping criterion is met. The method is depicted in Figure 6.7.

Consider that x_i is the current estimate for the root and that x_{i+1} is an improved estimate for the root. Using the concept of a derivative, we have

$$\left(\frac{df}{dx}\right)_{x_i} = f'(x_i) = \frac{f(x_i) - 0}{x_i - x_{i+1}} \tag{6.23}$$

Solving for x_{i+1} produces the following Newton–Raphson iteration formula for an improved estimate:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{6.24}$$

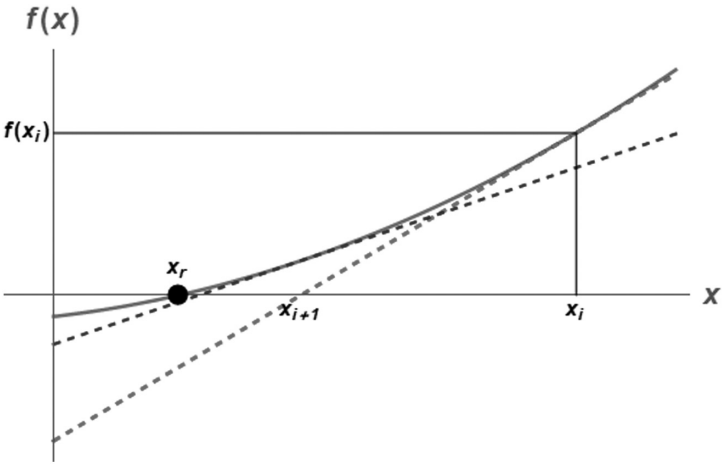


Figure 6.7 Visualization of the Newton–Raphson method.

This formula can also be derived using a Taylor series. The method requires one starting value, x_1. The Newton–Raphson is efficient when it converges but is not always guaranteed to converge. As a general-purpose algorithm for finding zeros of a function, it has three drawbacks:

- The function $f(x)$ must be smooth.
- It might not be convenient to compute the derivative $f'(x)$.
- The starting guess must be sufficiently close to the root.

As with all numerical approximations, caution must be exercised. Although the Newton–Raphson generally converges rapidly, it could also completely diverge from the root. This behavior is caused by the nature of the function and the initial guess. For instance, if the initial guess happens to be at a local maximum or minimum where the derivative is zero, the iteration formula suffers a division by zero error.

Two stopping criteria are generally used:

$$\text{Relative error: } \varepsilon_a = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| < \text{tol} \tag{6.25}$$

$$\text{Tolerance in } f(x): \left| f(x_{i+1}) \right| < \text{tol} \tag{6.26}$$

6.8 SECANT METHOD

For certain functions, the derivative required with the Newton–Raphson method may be inconvenient or impossible to evaluate. An alternative scheme is to use the *secant* rather than the *tangent* to locate an improved estimate, as shown in Figure 6.8.

For this scheme, the derivative is estimated with a finite difference approximation using the two most recent iterates:

$$f'(x_i) \cong \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \tag{6.27}$$

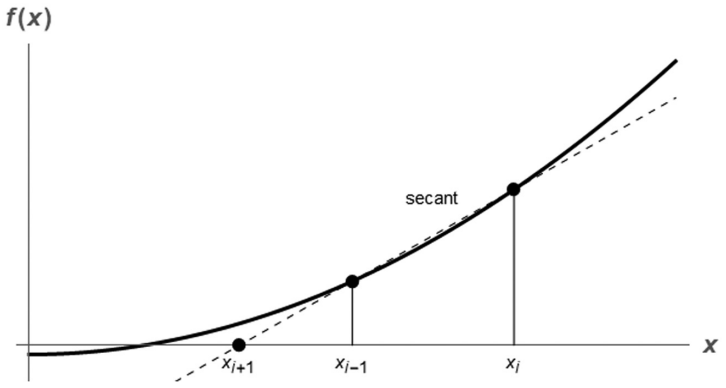


Figure 6.8 Visualization of the secant method.

This approximation is used in the Newton–Raphson formula to get the *secant method* for an improved estimate:

$$x_{i+1} = x_i - f(x_i)\left(\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})}\right) \tag{6.28}$$

This method requires two initial guesses but does not require an explicit evaluation of the derivative.

It should be pointed out that no single method is best for all situations. Even great, professionally developed software such as Mathematica or MATLAB® is not always foolproof. Sophisticated users understand the strengths and weaknesses of the available numerical techniques and are able to select an appropriate strategy.

6.9 ROOTS OF SIMULTANEOUS NONLINEAR EQUATIONS

Up to this point, we have studied methods to solve for the roots of a single equation. The next logical question is to ask about the roots of a set of simultaneous equations in the form

$$\begin{aligned} f_1(x_1, x_2, \cdots, x_n) &= 0 \\ f_2(x_1, x_2, \cdots, x_n) &= 0 \\ &\vdots \\ f_m(x_1, x_2, \cdots, x_n) &= 0 \end{aligned} \tag{6.29}$$

If these equations are linear, the methods of Chapter 5 for linear algebraic equations can be employed. Fortunately, the methods developed in this chapter for finding the root of a single nonlinear equation can be extended to simultaneous sets of nonlinear equations.

The Newton–Raphson method was based on following the derivative to the x-axis in order to obtain an improved estimate. This estimate was based on a first-order Taylor series:

$$f_{i+1} = f_i + (x_{i+1} - x_i)\frac{\partial f_i}{\partial x} \tag{6.30}$$

where

$f_i = f(x_i)$

$$\frac{\partial f_i}{\partial x} = \frac{\partial f(x_i)}{\partial x}$$

Since x_{i+1} is the point that intercepts the x-axis, $f_{i+1} = 0$, and the previous expression can be rearranged to get

$$x_{i+1} = x_i - \frac{f_i}{\dfrac{\partial f_i}{\partial x}} \tag{6.31}$$

An iteration procedure for simultaneous equations can be derived in an identical fashion using multivariable Taylor series. Consider two simultaneous equations:

$$\begin{aligned} f(x,y) &= 0 \\ g(x,y) &= 0 \end{aligned} \tag{6.32}$$

A first-order Taylor series can be written for both equations as

$$\begin{aligned} f_{i+1} &= f_i + (x_{i+1} - x_i)\frac{\partial f_i}{\partial x} + (y_{i+1} - y_i)\frac{\partial f_i}{\partial y} \\ g_{i+1} &= g_i + (x_{i+1} - x_i)\frac{\partial g_i}{\partial x} + (y_{i+1} - y_i)\frac{\partial g_i}{\partial y} \end{aligned} \tag{6.33}$$

The current estimate (x_{i+1}, y_{i+1}) corresponds to the values where $f_{i+1} = g_{i+1} = 0$. These equations provide a 2-by-2 linear system for (x_{i+1}, y_{i+1}). In matrix form,

$$\begin{bmatrix} \dfrac{\partial f_i}{\partial x} & \dfrac{\partial f_i}{\partial y} \\ \dfrac{\partial g_i}{\partial x} & \dfrac{\partial g_i}{\partial y} \end{bmatrix} \begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} x_i \dfrac{\partial f_i}{\partial x} + y_i \dfrac{\partial f_i}{\partial y} - f_i \\ x_i \dfrac{\partial g_i}{\partial x} + y_i \dfrac{\partial g_i}{\partial y} - g_i \end{bmatrix} \tag{6.34}$$

Solving this 2-by-2 system using Cramer's rule and simplifying gives

$$\begin{aligned} x_{i+1} &= x_i - \frac{f_i \dfrac{\partial g_i}{\partial y} - g_i \dfrac{\partial f_i}{\partial y}}{J_i} \\ y_{i+1} &= y_i - \frac{g_i \dfrac{\partial f_i}{\partial x} - f_i \dfrac{\partial g_i}{\partial x}}{J_i} \end{aligned} \tag{6.35}$$

$$J = \text{Jacobian} = \frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial g}{\partial x}\frac{\partial f_i}{\partial y} \tag{6.36}$$

This is the two-equation version of the Newton–Raphson method, and it can be used to iteratively hone in on the roots of two simultaneous equations. As with an equation of a single variable, the method can diverge if a suitable starting guess is not made. This often requires trial and error or a reasonable estimate based on intuition from the physical problem of interest.

The technique can be extended to any number of simultaneous equations.

Example

Consider the task of finding the values (x,y) such that the following simultaneous equations are satisfied:

$$\begin{aligned} f(x,y) &= x^2 + xy - 10 = 0 \\ g(x,y) &= y + 3xy^2 - 57 = 0 \end{aligned} \tag{6.37}$$

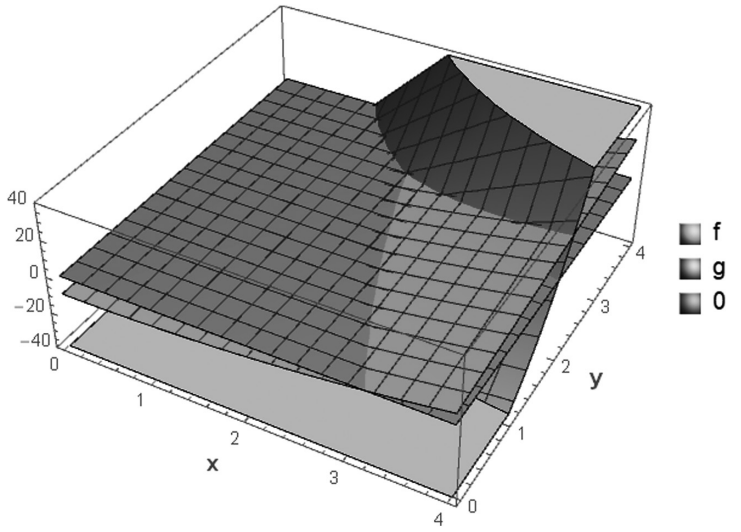


Figure 6.9 The functions **f**(**x**,**y**) and **g**(**x**,**y**). The plane corresponding to the zero is also shown.

These equations are plotted in Figure 6.9 along with the zero plane. The simultaneous solution of Equations 6.37 appears to be approximately $(x,y) = (2,3)$.

For this case, the partial derivatives are

$$\frac{\partial f}{\partial x} = 2x + y, \quad \frac{\partial f}{\partial y} = x$$
$$\frac{\partial g}{\partial x} = 3y^2, \quad \frac{\partial g}{\partial y} = 1 + 6xy \tag{6.38}$$

Substituting Equations 6.38 into the iteration formulas (Equations 6.35) results in

$$x_{i+1} = x_i - \frac{\left(x^2 + xy - 10\right)_i \left(1 + 6xy\right)_i - \left(y + 3xy^2 - 57\right)_i \left(x\right)_i}{J_i}$$
$$y_{i+1} = y_i - \frac{\left(y + 3xy^2 - 57\right)_i \left(2x + y\right)_i - \left(x^2 + xy - 10\right)_i \left(3y^2\right)_i}{J_i} \tag{6.39}$$

$$J = \text{Jacobian} = \left(2x + y\right)\left(1 + 6xy\right) - \left(3y^2\right)\left(x\right) \tag{6.40}$$

After iteration of Equations 6.39 for a sufficient number of times, the iteration converges on the root of $(x,y) = (2,3)$. This solution can be verified by substituting it back into the original Equations 6.37.

The classical bracketing methods, such as the bisection technique, and open methods, such as the Newton–Raphson technique, can be used for any number of simultaneous nonlinear equations. As with a single equation, reasonable initial guesses can make a difference in finding the roots. The issues of convergence become more critical as the number of simultaneous equations increases.

PROBLEMS

Problem 6.1

Consider the equation $e^{-x} = x$. Estimate the real root of this equation in the following ways:

(a) Graphically by plotting the functions e^{-x} and x.
(b) Graphically by plotting the function $f(x) = e^{-x} - x$.
(c) Using three iterations of the bisection method with initial guesses $x_l = 0$ and $x_u = 2$.
(d) Using three iterations of the false-position method with initial guesses $x_l = 0$ and $x_u = 2$.
(e) Using three iterations of the Newton–Raphson method with initial guess $x_0 = 2$.

Summarize your results for parts (c) and (d) with a table:

Iteration	x_l	x_u	x_r
1			
2			
3			

Problem 6.2

Determine the positive real root of the equation

$$\ln\left(x^2\right) = 0.7:$$

(a) Graphically.
(b) Using three iterations of the bisection method, with initial guesses $x_l = 0.5$ and $x_u = 2$.
(c) Using three iterations of the false-position method, with initial guesses $x_l = 0.5$ and $x_u = 2$.

Problem 6.3

Consider a metal plate exposed to the sun with an insulated (no heat flow) bottom surface, as described in Section 6.2.3.

a) Determine the plate temperature, T, on a day when $G_s = 900$ W/m^2 and $h = 15$ W/m^2 K.
b) Compute a table of the plate temperature T for values of G_s ranging from 0 to 1200 W/m^2 in increments of 100 for $h = 15$ W/m^2 K. Make a plot of T versus G_s.
c) Compute a table of the plate temperature, T, for values of the heat transfer coefficient, h, ranging from 10 to 200 W/m^2 K in increments of 10 when $G_s = 900$ W/m^2. Make a plot of T versus h. As h becomes extremely large, what is the plate temperature?

Problem 6.4

The Redlich–Kwong equation of state is given by

$$p = \frac{RT}{v - b} - \frac{a}{v(v + b)\sqrt{T}}$$

where

R = gas constant
T = absolute temperature (K)
p = absolute pressure (kPa)
v = specific volume (m^3/kg)

The parameters a and b are calculated by

$$a = 0.427\frac{R^2T_c^{2.5}}{p_c}, \quad b = 0.0866R\frac{T_c}{p_c}$$

For methane, $R = 0.518$ kJ/(kgK), $p_c = 4580$ kPa, and $T_c = 191$ K. As a chemical engineer, you are asked to determine the amount of methane fuel that can be held in a 3 m^3 tank at a temperature of –50 degrees Celsius with a pressure of 65,000 kPa.

a) Estimate v using the graphical method.
b) Use a root locating method to calculate v, and then determine the mass of methane contained in the tank.
c) How does the Redlich–Kwong equation of state compare with the ideal gas law?

Problem 6.5: Depth of Water in a Tank

You are designing a spherical tank to hold water. The volume of liquid in the tank is

$$V = \pi h^2\frac{(3R-h)}{3}$$

where

V = volume (m^3)
h = depth of water in the tank (m)
R = tank radius (m).

a) We wish to determine the required height h for given values of R and V. Derive the formula that uses the Newton–Raphson method to determine h.
b) What is the range of h values that would physically make sense for an initial guess?
c) Perform one iteration using your formula with parameters $R = 3$ m, $V = 30$ m^3. Use an initial guess of $h_0 = 1$ m.

Problem 6.6: Optimal Fin Spacing

Heat sinks are often attached to electronic devices to increase the cooling efficiency and thereby lower the temperature of the device. One common configuration of these heat sinks is an array of pin fins. Given the overall dimensions of a heat sink consisting of pin fins, it is desirable to know the optimal fin spacing, S_{opt}. The empirical formula for the optimal spacing is

$$\left(\frac{S_{opt}}{D}\right)\frac{2+S_{opt}/D}{\left(1+S_{opt}/D\right)^{2/3}} = 2.75\left(\frac{H}{D}\right)^{1/3}Ra^{-1/4}$$

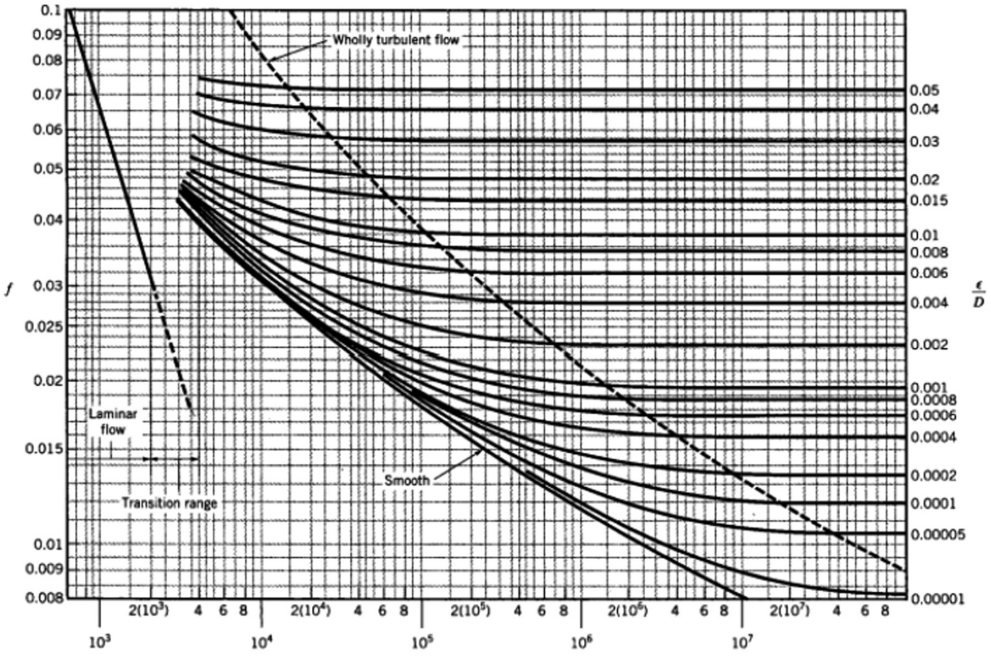
where

D = diameter (m)
H = height (m)
Ra = Rayleigh number

a) Write a function to compute the optimal spacing given D, H, and Ra. Include your algorithm in the form of a flowchart or pseudocode.
b) Use your function to plot S_{opt}/D versus Ra over the range $300 < Ra < 10{,}000$ for $H/D = 5$, 10, 15, and 20.

Problem 6.7: Friction Factor and Moody Diagram

The Moody diagram is a classic found in virtually all books on fluid mechanics and heat transfer. It was originally published by L. F. Moody in the *Transactions of the ASME* in 1944 and is one of those rare diagrams that have passed the test of time and are still used today. This diagram shows the friction factor f as a function of Reynolds number Re_D for various values of the relative roughness ε/D.



$$Re_D = u_m D/\nu$$

$$\mathrm{Re}_D = u_m D/\nu$$

Variable definitions are:

$$f = \text{friction factor} = -\frac{dP}{dx} D/\left(\rho u_m^2/2\right)$$

where

P = pressure (N/m^2)
x = distance along pipe (m)
ε = roughness (m)
D = pipe diameter (m)
$\mathrm{Re}_D = u_m D/\nu$
u_m = mean fluid velocity (m/s)
ν = viscosity (m^2/s)

A correlation between f, Re_D, and ε/D was developed by Colebrook for fully developed turbulent pipe flow:

$$\frac{1}{\sqrt{f}} = -2\log_{10}\left(\frac{\varepsilon/D}{3.7} + \frac{2.51}{\mathrm{Re}_D\sqrt{f}}\right)$$

This equation is valid for $\mathrm{Re}_D > 3000$. For $\mathrm{Re}_D < 3000$, the flow is laminar flow, and the friction factor for all ε/D values is $f = 64/\mathrm{Re}_D$.

a) Write a function to compute the friction factor as a function of ε/D and Re_D. For $\mathrm{Re}_D > 3000$, the Colebrook correlation should be used. An initial guess at the root can be obtained from

$$f = \left(1.8\log_{10}\left(\frac{6.9}{\mathrm{Re}_D} + \left(\frac{\epsilon/D}{3.7}\right)^{1.11}\right)\right)^{-2}$$

For $\mathrm{Re}_D < 3000$, the laminar flow formula, $f = 64/\mathrm{Re}_D$, should be used.

b) Use your function to plot the friction factor data on a log-log scale, just like the Moody diagram. Use the ε/D values shown in the figure.

Problem 6.8: Enzyme Kinetics

The Michaelis–Menten model describes the kinetics of enzyme-mediated reactions:

$$\frac{dS}{dt} = -v_m\frac{S}{k_s + S}$$

where

S = substrate concentration (moles/L)
v_m = maximum uptake rate (moles/L/d)
k_s = half saturation constant, which is the substrate level at which uptake level is half of the maximum (moles/L)

If the initial substrate level at $t = 0$ is S_0, this differential equation can be solved to get

$$S = S_0 - v_m t + k_s \ln\left(S_0/S\right)$$

a) Develop a function to determine S as a function of t, S_0, v_m, and k_s.
b) Use your function to plot S versus t for $S_0 = 0$, 10, and 20 moles/L. Use $v_m = 0.5$ moles/L/d and $k_s = 2$ moles/L. Experiment with the maximum time so that a steady state is reached. Put all three curves on a single graph.

Problem 6.9: Basics

a) Determine all the roots of $f(x) = -14 - 20x + 19x^2 - 3x^3$ graphically.
b) Determine the first root of the function with the bisection method.
c) Determine the first root of the function with the false position method.

For parts (b) and (c), use initial guesses of $x_l = -1$ and $x_u = 0$, and a stopping criterion of $|e_a| \le 1\%$.

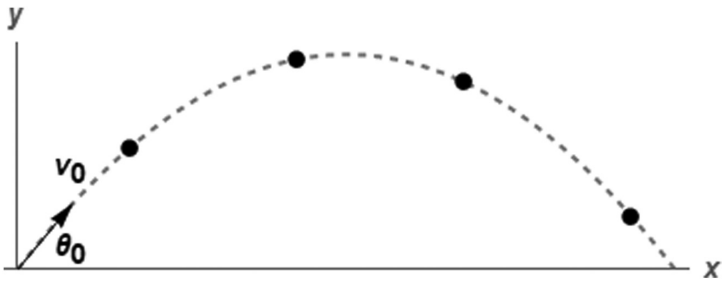
Problem 6.10: Fixed Points of an Ordinary Differential Equation (ODE)

Consider the first-order ODE $dx/dt = e^x - \cos(x)$.

a) Can you find the fixed points (steady state) explicitly?
b) Show the location of the fixed points graphically.
c) For this problem, what is the iteration formula needed to find the fixed points using the Newton–Raphson method?
d) Starting with an initial guess of $x_0 = 10$, what result would the Newton–Raphson converge to after many iterations?

Problem 6.11: Projectiles

Aerospace engineers sometimes compute the trajectories of projectiles such as rockets. A related problem deals with the trajectory of a thrown baseball. The trajectory of a ball thrown by a fielder is defined by the (x, y) coordinates displayed in the figure.



The trajectory can be derived as

$$y = y_0 + x\tan(\theta_0) - \frac{g}{2v_0^2\cos(\theta_0)^2}x^2$$

a) Derive this formula from physical principles.
b) Create an animation of y versus x with controls for the parameters.
c) Create an animation of a parametric plot of $y(t)$ versus $x(t)$, where t is a parameter, along with the other system parameters. This is a "shooting star" animation.
d) Create a function to compute the initial angle θ_0 as a function of the other variables.
e) Use your function to study the effect of a baseball thrown by a fielder to home plate. Create a plot of θ_0 (in degrees) versus x. Use an x-range of 1 to 130 m (deep outfield). On a single graph, plot curves corresponding to $v_0 = 25$, 35, and 45 m/s. Note that 45 m/s is about 100 mph, which is the limit for a top major league baseball player. Also assume that the fielder releases the ball at an elevation of 2 m and the catcher receives it at 0 m.

Problem 6.12: Simultaneous Equations

Consider the following equations. Sketch solutions to each of these equations in the x-y plane and indicate the x, y points that are solutions to both equations. Apply one step of the Newton–Raphson formula using initial guess $(x_0, y_0) = (0, 0)$.

a) $f(x,y) = 2x - y = 0$
$g(x,y) = x + y - 3 = 0$

b) $f(x,y) = y - 2 = 0$
$g(x,y) = x^2 - y = 0$

Problem 6.13: Graphical Solutions

Consider the function

$$f(x) = ax + b - \sin(x)$$

a) Graphically show the roots of $f(x) = 0$ when $a = b = 0$. What are the roots?
b) Graphically show the roots $f(x) = 0$ when $a = 0$ for various values of b. Show the roots by plotting $f(x)$ versus x. On a separate plot, show the roots by plotting both $ax + b$ and $\sin(x)$ versus x. What is the range of b for which roots exist?
c) Graphically show the roots when $b = 0$ for various values of a. Are roots always possible? Are there any ranges of a for which multiple roots exist?
d) Find the root of $f(x) = 0$ for
   - $a = 0.1$ and $b = 0$ with an initial guess of $x_0 = 4$
   - $a = 0.5$ and $b = 1$ with an initial guess of $x_0 = 4$
   - $a = 0.5$ and $b = 1$ with an initial guess of $x_0 = -4$

Problem 6.14: Simple Interest

The formula for simple interest is

$$A = P\frac{i(1+i)^n}{(1+i)^n - 1}$$

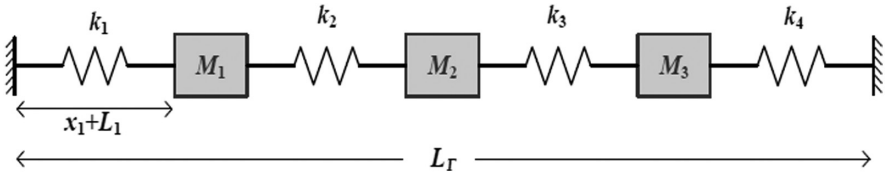
where

P = present worth
A = annual payments
n = number of years
i = interest rate

a) Write a function to compute i as a function of A, P, and n.
b) For $P = \$25{,}000$, $A = \$5000$, and $n = 30$ years, determine i.
c) With $P = \$25{,}000$, plot i versus n for $A = \$1000$ to $\$10{,}000$.

Problem 6.15: Nonlinear Springs

Consider the spring and block system shown with nonlinear springs:



Nonlinear springs can be described by $F_{\text{spring}} = kx + \gamma x^3$, where x is the displacement of the spring from its equilibrium length and k and γ are dependent upon the properties of the spring. For relatively small values of x, the nonlinear term is small, and the spring behaves as a linear spring, that is, $F_{\text{spring}} = kx$. If $\gamma > 0$, the spring is called a *hardening spring*, because it takes more force to cause the same displacement. If $\gamma < 0$, the spring is called a *softening spring*, meaning that it loses its strength after being stretched or compressed.

a) Considering all the k_i, γ_i, L_i, and W as known, derive the mathematical model for the stretched lengths x_i.
b) Write a function to solve the system of equations in part (a).
c) Use the function developed in part (b) to determine the x_is in the following cases.

Case 1	*Case 2*	*Case 3*
$W = 0.2$ m	$W = 0.2$ m	$W = 0.2$ m
$L_T = 8$ m	$L_T = 8$ m	$L_T = 8$ m
$L_i = 1$ m	$L_i = 1$ m	$L_i = 1$ m
$k_i = 2$ N/m	$k_i = 2$ N/m	$k_1 = 1, k_2 = 2, k_3 = 3, k_4 = 4$ N/m
$\gamma_i = 0$ N/m^3	$\gamma_1 = 0.1, \gamma_2 = 0.2, \gamma_3 = 0.3, \gamma_4 = 0.4$ N/m^3	$\gamma_1 = 0.1, \gamma_2 = 0.2, \gamma_3 = 0.3, \gamma_4 = 0.4$ N/m^3

Chapter 7

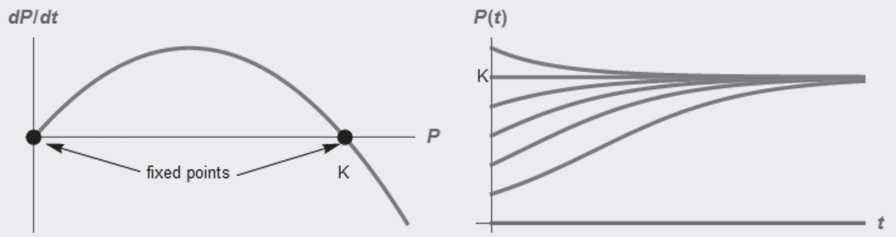
Introduction to Ordinary Differential Equations

CHAPTER OBJECTIVES

The primary objectives of this chapter are to present a logical classification of ordinary differential equations, to visualize their behavior using phase portraits, and to motivate the study of them through physical applications.

The selected topics are

- Classification of ordinary differential equations (ODEs)
- Initial versus boundary value problems
- Phase portraits for first-order ODEs
- First-order linear ODEs with application to thermal and electrical models
- First-order nonlinear ODEs with application to population models
- Phase portraits for second-order ODEs
- Second-order linear ODEs with application to mechanical vibrations and electrical circuits
- Second-order nonlinear ODEs with application to pendulums and predator–prey models
- Second-order boundary value problems typical of steady-state heat conduction



7.1 CLASSIFICATION OF ORDINARY DIFFERENTIAL EQUATIONS

An *ordinary differential equation* (ODE) is a differential equation where the dependent variable or variables depend on only one independent variable (usually *time* or *space*). The order of an ODE refers to the highest derivative or equivalently, to the number of simultaneous equations. ODEs can be classified by the order of the equation as well as whether the system is *linear* or *nonlinear*. Figure 7.1 shows the mathematical forms of these various types of ordinary differential equations.

	Linear	Nonlinear
First-order ODE	$\frac{d\theta}{dt} = a(t)\theta + b(t)$	$\frac{d\theta}{dt} = f(t,\theta)$
Second-order ODE	$\frac{d^2\theta}{dt^2} + c_1\frac{d\theta}{dt} + c_2\theta = F(t)$ or $\frac{d\theta_1}{dt} = a_{11}\theta_1 + a_{12}\theta_2 + b_1$ $\frac{d\theta_2}{dt} = a_{21}\theta_1 + a_{22}\theta_2 + b_2$	$\frac{d^2\theta}{dt^2} = f\left(t,\theta,\frac{d\theta}{dt}\right)$ or $\frac{d\theta_1}{dt} = f_1(t,\theta_1,\theta_2)$ $\frac{d\theta_2}{dt} = f_2(t,\theta_1,\theta_2)$
n^{th}-order ODE	$\frac{d\theta_1}{dt} = a_{11}\theta_1 + a_{12}\theta_2 + \cdots + a_{1n}\theta_n + b_1$ $\frac{d\theta_2}{dt} = a_{21}\theta_1 + a_{22}\theta_2 + \cdots + a_{2n}\theta_n + b_2$ $\vdots$ $\frac{d\theta_n}{dt} = a_{n1}\theta_1 + a_{n2}\theta_2 + \cdots + a_{nn}\theta_n + b_n$	$\frac{d\theta_1}{dt} = f_1(t,\theta_1,\theta_2,\cdots,\theta_n)$ $\frac{d\theta_2}{dt} = f_2(t,\theta_1,\theta_2,\cdots,\theta_n)$ $\vdots$ $\frac{d\theta_n}{dt} = f_n(t,\theta_1,\theta_2,\cdots,\theta_n)$

Figure 7.1 Classification of ordinary differential equations.

7.1.1 Autonomous versus Nonautonomous Systems

In general, the derivative of the dependent variable can be an explicit function of both t and θ, such as the first-order equation $d\theta/dt = f(t,\theta)$. This is referred to as *nonautonomous*. On the other hand, when the derivative has no explicit t dependence, the system is called *autonomous*. The first-order equation now has the form $d\theta/dt = f(\theta)$. With the linear first-order ODE, this means that a and b do not depend on t.

An autonomous system evolves in time but without external sources or interference. The same can be said for all other types of systems listed in Figure 7.1.

7.1.2 Initial Value and Boundary Value Problems

A unique solution for any system of differential equations requires auxiliary conditions. The need for auxiliary conditions can be seen by considering the simplest differential equation,

$$\frac{d\theta}{dt} = 8 \tag{7.1}$$

Multiply by dt and integrate $\int d\theta = 8\int dt$. The result is $\theta = 8t + c$, where c is some unknown constant. A unique solution can only be obtained if an initial condition is specified at the start of the process. For instance, if the process starts at $t = 0$, at which time θ has the value 3, we require

$$\theta = 3, \quad t = 0 \tag{7.2}$$

In order to satisfy this initial condition, c must be 3, and now the unique solution is

$$\theta = 8t + 3 \tag{7.3}$$

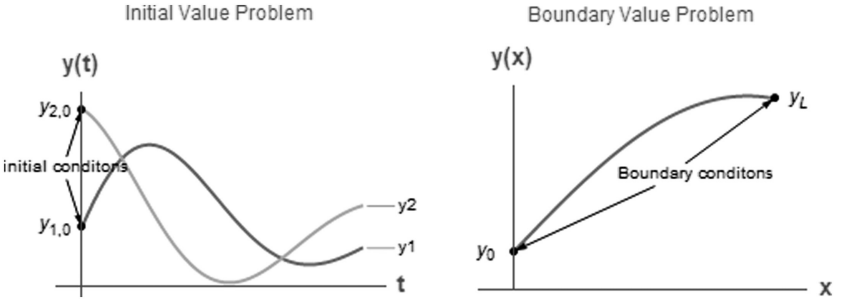


Figure 7.2 Initial versus boundary value problems.

In general, an *n*th-order system requires *n* conditions. That is, one condition is required for a first-order equation, two conditions are required for second-order systems, and so on.

First-order equations always require a single initial condition. However, for second-order or higher systems, the required conditions can be either *initial conditions* or *boundary conditions*. If all the conditions are specified at the same value of the independent variable, then we have an initial value problem. In contrast, if conditions are known at different locations of the independent variable, then we have a boundary value problem. Figure 7.2 contrasts these two scenarios for second-order systems.

The independent variable in the initial value problem is t, and it usually represents time in a transient problem. The independent variable in the boundary value problem is x, since it usually represents position in a spatially distributed problem. These various types of systems are introduced in the following sections.

7.2 FIRST-ORDER ORDINARY DIFFERENTIAL EQUATIONS

7.2.1 First-Order Phase Portraits

Phase portraits are used to show qualitative behavior. Consider a first-order autonomous differential equation.

$$\frac{d\theta}{dt} = f(\theta) \tag{7.4}$$

For now, we will consider only *autonomous* systems, where the function f does not depend explicitly on time. Time-dependent or *nonautonomous* equations of the form $d\theta/dt = f(t,\theta)$ are more complicated, because two pieces of information, θ and t, are needed to predict the future state of the system. These will be discussed later.

Pictures are often more helpful than formulas, especially for analyzing nonlinear systems. We will use a basic technique from dynamics: interpreting a differential equation as a vector field. Consider any autonomous differential equation of the form $d\theta/dt = f(\theta)$. We wish to develop a graphical interpretation of the behavior of this equation. To do this, we simply plot $d\theta/dt$ versus θ, as shown in Figure 7.3.

We think of t as time, θ as the position of an imaginary particle moving along the real line, and $d\theta/dt = f(\theta)$ as the velocity of the particle. Then, the differential equation represents a vector field on the line: it indicates the velocity vector $d\theta/dt$ at each θ. The direction of the

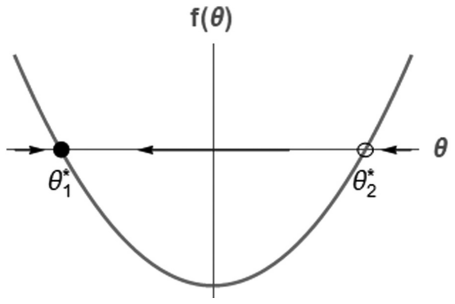


Figure 7.3 Phase portrait for **dθ/dt = f(θ)**.

arrows represents the direction of the motion. At points where $d\theta/dt = 0$, there is no motion. Such points are called *fixed points*. Note the following features shown on this graph.

1. Arrows to the left are drawn in regions where $f(\theta) < 0$, since θ is getting smaller.
2. Arrows to the right are draw in regions where $f(\theta) > 0$, since θ is getting bigger.
3. Fixed points θ^* are values where $f(\theta^*) = 0$. This particular case has two fixed points, θ_1^* and θ_2^*.
4. From this diagram, we conclude that the fixed point θ_1^* is *stable*, since initial θ values on either side move toward or are attracted to θ_1^*. Stable fixed points are indicated by a solid black dot.
5. On the other hand, θ_2^* is an *unstable* fixed point, since points on either side move away from or are repelled by θ_2^*. Unstable fixed points are indicated by an open circle.

This diagram is called a *phase portrait*. From it, we can deduce qualitatively the behavior of the differential equation, including the curvature, starting from any initial condition. A sketch of solution $\theta(t)$, starting from several different initial conditions, must look like Figure 7.4. Note that the correct curvature can be deduced from the phase portrait.

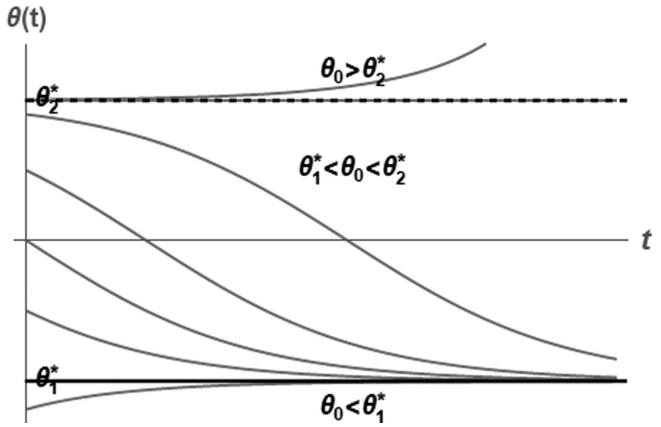


Figure 7.4 Anticipated solution based on phase portrait for **dθ/dt = f(θ)**.

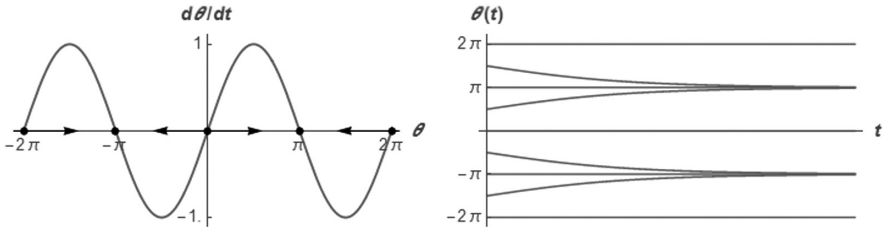


Figure 7.5 Phase portrait and solution for the nonlinear system $\mathbf{d\theta/dt = \sin(\theta)}$.

As another example, we tackle a nonlinear first-order ODE subject to a prescribed initial condition.

$$\frac{d\theta}{dt} = \sin(\theta)$$

$$\theta = \theta_0, \quad t = 0 \tag{7.5}$$

This is a rare nonlinear system in that we can actually solve it analytically. The solution is

$$t = \ln\left|\frac{\csc(\theta_0) + \cot(\theta_0)}{\csc(\theta) + \cot(\theta)}\right| \tag{7.6}$$

Although this result is exact, it is still extremely difficult to envision the nature of the solution. Could you start from some known initial condition and qualitatively sketch the features of the solution, especially as $t \to \infty$? The answer is almost surely no!

In contrast, the graphical analysis is simple and clear, as shown in the following plot of $d\theta/dt = \sin(\theta)$ versus θ in Figure 7.5. Some solutions, $\theta(t)$ versus t, are also plotted. Note that except for specific numerical values, all the qualitative behavior is contained in the phase portrait.

7.2.2 Nonautonomous Systems

The previous discussion dealt with *autonomous* systems, where the function f does not depend explicitly on time. We now turn to time-dependent or *nonautonomous* equations of the form

$$\frac{d\theta}{dt} = f(t,\theta) \tag{7.7}$$

This type of differential is a bit more complicated, because two pieces of information, θ and t, are needed to predict the future state of the system. The time dependency usually results from external sources acting on the system Figure 7.6.

In order to visualize the solution, we would need a three-dimensional graph of $f(t,\theta)$ versus θ and t. An alternative visualization is to plot the *slope field*. That is, for each point (θ, t), the differential equation gives the slope $d\theta/dt$ of the solution passing through that point. The solution always follows a path that is tangent to the local slope.

Some examples are displayed in Figure 7.6. In each case, we can sketch the approximate solution by following the slope at any location.

7.2.3 First-Order Linear Equations

A first-order linear ODE has the specific form

$$\frac{d\theta}{dt} = -a \cdot \theta + b \tag{7.8}$$

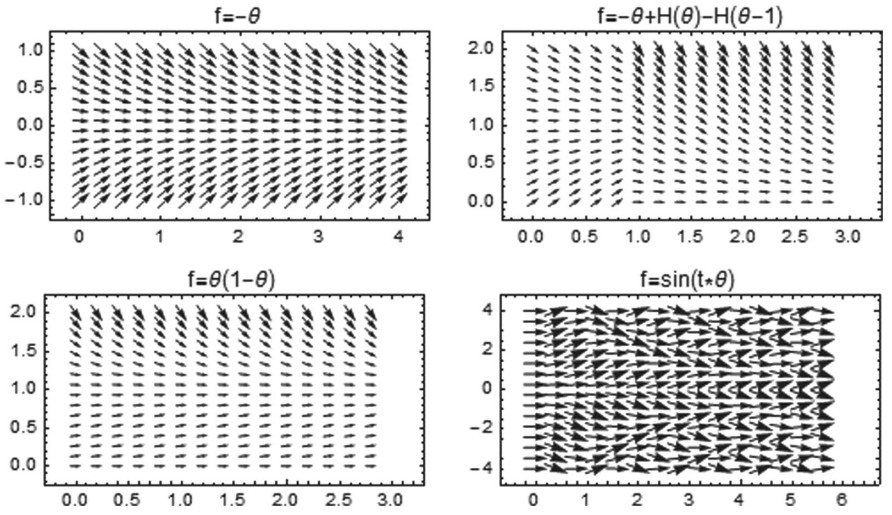


Figure 7.6 Slope fields for some selected ODEs.

If there is no explicit time dependence in the right-hand side (i.e., a and b are constants, independent of t), the system is called *autonomous*. For autonomous systems, the derivative can be plotted as a function of the right-hand side to get a phase plot or phase portrait.

Note that the value of θ where $d\theta/dt = 0$ is the steady state or fixed point. Setting the right-hand side of the differential equation to zero reveals that the fixed point is $\theta^* = b/a$. Since θ increases when $d\theta/dt$ is positive and decreases when $d\theta/dt$ is negative, the solution for $\theta(t)$ must look something like the graph in Figure 7.7(b).

Linear ODEs have applications in all branches of physics and engineering, including the lumped thermal model described next.

7.2.4 Lumped Thermal Models

Consider an object heating up or cooling down due to exposure to a fluid at temperature T_∞ with heat transfer coefficient h and a heat source $g(t)$ (Figure 7.8).

The key assumption is that at any time, the object's temperature $T(t)$ is spatially uniform. This is known as the *lumped capacity approximation*.

An energy balance on the object requires that

$$\begin{aligned}\text{Rate of change of energy} &= -\text{Rate of heat loss to the fluid} \\ &\quad + \text{Rate of heat added by the source}\end{aligned}$$

$$\rho c V \frac{dT}{dt} = -hA_s\left(T - T_\infty\right) + gV \tag{7.9}$$

This first-order linear ODE describes the transient temperature history of the object. A unique solution requires knowledge of the initial temperature at the start of the process.

$$T = T_0, \quad t = 0 \tag{7.10}$$

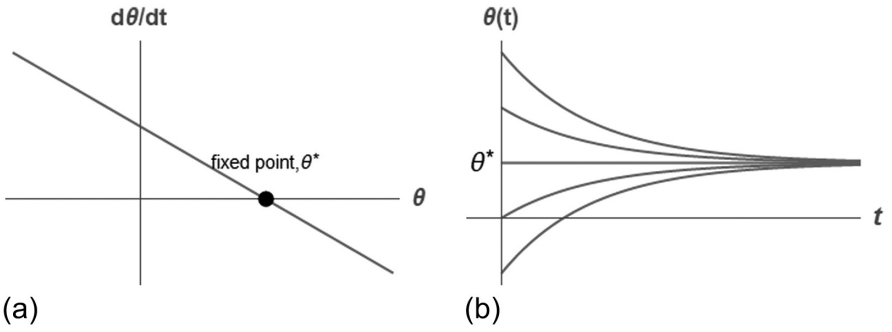
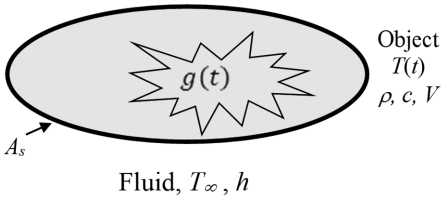


Figure 7.7 (a) Phase portrait and (b) anticipated solution for a linear first-order autonomous ODE.



T(t) = temperature (K or °C)
g(t) = heat source per volume (W/m^3=J/s·m^3)
V = volume (m^3)
A_s = surface area (m^2)
ρ = mass density (kg/m^3)
c = specific heat (J/kg·K or J/kg·°C)
T_∞ = fluid temperature (K or °C)
h = heat transfer coefficient (W/m^2·K or W/m^2·°C)

Figure 7.8 Schematic for the transient thermal analysis of a lumped mass.

Now recast this differential equation in the form

$$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S \tag{7.11}$$

The new variables are:

$$\theta(t) = T(t) - T_\infty = \text{temperature rise}$$

$$\tau = \frac{\rho c V}{h A_s} = \text{time constant}$$

$$S(t) = \frac{g(t)}{\rho c} = \text{source}$$

For a constant S, this equation has only one fixed point at $\theta^* = T^* - T_\infty = \tau S$. The phase portrait and anticipated solution are qualitatively exactly like those shown in Figure 7.7.

Note that regardless of the starting temperature, the system always gravitates toward the stable fixed point, $T^* = T_\infty + \tau S$. This point could be called an attractor for the system. In the absence of any external heat source, $S = 0$, the object always cools down to or heats up to room temperature T_∞. In the following chapters, methods to derive the exact analytical solutions and numerical solutions are presented. The purpose here, however, is visualization and insight.

7.2.5 RC Electrical Circuit

Consider the series RC circuit with a battery of constant direct current (dc) voltage V shown in Figure 7.9.

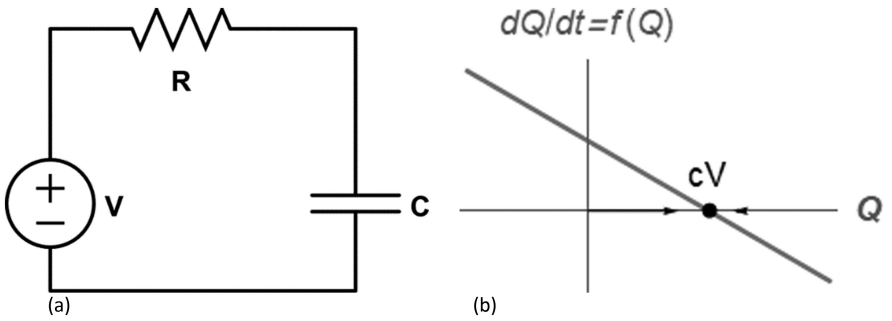


Figure 7.9 An RC circuit: (a) schematic, (b) phase portrait.

Kirchhoff's law applied around the closed loop gives the following equation for the charge Q in the circuit:

$$R\frac{dQ}{dt}+\frac{1}{C}Q=V \tag{7.12}$$

This relation can be rearranged in the form

$$\frac{dQ}{dt}=-\frac{1}{RC}Q+\frac{V}{R}=f(Q) \tag{7.13}$$

Equation 7.13 is plotted in Figure 7.9b. The only fixed point is $Q^*=CV$, and from Figure 7.9b, this can be identified as a *stable* fixed point, as expected from the physics of this elementary circuit.

7.2.6 First-Order Nonlinear Equations

A first-order nonlinear ODE has the general form

$$\frac{d\theta}{dt}=f(t,\theta) \tag{7.14}$$

Although the right-hand side can be any arbitrary and possibly complicated function, the differential equation still simply expresses the rate of change of the dependent variable θ for any t and current value of θ. Thus, for the autonomous case with no explicit t dependence, a phase portrait and anticipated solution can be drawn.

Unlike linear first-order equations with one fixed point, a nonlinear equation might have anywhere from zero to an infinite number of fixed points. In addition, the behavior could possibly be much more complex due to bifurcations, as explored in detail in Section 10.4.

7.2.7 Population Dynamics

A classic example is the modeling of population dynamics. Although population dynamics is an inexact science, many reasonable models have been proposed. Consider the population, P (number of individuals), of a particular species. A population balance suggests that

$$\text{Rate of change of } P = \text{Birth rate} - \text{Death rate} + \text{Immigration rate}$$

Plausible models for the various terms are:

Birth rate = $a\,P$
Death rate = $b\,P^2$
Immigration rate = S

where a and b are the birth rate and death rate constants, respectively. The population balance becomes

$$\frac{dP}{dt} = aP - bP^2 + S \tag{7.15}$$

We also need to specify the initial population at time $t = 0$ to obtain a unique solution.

$$P = P_0, \quad t = 0 \tag{7.16}$$

We now have a candidate mathematical model to predict the population as a function of time.

The simplest model for growth is $dP/dt = rP$, where r is the growth rate. This model predicts exponential behavior, which cannot go on forever. An improved model incorporates the growth rate as a function of P in the form $r(1 - P/K)$, where K is the carrying capacity or maximum sustainable population of the species. This produces the so-called *logistic equation*,

$$\frac{dP}{dt} = r \cdot P\left(1 - \frac{P}{K}\right) \tag{7.17}$$

This is similar to the previous model but based on a different line of reasoning. The phase portrait and anticipated solution, considering only positive values of P (no anti-people), are as shown in Figure 7.10.

There are two fixed points:

- $P^* = 0$: unstable
- $P^* = K$: stable

Based on the phase portrait, we can deduce the qualitative nature of the solution. For any nonextinct starting population, the population will always grow or die toward the carrying capacity; $P^* = K$. If the population starts below $K/2$, the growth rate is slow at first and gradually accelerates until P reaches $K/2$. The growth rate then slows as the population gradually approaches K. In the language of mathematics, the approach to steady state is *asymptotic*.

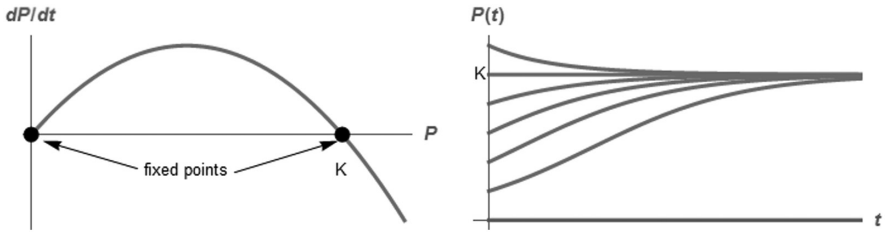


Figure 7.10 Phase portrait and anticipated solution for the logistic equation.

7.3 SECOND-ORDER INITIAL VALUE PROBLEMS

7.3.1 Second-Order Phase Portraits

First, we consider an autonomous second-order system.

$$\begin{aligned}\frac{d\theta_1}{dt} &= f_1(\theta_1,\theta_2)\\ \frac{d\theta_2}{dt} &= f_2(\theta_1,\theta_2)\end{aligned} \tag{7.18}$$

This system can be linear or nonlinear. For first-order autonomous ODEs, it was possible to clearly visualize the dynamics using phase portraits as shown in the previous sections. In principle, this is possible with all systems of ODEs. The difficulty is that for second-order autonomous systems, it is necessary to plot both derivatives as a function of both dependent variables (θ_1,θ_2). This is difficult on a two-dimensional surface.

An alternative is to visualize trajectories moving in the (θ_1,θ_2) plane, referred to as the *phase plane*. We can visualize the general solution $(\theta_1(t),\theta_2(t))$ by thinking of the vector field in terms of the motion of an imaginary fluid. Then, to find the trajectory starting at some point $(\theta_{1,0},\theta_{2,0})$, we place an imaginary particle at that point and watch how it is carried around by the flow. The vector field indicates the trajectories that our imaginary particle would follow. These trajectories are the parametric representation of the solution $(\theta_1(t),\theta_2(t))$, called the *phase portrait*. This shows the overall picture of trajectories in phase space. It is a valuable tool in understanding the fundamental characteristics and getting a feel for the qualitative behavior of second-order systems, particularly *nonlinear* systems where analytical solutions are usually impossible.

Equations 7.18 can be written as a vector field on the phase plane in matrix form as

$$\begin{aligned}\frac{d\boldsymbol{\theta}}{dt} &= \mathbf{f}(\boldsymbol{\theta})\\ \boldsymbol{\theta} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix},\ \mathbf{f} &= \begin{bmatrix}f_1\\ f_2\end{bmatrix}\end{aligned} \tag{7.19}$$

The vector $\boldsymbol{\theta}$ represents a point in the phase plane, and $d\boldsymbol{\theta}/dt$ represents the velocity of that point in the phase plane. By flowing along the vector field, a phase point traces out a solution, $\boldsymbol{\theta}(t)$, corresponding to a trajectory winding through phase space. Furthermore, the entire phase plane is filled with trajectories, since each point can play the role of an initial condition.

Just as in the case of first-order systems, the steady-state or fixed points are crucial. They are determined from the solution of the algebraic equations

$$\begin{aligned}f_1(\theta_1^*,\theta_2^*) &= 0\\ f_2(\theta_1^*,\theta_2^*) &= 0\end{aligned} \tag{7.20}$$

The rigorous mathematical analysis of the stability of the fixed points is presented later.

The *nullclines* are defined as the curves where either $d\theta_1/dt = 0$ or $d\theta_2/dt = 0$. The trajectories in the phase plane are parallel to the θ_2-axis when $d\theta_1/dt = 0$ and are parallel to the

θ_1-axis when $d\theta_2/dt = 0$. The intersections of the nullclines are the fixed points. Graphing the nullclines can be a helpful start in trying to form a phase portrait.

At each point (θ_1,θ_2) in the phase plane, the differential equations define the velocity. Thus, the velocity vector field, indicating the direction and magnitude of the trajectories at each point, can be directly plotted from the system of differential equations. The solution follows the velocity vectors.

A complete phase portrait shows the trajectories moving in the phase plane. To get a feel for the dynamics, begin with a velocity vector field plot. Include the fixed points and nullclines on this plot. Trajectories can be drawn approximately by hand or directly from numerical solutions of the governing equations.

Example

As an example, consider the following system and corresponding phase portrait (Figure 7.11). This system has only one fixed point: $\left(\theta_1^*, \theta_2^*\right) = (0,0)$. Graphing the nullclines can be a helpful start in trying to form a phase portrait. The system assigns a vector ($d\theta_1/dt$, $d\theta_2/dt$) at each point (θ_1,θ_2) and therefore represents a vector field. The vector field can be directly plotted from the system of differential equations, as shown. Starting from any location in the plane, the trajectory is shown by the arrows, and the solution of the system can be anticipated. Several linear and nonlinear applications are described next.

7.3.2 Second-Order Linear Equations

A typical second-order linear ODE has the form

$$\frac{d^2\theta}{dt^2} + c_1\frac{d\theta}{dt} + c_2 \cdot \theta = F \tag{7.21}$$

An equivalent form for an arbitrary second-order system is

$$\begin{aligned} \frac{d\theta_1}{dt} &= a_{11}\theta_1 + a_{12}\theta_2 + b_1 \\ \frac{d\theta_2}{dt} &= a_{21}\theta_1 + a_{22}\theta_2 + b_2 \end{aligned} \tag{7.22}$$

$$\begin{aligned} \frac{d\theta_1}{dt} &= -\theta_1 + 2\theta_2 \\ \frac{d\theta_2}{dt} &= 3\theta_1 - \theta_2 \end{aligned}$$

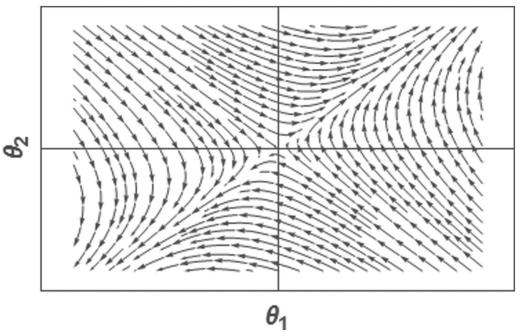


Figure 7.11 Phase portrait (vector field).

It is evident that these two forms are equivalent by defining the new variables

$$\theta_1 = \theta$$
$$\frac{d\theta_1}{dt} = \theta_2 \tag{7.23}$$

The original second-order ODE can be decomposed into simultaneous first-order equations by substituting the variables defined by Equations 7.23 into Equation 7.21 to get

$$\frac{d\theta_1}{dt} = \theta_2$$
$$\frac{d\theta_2}{dt} = -c_1\theta_2 - c_2 \cdot \theta_1 + F \tag{7.24}$$

7.3.3 Mechanical Vibrations

Consider a mass-spring-damper system with an applied force $f(t)$, shown in Figure 7.12.

Apply Newton's second law to the associated free body diagram.

$$m \cdot a = \sum \text{Forces}$$

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = f(t) \tag{7.25}$$

The initial position and velocity need to be specified to complete the mathematical model.

$$\left.\begin{aligned} x &= x_0 \\ \frac{dx}{dt} &= v_0 \end{aligned}\right\} t = 0 \tag{7.26}$$

This differential equation and the two initial conditions describe the motion of the mass. This type of problem is an *initial value problem*, since two conditions at the start of the process are specified.

7.3.4 Mechanical and Electrical Circuits

Figure 7.13 shows a variety of mechanical and electrical second-order systems with identical mathematical structure.

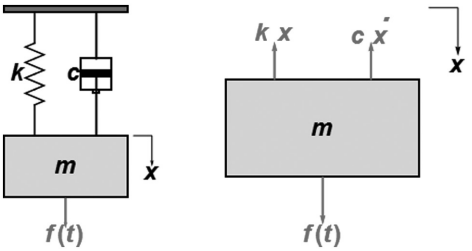


Figure 7.12 Schematic and free body diagram of a mass-spring-damper system. The variables are: $x(t)$ = displacement measured from static equilibrium (m), m = mass (kg), c = damping coefficient (N•s/m), k = spring constant (N/m), $f(t)$ = applied force (N).

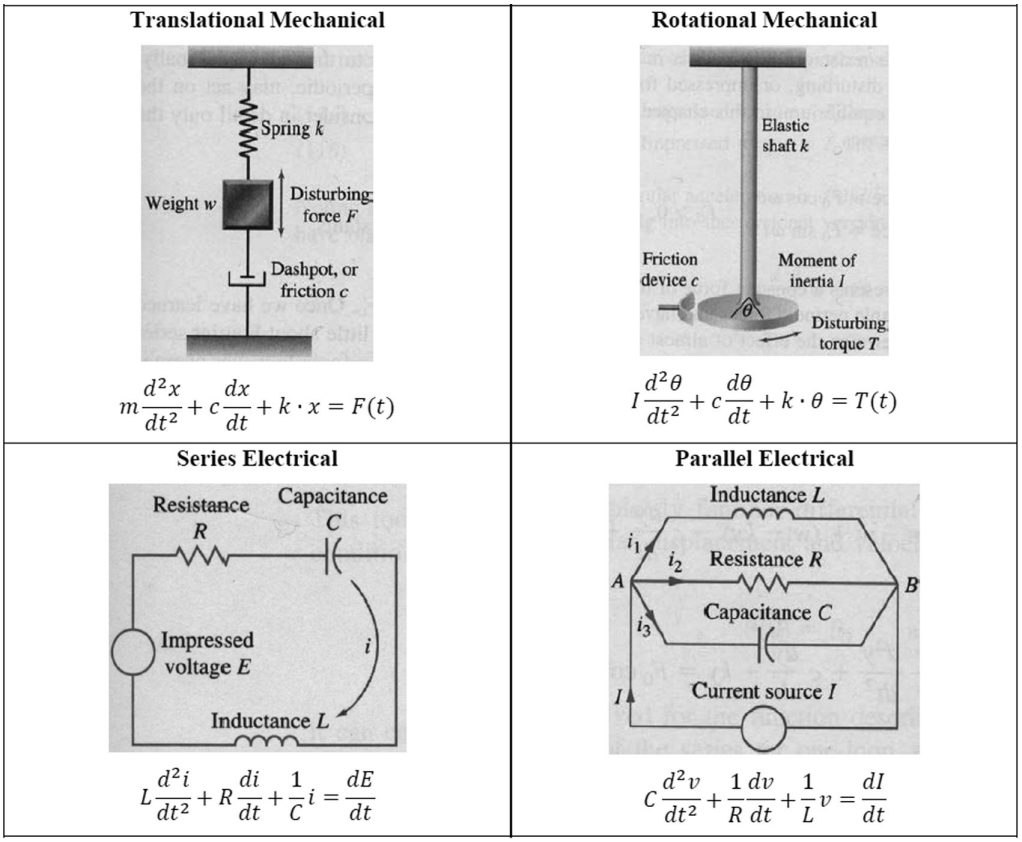


Figure 7.13 The mathematical connection between mechanical and electrical circuits.

7.3.5 Second-Order Nonlinear Equations

Second-order nonlinear systems can be visualized using trajectories in the phase plane, just like linear equations. The equations and resulting dynamics of nonlinear systems can be much more complex, however. There can be several fixed points, and bifurcations can occur. These issues are discussed in detail in Section 11.8. Several nonlinear applications are presented next.

7.3.6 The Pendulum

A damped pendulum with an applied torque is pictured (Figure 7.14).

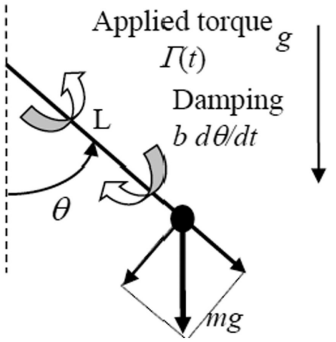
Apply Newton's law for rotational motion.

$$\text{Mass moment of inertia} \times \text{Angular acceleration} = \sum \text{Moments}$$

$$L^2 m\frac{d^2\theta}{dt^2} = -b\frac{d\theta}{dt} - m\cdot g\cdot L\cdot\sin(\theta)+\Gamma(t) \tag{7.27}$$

The initial position and angular velocity need to be specified.

$$\left.\begin{aligned}\theta &= \theta_0\\ \frac{d\theta}{dt} &= v_0\end{aligned}\right\} t=0 \tag{7.28}$$



m = mass (kg)
L = pendulum length (m)
b = viscous damping coefficient (N·m·s)
g = gravitational acceleration (m/s^2)
$\Gamma(t)$= applied torque (N·m)

Figure 7.14 The pendulum.

This is a second-order, nonlinear ODE. This differential equation, along with the two initial conditions, describes the motion of the pendulum.

For highly damped conditions, such as a pendulum swinging in molasses, the damping term dominates the angular acceleration term and Equation 7.27 reduces to

$$b\frac{d\theta}{dt} = -m \cdot g \cdot L \cdot \sin(\theta) + \Gamma(t) \tag{7.29}$$

This reduces to a first-order, nonlinear ODE.

$$L^2 m\frac{d^2\theta}{dt^2} = -b\frac{d\theta}{dt} - m \cdot g \cdot L \cdot \theta + \Gamma(t) \tag{7.30}$$

This reduces to the linear spring-mass system with the same mathematical structure as Equation 7.26.

7.3.7 Predator–Prey Models

Consider the population dynamics of two interdependent species of animals. One species, the prey, is the primary food source for the other species, the predator. For instance, we might have rabbits and wolves. One plausible mathematical model is

$$\begin{aligned}\frac{dR}{dt} &= \alpha_1 R - \delta_1 R \cdot W \\ \frac{dW}{dt} &= -\delta_2 W + \alpha_2 R \cdot W\end{aligned} \tag{7.31}$$

where

R = population of the prey (Rabbits)
W = population of the predator (Wolves)
α_1 and α_2 are the growth rate coefficients
δ_1 and δ_2 are the death rate coefficients.

If left to themselves, the rabbits would grow in proportion to the number of rabbits, while the wolves would starve at a rate proportional to the current population. However, both species could coexist, since the nonlinear interaction terms cause the rabbits to perish and the wolves to thrive in proportion to the product $R{\cdot}W$.

This system is a *second-order* system, since two independent variables (R and W) evolve simultaneously. It is a *nonlinear* set of equations due to the interaction terms. Once the initial populations R_0 and W_0 are specified, the system of ODEs is properly formulated.

7.4 SECOND-ORDER BOUNDARY VALUE PROBLEMS

Consider a plane wall shown in Figure 7.15 with some heat generation $g(x)$. Steady-state conditions exist, and heat flows in only one direction.

Applying the conservation of energy principle to a differential control volume with constant thermal conductivity leads to the following steady-state heat conduction equation:

$$\frac{d^2T}{dx^2} + \frac{g}{k} = 0 \tag{7.32}$$

To complete the mathematical model, boundary conditions at both surfaces must be specified. Specified temperature boundary conditions are

$$\begin{aligned} T &= T_0, \quad x = 0 \\ T &= T_L, \quad x = L \end{aligned} \tag{7.33}$$

Specified heat flux or convection boundary conditions could also be used. In any case, this system is a *boundary value problem*, since one condition at each boundary is specified, as opposed to two conditions at the same starting point required for *initial value problems*, such as the spring-mass system.

7.5 HIGHER-ORDER SYSTEMS

Simultaneous systems of differential equations with any number of equations have applications in physics and engineering. Higher-order systems are difficult to visualize using tools such as phase plots employed earlier for first- and second-order systems, since too many variables need to be graphed. However, much of the intuition developed for linear systems using phase plots is still applicable. The amount of computational expense increases greatly for large numbers of simultaneous equations, but the dynamics are well understood.

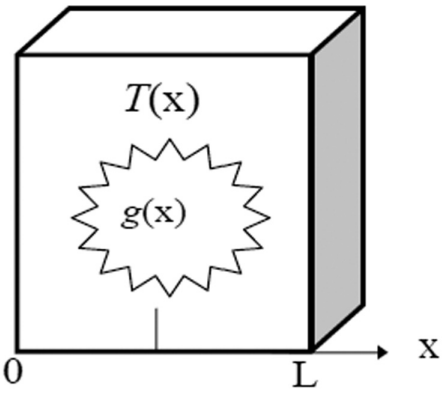


Figure 7.15 Steady heat conduction in a plane wall.

On the other hand, simultaneous systems of nonlinear equations with three or more variables can have surprisingly complex behavior compared with systems with one or two equations. For example, consider the Lorenz equations:

$$\frac{dx}{dt} = \sigma\left(-x + y\right)$$

$$\frac{dy}{dt} = r \cdot x - y - x \cdot z \tag{7.34}$$

$$\frac{dz}{dt} = x \cdot y - b \cdot z$$

This system of equations has only two nonlinear terms ($x{\cdot}z$ in the y equation and $x{\cdot}y$ in the z equation). However, it has extremely complex dynamics once the driving parameter r crosses a certain threshold. The behavior is termed *chaotic*.

PROBLEMS

For each of the first-order ODEs in problems 7.1–7.6:

a) Find all fixed points and classify their stability (stable or unstable).
b) Sketch the phase portrait ($d\theta/dt$ vs. θ)
c) Sketch the anticipated solution ($\theta(t)$ vs. t), starting from several initial conditions.

Clearly label your sketches.

Problem 7.1

$$\frac{d\theta}{dt} = b, \quad b \text{ is a constant.}$$

Draw sketches for $b < 0$, $b = 0$, and $b > 0$.

Problem 7.2

$$\frac{d\theta}{dt} = -\theta$$

Problem 7.3

$$\frac{d\theta}{dt} = -a \cdot \theta$$

Draw sketches corresponding to a zero, a low, and a high value of the parameter a.

Problem 7.4

$$\frac{d\theta}{dt} = -a \cdot \theta + b$$

Draw sketches corresponding to zero, low and high values of the parameter b.

Problem 7.5

$$\frac{d\theta}{dt} = r \cdot \theta(1-\theta), \quad r \text{ is a positive constant}$$

Draw anticipated solution ($\theta(t)$ vs. t) starting from several initial conditions.

Problem 7.6

$$\frac{d\theta}{dt} = r \cdot \theta(1-\theta) + S, \quad r \text{ is a positive constant}$$

Are there any critical values of S? Draw sketches starting from different initial positions.

Problem 7.7

Consider the linear first-order ODE

$$\frac{d\theta}{dt} + \frac{1}{\tau}(\theta - \theta_\infty) = S_c$$

$$\theta = \theta_0, \quad t = 0$$

The parameters τ, θ_∞, and S_c are constants. Assume $\tau > 0$.

a) Find and classify the stability of the fixed points.
b) Sketch the phase portrait for $S_c = 0$ and $S_c > 0$.
c) Sketch the solution based on the phase portrait for different initial conditions. Consider $S_c = 0$ and $S_c > 0$.
d) Based on the previous findings, sketch the anticipated solution for $\theta_\infty = 0$ with a pulsed, time-dependent source

$$S(t) = S_c\left(H(t) - H(t - t_1)\right)$$

Problem 7.8: Tumor Growth

The growth of cancerous tumors can be modeled by the Gompertz law:

$$\frac{dN}{dt} = -a \cdot N \cdot \ln(bN)$$

where
$N(t)$ is proportional to the number of cells in the tumor
$a, b > 0$ are parameters

a) Interpret a and b biologically.
b) Sketch the phase portrait.
c) Sketch $N(t)$ for various initial values.

Problem 7.9

Consider the model chemical reaction

$$A + X \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} 2X$$

in which one molecule of X combines with one molecule of A to form two molecules of X. This means that the chemical X stimulates its own production, a process called *autocatalysis*. This positive feedback process leads to a chain reaction, which eventually is limited by a "back reaction" in which $2X$ returns to $A + X$.

According to the *law of mass action* of chemical kinetics, the rate of an elementary reaction is proportional to the product of the concentrations of the reactants. We denote the concentrations by lowercase letters $x = [X]$ and $a = [A]$. Assume that there's an enormous surplus of chemical A, so that its concentration a can be regarded as constant. Then, the equation for the kinetics of x is

$$\frac{dx}{dt} = k_1 ax - k_{-1} x^2$$

where k_1 and k_{-1} are positive parameters called *rate constants*.

a) Sketch the phase portrait.
b) Find all the fixed points and classify their stability.
c) Sketch the solution based on the phase portrait for different initial conditions.

Problem 7.10: Chemical Kinetics

Consider the chemical reaction system

$$A + X \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} 2X$$

$$X + B \overset{k_2}{\rightarrow} C$$

This is a generalization of Problem 7.9. The new feature is that X is used up in the production of C.

a) Assuming that both A and B are kept at constant concentrations a and b, show that the law of mass action leads to an equation of the form $dx/dt = c_1 x - c_2 x^2$, where x is the concentration of X, and c_1 and c_2 are constants to be determined.
b) Find and classify all the fixed points. Show that $x^* = 0$ is stable when $k_2 b > k_1 a$, and explain why this makes sense chemically.
c) Sketch the phase portraits and anticipated solutions for all the qualitatively different possibilities.

Problem 7.11: Allee Effect

For certain species of organisms, the effective growth rate $\dot{N}/N$ is highest at intermediate N. This is called the *Allee effect*. For example, imagine that it is too hard to find mates when N is very small, and there is too much competition for food and other resources when N is large.

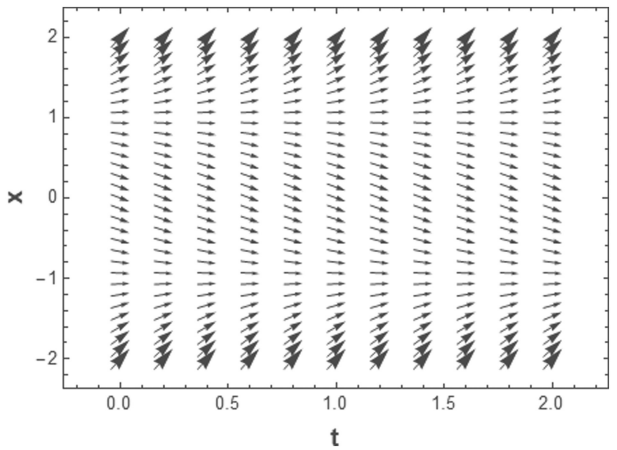
a) Show that $\frac{\dot{N}}{N} = r - a(N-b)^2$ provides an example of the Allee effect if r, a, and b satisfy certain constraints, to be determined.
b) Find all the fixed points of the system and classify their stability.
c) Sketch the solutions $N(t)$ for different initial conditions.
d) Compare the solutions $N(t)$ with those found for the logistic equation. What are the qualitative differences, if any?

Problem 7.12: Slope Fields

Consider an autonomous first-order ODE:

$$\frac{dx}{dt} = f(x)$$

The slope field is shown.



a) Sketch the solution $x(t)$ versus t starting from $x_0 = -2$, -1, 0, 1, and 2.
b) Sketch the phase portrait, dx/dt versus x.
c) Propose a possible function $f(x)$ consistent with the slope field.

Chapter 8

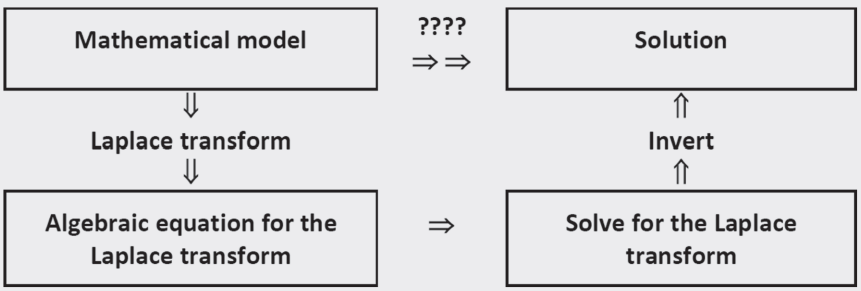
Laplace Transforms

CHAPTER OBJECTIVES

The basics of the Laplace transform are described. The Laplace transform is a powerful method for solving linear ordinary and partial differential equations.

Specific objectives and topics covered are

- The basic definition of the Laplace transform
- Laplace transform pairs
- Properties of the Laplace transform
- Inverting Laplace transforms using partial fraction expansion
- Solutions of ordinary differential equations using Laplace transforms
- Transfer functions



8.1 DEFINITION OF THE LAPLACE TRANSFORM

The Laplace transform method is a mathematical technique that can be used to obtain solutions to linear, time-invariant systems of differential equations. The advantage is that the method reduces the differential equation in time to an algebraic equation, which can be inverted to get the solution.

The Laplace transform of a function $f(t)$ is defined as

$$\mathcal{L}\left[f(t)\right] = F(s) = \int_{t=0}^{\infty} f(t)e^{-st}dt \tag{8.1}$$

The inverse Laplace transform recovers the function $f(t)$:

$$f(t) = \mathcal{L}^{-1}\left[F(s)\right] = \frac{1}{2\pi j}\int_{s=\gamma-j\infty}^{\gamma+j\infty} F(s)e^{st}ds \tag{8.2}$$

$$j = \sqrt{-1}$$

We seldom use this inversion integral. Usually, the inverse is obtained by matching the transform to known Laplace transform pairs.

8.2 LAPLACE TRANSFORM PAIRS

Some important cases of Laplace transform pairs are listed in Table 8.1. The value of having this table is that if we can match the Laplace transform $F(s)$ with one of these cases, we immediately know the function $f(t)$.

8.3 PROPERTIES OF THE LAPLACE TRANSFORM

There are some useful properties of Laplace transforms. One of these is the addition of functions given by

$$\mathcal{L}\left[f_1(t)+f_2(t)\right] = \int_{t=0}^{\infty}\left(f_1(t)+f_2(t)\right)e^{-st}dt = \mathcal{L}\left[f_1(t)\right]+\mathcal{L}\left[f_2(t)\right] \tag{8.3}$$

Table 8.1 Laplace transform pairs

Inverse Laplace Transform $f(t) = \mathcal{L}^{-1}\left[F(s)\right]$	*Laplace Transform* $F(s) = \mathcal{L}\left[f(t)\right]$
Delta function, $\delta(t)$	1
Unit step, H(t)	$1/s$
Ramp, t	$1/s^2$
e^{-at}	$\frac{1}{a+s}$
te^{-at}	$\frac{1}{(a+s)^2}$
$t^n e^{-at}$	$\frac{n!}{(a+s)^{n+1}}$
$\mathrm{Sin}(\omega t)$	$\frac{\omega}{s^2+\omega^2}$
$\mathrm{Cos}(\omega t)$	$\frac{s}{s^2+\omega^2}$
$e^{-at}\,\mathrm{Sin}(\omega t)$	$\frac{\omega}{(a+s)^2+\omega^2}$
$e^{-at}\,\mathrm{Cos}(\omega t)$	$\frac{a+s}{(a+s)^2+\omega^2}$

Another useful property is the *translation of functions*. Consider a translated function defined using the step function in the form

$$f(t-t_o)H(t-t_o) \tag{8.4}$$

The Laplace transform is

$$\mathcal{L}\left[f(t-t_o)H(t-t_o)\right] = \int_{t=0}^{\infty} f(t-t_o)H(t-t_o)e^{-st}dt = e^{-st_o}F(s) \tag{8.5}$$

This property is useful for forcing functions that are applied at times other than $t=0$.

Sometimes we have a function multiplied by e^{-at}. For this case the Laplace transform becomes

$$\mathcal{L}\left[e^{-at}f(t)\right] = \int_{t=0}^{\infty} e^{-at}f(t)e^{-st}dt = \int_{t=0}^{\infty} f(t)e^{-(s+a)t}dt = F(s+a) \tag{8.6}$$

An extremely important property for solving differential equations is the Laplace transform of derivatives. For first and second derivatives the relationships are

First Derivative

$$\mathcal{L}\left[\dot{f}(t)\right] = s\mathcal{L}\left[f(t)\right] - f(0) \tag{8.7}$$

$$= sF(s) - f(0)$$

Second Derivative

$$\mathcal{L}\left[\ddot{f}(t)\right] = s^2\mathcal{L}\left[f(t)\right] - sf(0) - f'(0) \tag{8.8}$$

$$= s^2F(s) - sf(0) - f'(0)$$

This is the important property of Laplace transforms, since it makes the solution of ordinary differential equations (ODEs) possible.

8.4 THE INVERSE LAPLACE TRANSFORMATION

The inverse Laplace transform is the process of finding the time function $f(t)$ from the corresponding transform $F(s)$. The methods of finding the inverse Laplace transform are to

- Use the basic definition, Equation 8.2. This is usually difficult and is seldom used.
- Use tables of functions $f(t)$ corresponding to given Laplace transforms $F(s)$. Some important cases are listed in Table 8.1.
- Use the partial-fraction expansion method. This method is emphasized in many books.
- Use Mathematica's "InverseLaplaceTransform" command.

8.4.1 Partial-Fraction Expansion Method

For problems involving dynamical systems, $F(s)$ frequently occurs in the form

$$F(s) = \frac{B(s)}{A(s)} \tag{8.9}$$

where $A(s)$ and $B(s)$ are polynomials in s. The degree of $B(s)$ is not higher than that of $A(s)$. The advantage of the partial-fraction approach is that the individual terms of $F(s)$ resulting from the expansion are simple functions of s that have well-known inversions and can be found in standard Laplace transform tables such as Table 8.1.

In order to apply the partial-fraction approach, the roots of the denominator polynomial must be determined. Then, Equation 8.9 for $F(s)$ can be written as

$$F(s) = \frac{B(s)}{(s+p_1)(s+p_2)\cdots(s+p_n)} \tag{8.10}$$

The parameters $p_1, p_2, \ldots, p_n$ are called *poles*. They may be real or complex. Complex poles always occur as a pair of complex conjugates, that is, $a + bj$ and $a - bj$.

8.4.2 Partial-Fraction Expansion for Distinct Poles

Cases with distinct poles always allow a simple expansion:

$$F(s) = \frac{B(s)}{A(s)} = \frac{a_1}{s+p_1} + \frac{a_2}{s+p_2} + \cdots + \frac{a_n}{s+p_n} \tag{8.11}$$

where the constants a_k, $k = 1, \ldots, n$ are called the *residues*. A typical residue a_k can be found by multiplying both sides by $s + p_k$ and evaluating the resulting expression at $s = -p_k$.

$$\begin{aligned}
&\left(\frac{B(s)}{A(s)}(s+p_k)\right)_{s=-p_k} \\
&= \left(\frac{a_1}{s+p_1}(s+p_k) + \cdots + \frac{a_k}{s+p_k}(s+p_k) + \cdots + \frac{a_n}{s+p_n}(s+p_k)\right)_{s=-p_k} \\
&= 0 + \cdots + a_k + \cdots + 0 = a_k
\end{aligned} \tag{8.12}$$

From Table 8.1 of Laplace transforms, we find

$$\mathcal{L}^{-1}\left(\frac{a_k}{s+p_k}\right) = a_k \mathcal{L}^{-1}\left(\frac{1}{s+p_k}\right) = a_k e^{-p_k t} \tag{8.13}$$

Now, we can invert the entire polynomial.

$$\begin{aligned}
f(t) &= \mathcal{L}^{-1}\left[F(s)\right] \\
&= \mathcal{L}^{-1}\left[\frac{a_1}{s+p_1} + \frac{a_2}{s+p_2} + \cdots + \frac{a_n}{s+p_n}\right] \\
&= \mathcal{L}^{-1}\left[\frac{a_1}{s+p_1}\right] + \mathcal{L}^{-1}\left[\frac{a_2}{s+p_2}\right] + \ldots + \mathcal{L}^{-1}\left[\frac{a_n}{s+p_n}\right]
\end{aligned} \tag{8.14}$$

$$f(t) = a_1 e^{-p_1 t} + a_2 e^{-p_2 t} + \ldots + a_n e^{-p_n t} \tag{8.15}$$

Note that when a quadratic factor in the denominator has a pair of complex roots, it is better not to factor the quadratic portion.

8.4.3 Partial-Fraction Expansion for Multiple Poles

Consider a denominator with multiple roots such as

$$F(s) = \frac{B(s)}{A(s)} = \frac{2s+3}{(s+1)^2} \tag{8.16}$$

The partial-fraction expansion involves two terms.

$$F(s) = \frac{B(s)}{A(s)} = \frac{2s+3}{(s+1)^2} = \frac{b_2}{(s+1)^2} + \frac{b_1}{s+1} \tag{8.17}$$

The coefficients b_1 and b_2 are determined by multiplying by $(s+1)^2$. Now, evaluate at $s=-1$.

$$\left((s+1)^2 \frac{B(s)}{A(s)}\right)_{s=-1} = \left(b_2 + b_1(s+1)\right)_{s=-1} \tag{8.18}$$

$$1 = b_2$$

Next, take the derivative of Equation 8.18.

$$\frac{d}{dt}\left((s+1)^2 \frac{B(s)}{A(s)}\right) = \frac{d}{dt}\left(b_2 + b_1(s+1)\right) \tag{8.19}$$

$$\frac{d}{dt}(2s+3) = \frac{d}{dt}\left(b_2 + b_1(s+1)\right)$$

$$2 = b_1$$

We thus have obtained the partial-fraction expansion.

$$F(s) = \frac{B(s)}{A(s)} = \frac{2s+3}{(s+1)^2} = \frac{1}{(s+1)^2} + \frac{2}{s+1} \tag{8.20}$$

Now, we can invert the entire polynomial.

$$f(t) = \mathcal{L}^{-1}\left[F(s)\right] = \mathcal{L}^{-1}\left[\frac{1}{(s+1)^2} + \frac{2}{s+1}\right] = \mathcal{L}^{-1}\left[\frac{1}{(s+1)^2}\right] + \mathcal{L}^{-1}\left[\frac{2}{s+1}\right] \tag{8.21}$$

$$f(t) = te^{-t} + 2e^{-t} = e^{-t}(t+2) \tag{8.22}$$

8.5 SOLUTIONS OF LINEAR ORDINARY DIFFERENTIAL EQUATION

8.5.1 General Strategy

Now, we can address the most important use of the Laplace transform method—solving linear differential equations. The basic objective is to find the solution to a differential equation with associated initial conditions. We would like to go directly from a mathematical model to the solution. However, it is often not clear how to take this giant step. Often, a

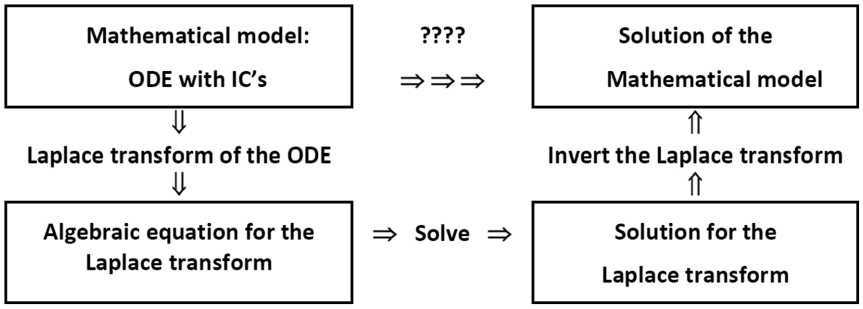


Figure 8.1 Steps in the Laplace transform method.

direct solution is quite difficult, as indicated in Figure 8.1. Thus, the Laplace transform method is used. It yields the complete solution (both homogeneous and particular) of linear, time-invariant differential equations. The steps involved in the Laplace transform method are outlined in Figure 8.1.

8.5.2 First-Order Ordinary Differential Equations

Mathematical Model

Consider the following classic first-order ODE and initial condition.

$$\begin{aligned} \dot{x} &= -ax \\ x(0) &= x_0 \end{aligned} \tag{8.23}$$

Note that the derivative has several commonly used designations:

$$\frac{dx}{dt} = \dot{x}(t) = x'(t) \tag{8.24}$$

This is a first-order ODE in time and thus requires one initial condition for a unique solution. The objective is to find and examine the solution, $x(t)$.

The first step in the solution process is to take the Laplace transform. Keep in mind the important properties of the derivative:

$$\begin{aligned} \mathcal{L}\left[x(t)\right] &= X(s) \\ \mathcal{L}\left[\dot{x}(t)\right] &= sX(s) - x_0 \end{aligned} \tag{8.25}$$

The Laplace transform of the differential equation is

$$\begin{aligned} \mathcal{L}\left[\dot{x} + ax\right] &= \mathcal{L}\left[0\right] \\ \mathcal{L}\left[\dot{x}\right] + \mathcal{L}\left[ax\right] &= 0 \\ \left(sX(s) - x_0\right) + aX(s) &= 0 \end{aligned} \tag{8.26}$$

Solving this algebraic equation for $X(s)$ gives

$$X(s) = x_0 \frac{1}{s+a} \tag{8.27}$$

The solution can now be determined by taking the inverse Laplace transform:

$$x(t) = \mathcal{L}^{-1}\left[X(s)\right] \tag{8.28}$$

This inverse can be obtained directly from the table of Laplace transform pairs in Table 8.1 or by using Mathematica's "InverseLaplaceTransform" command. Either way, the result is

$$x(t) = x_0 e^{-at} \tag{8.29}$$

This solution summary is shown in Figure 8.2, and the effects of parameters a and x_0 are displayed in Figure 8.3.

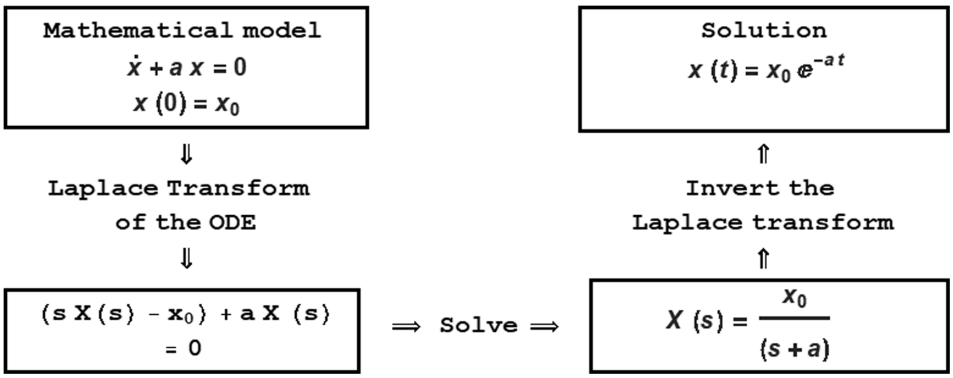


Figure 8.2 Solution summary for Equation 8.24.

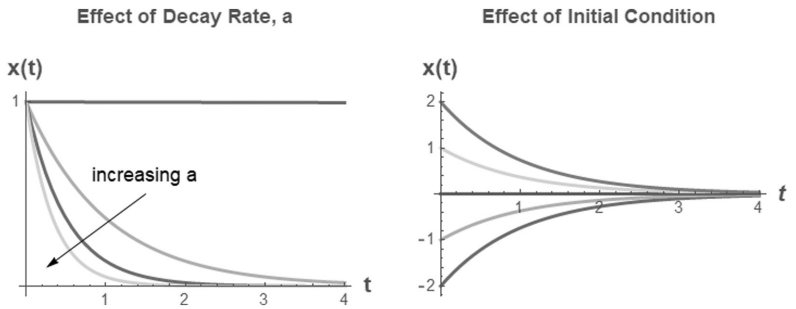


Figure 8.3 Classic behavior of a first-order ODE.

8.5.3 Second-Order Ordinary Differential Equations

A specific mathematical model is selected in order to demonstrate the procedure. The objective is to find the solution of the following system of equations:

$$\begin{aligned} &\ddot{x} + 3\dot{x} + 2x = 0 \\ &x(0) = x_0 \\ &\dot{x}(0) = v_0 \end{aligned} \tag{8.30}$$

This is a second-order ODE in time and thus requires two initial conditions for a unique solution.

The Laplace transform of the differential equation is

$$\mathcal{L}\left[\ddot{x} + 3\dot{x} + 2x\right] = \mathcal{L}[0] \tag{8.31}$$

$$\left(s^2X(s) - sx_0 - v_0\right) + 3\left(sX(s) - x_0\right) + 2X(s) = 0$$

Solve this algebraic equation for $X(s)$.

$$X(s) = \frac{sx_0 + v_0 + 3x_0}{s^2 + 3s + 2}$$

The solution is determined by taking the inverse Laplace transform:

$$x(t) = \mathcal{L}^{-1}\left[X(s)\right] \tag{8.32}$$

This inverse can be obtained by using the partial fraction expansion method (Section 8.5.2) or by using Mathematica's "InverseLaplaceTransform" command.

$$\begin{aligned} &x(t) = (2x_0 + v_0)e^{-t} - (x_0 + v_0)e^{-2t} \\ &\qquad \text{or} \\ &x(t) = \underbrace{x_0\left(2e^{-t} - e^{-2t}\right)}_{\text{due to } x_0} + \underbrace{v_0\left(e^{-t} - e^{-2t}\right)}_{\text{due to } v_0} \end{aligned} \tag{8.33}$$

Note that without Mathematica, the inverse Laplace transform is difficult and requires a lot of partial fractions and such. The solution is displayed in Figure 8.4.

8.6 THE TRANSFER FUNCTION

In system dynamics, transfer functions are used to characterize the input–output relationships of components or systems. They are useful for linear, time-invariant differential equations. The transfer function is the ratio of the Laplace transform of the output (response) to the Laplace transform of the input (driving function) for the case of zero initial conditions.

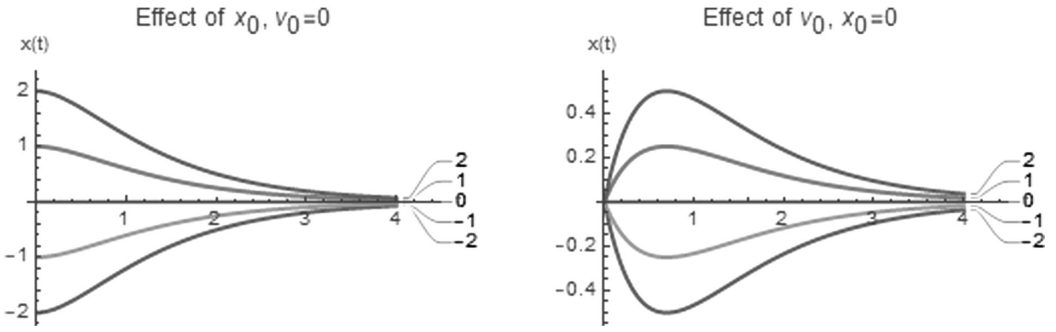


Figure 8.4 Solution of Equations 8.30.

Consider a general linear, time-invariant, nth-order differential equation where $y(t)$ is the output and $x(t)$ is the input.

$$a_0 \overset{(n)}{y} + a_1 \overset{(n-1)}{y} + \ldots + a_{n-1}\dot{y} + a_n y = b_0 \overset{(m)}{x} + b_1 \overset{(m-1)}{x} + \ldots + b_{m-1}\dot{x} + b_m x \tag{8.34}$$

The transfer function $G(s)$ for this system is

$$G(s) = \frac{\mathcal{L}[\text{output}]}{\mathcal{L}[\text{input}]} = \frac{Y(s)}{X(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \ldots + b_{m-1}s + b_m}{a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1}s + a_n} \tag{8.35}$$

The solution for the dynamic response is

$$y(t) = \mathcal{L}^{-1}\left[Y(s)\right] = \mathcal{L}^{-1}\left[G(s)X(s)\right] \tag{8.36}$$

Observe that this is simply the Laplace transform method cast in a different form. The advantage is that we can determine the transfer function once for a given system and use it to solve for and examine the solution for a variety of different special cases of the input.

8.6.1 The Impulse Response

The impulse response is the response to an impulse (Delta function) force. The impulse response is the most fundamental behavior that the system can exhibit. All other responses can be derived by adding (or integrating) the impulse response function.

Specifically, consider the situation where the input function is the impulse function (Delta function):

$$x(t) = \delta(t) \tag{8.37}$$

Take the Laplace transform:

$$X(s) = \mathcal{L}\left[x(t)\right] = \int_{t=0}^{\infty} \delta(t) e^{-st} dt = e^0 = 1 \tag{8.38}$$

The solution for the dynamic response, referred to as the *impulse response*, is

$$y(t) = \mathcal{L}^{-1}\left[Y(s)\right] = \mathcal{L}^{-1}\left[G(s)\cdot 1\right] = g(t) \tag{8.39}$$

8.6.2 First-Order Ordinary Differential Equations

Consider the first-order ODE with a forcing function, $x(t)$.

$$\dot{y} + a\cdot y = x(t) \tag{8.40}$$

$$y(0) = 0 \tag{8.41}$$

In order to determine the transfer function, take the Laplace transform of Equation 8.40 and use initial condition (Equation 8.41) to get

$$\mathcal{L}\left[\dot{y} + ay\right] = \mathcal{L}\left[x(t)\right]$$

$$sY(s) + aY(s) = X(s)$$

Solve this algebraic equation for the transfer function, $G(s)$.

$$G(s) = \frac{Y(s)}{X(s)} = \frac{1}{s+a} \tag{8.42}$$

The impulse response is the special case where $x(t) = \delta(t)$.

$$y_{\text{impulse}}(t) = \mathcal{L}^{-1}\left[Y(s)\right] = \mathcal{L}^{-1}\left[G(s)\cdot X(s)\right] = \mathcal{L}^{-1}\left[\frac{1}{s+a}\cdot 1\right] \tag{8.43}$$

The inversion gives the solution.

$$y_{\text{impulse}}(t) = e^{-at}H(t) \tag{8.44}$$

This solution is one of the all-time classics and is shown in Figure 8.5.

The impulse response is exactly the same solution as an initial condition with $x_0 = 1$, as shown in Figure 8.2. This special case is also the Green's function for this problem.

The step response is the solution to an applied step force, $x(t) = H(t)$.

$$y_{\text{step}}(t) = \mathcal{L}^{-1}\left[Y(s)\right] = \mathcal{L}^{-1}\left[G(s)\cdot X(s)\right] = \mathcal{L}^{-1}\left[\frac{1}{s+a}\cdot\frac{1}{s}\right] \tag{8.45}$$

Inverting this gives the solution as

$$y_{\text{step}}(t) = \frac{1}{a}\left(1 - e^{-at}\right)H(t) \tag{8.46}$$

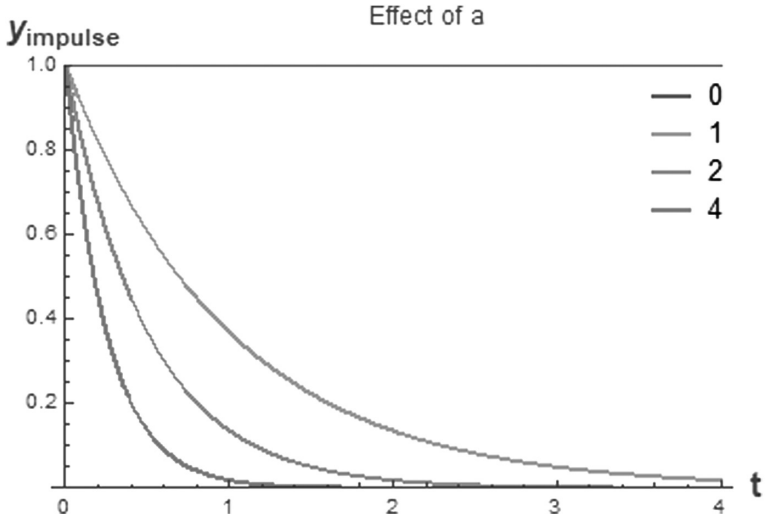


Figure 8.5 The impulse response.

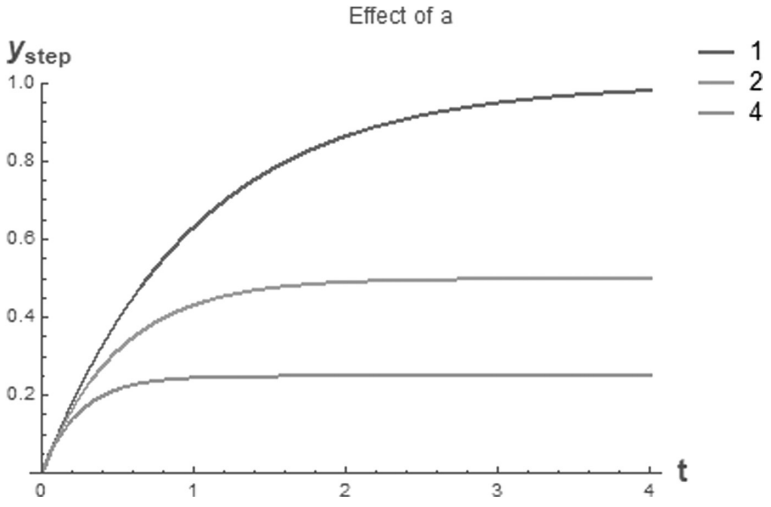


Figure 8.6 The unit step response.

This important case is shown in Figure 8.6 for various values of the parameter a.

The advantage of the transfer function is that the same transfer function can be used for any special case of the forcing function, as demonstrated in the previous two examples of an impulse and a step forcing function.

PROBLEMS

Problem 8.1

Consider the linear first-order ODE where τ is constant.

$$\frac{d\theta}{dt}+\frac{1}{\tau}\theta(t)=f(t)$$

$$\theta = \theta_0, \quad t = 0$$

1. Determine the analytical solution using the Laplace transform method when $f(t) = 0$. Create a meaningful plot of this solution.
2. Determine the transfer function $G(s)$.
3. Apply the results from part (2) to determine the solution for the following forcing functions with $\theta_0 = 0$. Plot each case.
   a. Step input, $f(t) = f_c H(t - t_0)$
   b. Pulsed input, $f(t) = f_c\left(H\left(t - t_1\right) - H\left(t - t_2\right)\right)$
   c. Impulse or Delta function, $f(t) = f_c\,\delta\left(t - t_1\right)$
   d. Harmonic forcing, $f(t) = f_c \sin(\omega t)$

Problem 8.2

Consider the linear second-order ODE where k is constant.

$$\ddot{x} + kx\left(t\right) = f\left(t\right)$$

$$x\left(0\right) = x_0$$

$$\dot{x}\left(0\right) = v_0$$

1. Determine the analytical solution using the Laplace transform method when $f(t) = 0$. Create a meaningful plot of this solution.
2. Determine the transfer function $G(s)$.
3. Apply the results from part (2) to determine the solution for the following forcing functions with zero initial conditions. Plot each case.
   a. Step input, $f(t) = f_c H(t - t_0)$
   b. Pulsed input, $f(t) = f_c\left(H\left(t - t_1\right) - H\left(t - t_2\right)\right)$
   c. Impulse or Delta function, $f(t) = f_c\,\delta\left(t - t_1\right)$
   d. Harmonic forcing, $f(t) = f_c \sin(\omega t)$

Chapter 9

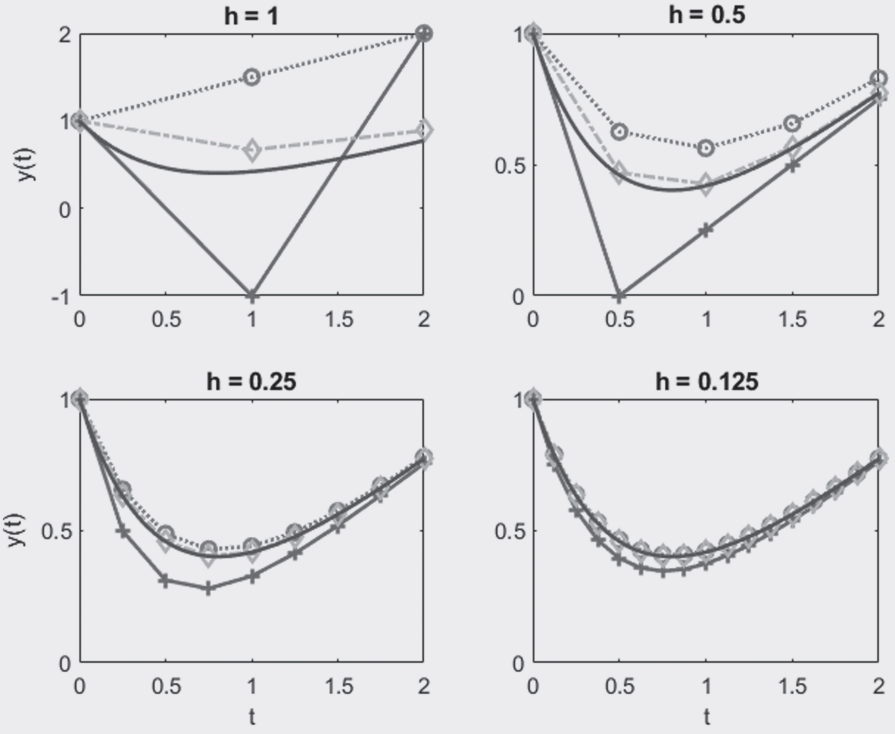
Numerical Solutions of Ordinary Differential Equations

CHAPTER OBJECTIVES

The objective of this chapter is to develop approximate numerical solutions for ordinary differential equations using the family of solutions known as Runge–Kutta methods.

Specific topics covered are

- Euler's method (first-order Runge–Kutta method)
- Heun's method (second-order Runge–Kutta method)
- General Runge–Kutta methods
- Coupled ordinary differential equations
- Second-order initial value problems
- Second-order boundary value problems
- Implicit methods



9.1 INTRODUCTION TO NUMERICAL SOLUTIONS

Consider a typical first-order ordinary differential equation (ODE) and initial condition in the form

$$\frac{dy}{dt} = f(t, y)$$
$$y = y_0, \quad t = 0 \tag{9.1}$$

As depicted in Figure 9.1 for the special case $dy/dt = -a \cdot y$, the solution can be derived analytically using methods of calculus or numerically using a computer.

The analytical method produces an exact mathematical formula for the solution $y(t)$ as a function of t. In contrast, the numerical method results in approximate values y_i of the solution only at discrete values t_i. The exact analytical solution in the form of a mathematical function for all t values is certainly more desirable than a sequence of approximate solution values at a limited number of t values. However, the exact solution is often difficult or impossible to find, and numerical solutions are the only alternative. Generally, exact solutions can only be found for certain linear differential equations.

In the numerical solution, the independent variable t and dependent variable y are discretized as shown in Figure 9.2. The first element has subscript 1 instead of 0, since matrices do not have a 0th element.

Starting from the known initial state $y = y_0$ at $t = 0$, a series of approximations for the solutions at ensuing times is developed in the form of algebraic expressions. A whole family of approximations, referred to as *Runge–Kutta methods*, can be developed. The simplest method is called the *Euler method*.

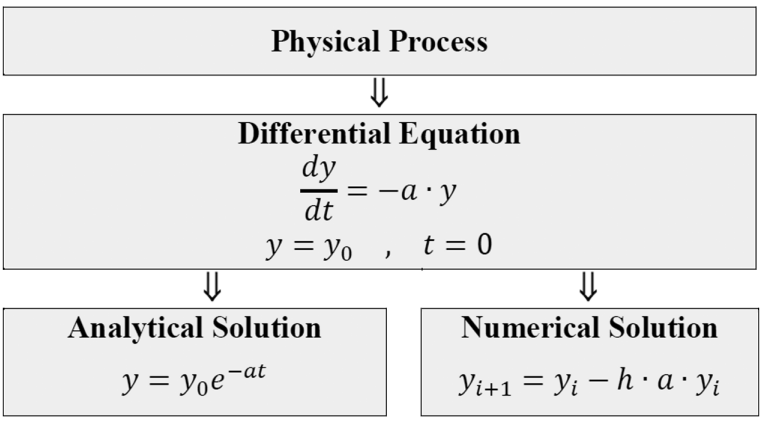


Figure 9.1 Analytical versus numerical solution of ordinary differential equations.

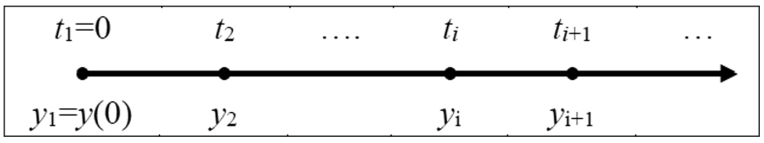


Figure 9.2 Time discretization.

9.2 RUNGE–KUTTA METHODS

9.2.1 Euler's Method

Euler's method, depicted in Figure 9.3, is the simplest but lowest order of the Runge–Kutta methods. The first derivative provides a direct estimate of the slope.

Using a forward finite difference scheme to approximate Equation 9.1 at t_i gives

$$\left(\frac{dy}{dt}\right)_i \cong \frac{y_{i+1}-y_i}{t_{i+1}-t_i} = \frac{y_{i+1}-y_i}{h} = f\left(t_i, y_i\right) \tag{9.2}$$

where the time step is $h = t_{i+1} - t_i = \Delta t$. Solve Equation 9.2 for y_{i+1} to get

$$y_{i+1} = y_i + h \cdot f\left(t_i, y_i\right) \tag{9.3}$$

This scheme, known as *Euler's method*, can be rewritten as

$$\begin{aligned} k_1 &= f\left(t_i, y_i\right) \\ y_{i+1} &= y_i + h \cdot k_1 \end{aligned} \tag{9.4}$$

Starting from the initial condition, y_0, a time marching scheme can be executed by the recursive use of Equation 9.3 to compute approximate solutions at times t_2, t_3, t_4, … . The solution is computed only at the discrete values t_i.

Accuracy is of order h, designated as O(h). That is, when the step size h is cut in half, we expect approximately half the numerical error.

Although the Euler scheme is simple and easy to program, numerical stability can be a problem. On the other hand, the implicit schemes presented in Section 9.5 are unconditionally stable but are generally more difficult to use.

9.2.2 Heun's Method

A fundamental problem with Euler's method is that the derivative at the beginning of an interval is assumed to apply across the entire interval. One possible improvement of Euler's method is called *Heun's method*. It involves the determination of two derivatives for the

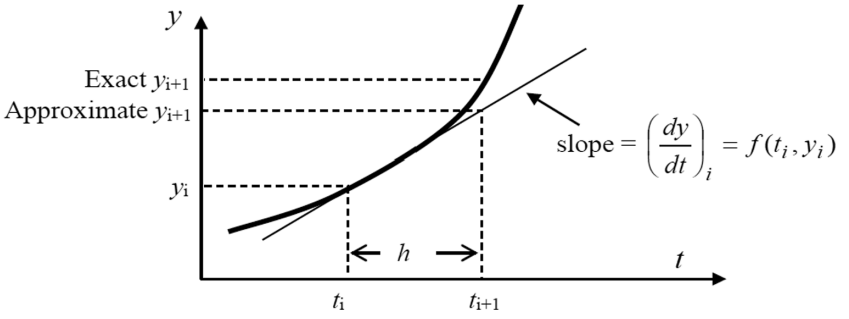


Figure 9.3 Euler's method.

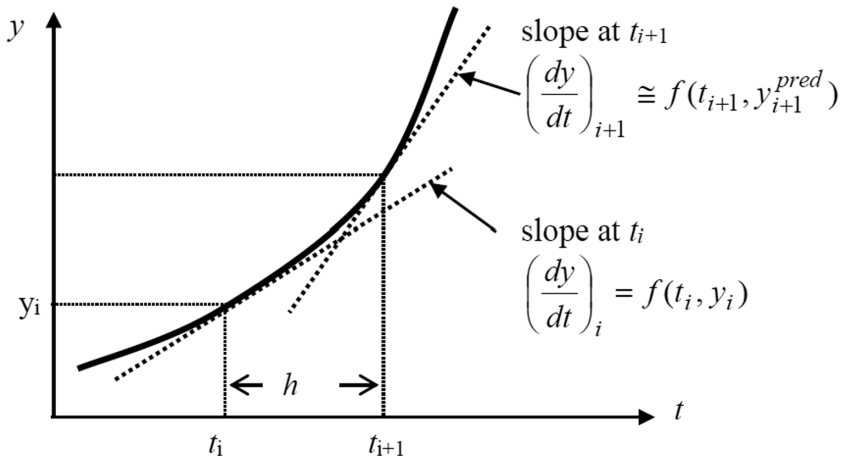


Figure 9.4 Heun's method.

interval—one at the beginning and one at the end, as shown in Figure 9.4. The two derivatives are then averaged.

The slope in the interval t_i to $t_{i+1} = t_i + h$ is estimated as the average of the slopes at t_i and t_{i+1} as

$$\left(\frac{dy}{dt}\right)_{\text{ave}} \cong \frac{\left(\frac{dy}{dt}\right)_i + \left(\frac{dy}{dt}\right)_{i+1}}{2} = \frac{f\left(t_i, y_i\right) + f\left(t_{i+1}, y_{i+1}^{\text{pred}}\right)}{2} \tag{9.5}$$

Since we do not know the slope y_{i+1} at t_{i+1}, the Euler method at t_i is used as a *predictor* or estimated value for y_{i+1}.

$$\text{Predictor: } y_{i+1} = y_i + h \cdot f\left(t_i, y_i\right) \tag{9.6}$$

The differential Equation 9.1 is discretized in the interval t_i to t_{i+1} using the following average derivative estimate:

$$\left(\frac{dy}{dt}\right)_{\text{ave}} \cong \frac{y_{i+1} - y_i}{h} = \frac{f\left(t_i, y_i\right) + f\left(t_{i+1}, y_{i+1}^{\text{pred}}\right)}{2} \tag{9.7}$$

Solving Equation 9.7 for y_{i+1} provides an improved estimate or *corrector*.

$$\text{Corrector: } y_{i+1} = y_i + \frac{h}{2}\left(f\left(t_i, y_i\right) + f\left(t_{i+1}, y_{i+1}^{\text{pred}}\right)\right) \tag{9.8}$$

This method is one of the *predictor–corrector* schemes. The numerical solution must be executed as a two-step process whereby first the *predictor* and then the *corrector* are computed. Since y_{i+1} is on both the left- and the right-hand side of the corrector formula, it provides a recursive formula for iteration.

We can rewrite Equations 9.6 and 9.8 in the Runge–Kutta format as

$$\begin{aligned} k_1 &= f\left(t_i, y_i\right) \\ k_2 &= f\left(t_i + h, y_i + hk_1\right) \\ y_{i+1} &= y_i + h\left(\frac{k_1 + k_2}{2}\right) \end{aligned} \tag{9.9}$$

In this formulation,

- k_1 is the slope at t_i
- k_2 is the slope at t_{i+1} estimated using k_1
- y_{i+1} is computed by averaging these two slopes

9.2.3 Higher-Order Runge–Kutta Methods

The Euler and Heun algorithms are actually special cases of a more general class of solutions called *Runge–Kutta methods*. They are derived using a combination of Taylor series and intelligent choices. The general form is

$$y_{i+1} = y_i + h \cdot \phi(t_i, y_i, h) \tag{9.10}$$

where $\phi(t_i, y_i, h)$ is called the *increment function*, which can be interpreted as a representative slope over the interval. The increment function has the general form

$$\phi(t_i, y_i, h) = a_1 k_1 + a_2 k_2 + \cdots + a_n k_n \tag{9.11}$$

where
n is the order of the method
the a_is are constants
the k_is are

$$\begin{aligned} k_1 &= f(t_i, y_i) \\ k_2 &= f(t_i + h \cdot p_1, y_i + h \cdot q_{11} k_1) \\ k_3 &= f(t_i + h \cdot p_2, y_i + h \cdot q_{21} k_1 + h \cdot q_{22} k_2) \\ &\cdots \end{aligned} \tag{9.12}$$

Note that the k_is are recursive. That is, k_1 appears in k_2, which appears in k_3, and so on. Figure 9.5 summarizes the Runge–Kutta schemes up to the fourth order. The order of the method corresponds to the approximate global truncation error. Thus, decreasing h by a factor of 2 in the Euler method would result in approximately 1/2 the error. Similarly, decreasing h by a factor of 2 in the fourth-order scheme would result in approximately $1/2^4 = 1/16$ the error. The fourth-order scheme can be shown to be the most efficient in terms of accuracy per arithmetic operation. Thus, the fourth-order scheme is often used in practice.

9.2.4 Numerical Comparison of Runge–Kutta Schemes

Consider the differential equation and initial condition

$$\begin{aligned} \frac{dy}{dt} &= t - 2y \\ y &= 1, \quad t = 0 \end{aligned} \tag{9.13}$$

Euler's method given by Equation 9.3 applied to this special case gives

$$y_{i+1} = y_i + h(t - 2y_i) \tag{9.14}$$

Method	Discretizion of $\frac{dy}{dt} = f(t, y)$	k_i's
1st Order: Euler Method	$y_{i+1} = y_i + h \cdot k_1$	$k_1 = f(t_i, y_i)$
2nd Order: Heun Method	$y_{i+1} = y_i + h\frac{k_1 + k_2}{2}$	$k_1 = f(t_i, y_i)$ $k_2 = f(t_i + h, y_i + h \cdot k_1)$
2nd Order: Midpoint Method	$y_{i+1} = y_i + h \cdot k_2$	$k_1 = f(t_i, y_i)$ $k_2 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}h \cdot k_1)$
2nd Order: Ralston's Method	$y_{i+1} = y_i + h\left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right)$	$k_1 = f(t_i, y_i)$ $k_2 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}h \cdot k_1)$
3rd Order	$y_{i+1} = y_i + \frac{h}{6}(k_1 + 4k_2 + k_3)$	$k_1 = f(t_i, y_i)$ $k_2 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}h \cdot k_1)$ $k_3 = f(t_i + h, y_i - h \cdot k_1 + 2h \cdot k_2)$
4th Order	$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$	$k_1 = f(t_i, y_i)$ $k_2 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}h \cdot k_1)$ $k_3 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}h \cdot k_2)$ $k_4 = f(t_i + h, y_i + h \cdot k_3)$

Figure 9.5 Summary of Runge–Kutta methods.

This example is selected to test accuracy, since the exact analytical solution can be derived using an integrating factor, as described in Section 10.3.3. The solution is

$$y(t) = \left(2t - 1 + 5e^{-2t}\right)/4 \tag{9.15}$$

Figure 9.6 shows the solution using Euler's method, Heun's method, the fourth-order Runge–Kutta, and the exact solution for several step sizes. The higher-order methods clearly produce a more accurate solution. Also, a smaller step size produces more accurate results. Of course, higher-order methods and smaller step sizes require greater computation times.

9.3 COUPLED SYSTEMS OF FIRST-ORDER DIFFERENTIAL EQUATIONS

With small modifications, the same algorithms used for single first-order ODEs work for coupled systems. Consider two coupled first-order ODEs:

$$\begin{aligned} \frac{dy_1}{dt} &= f_1(t, y_1, y_2) \\ \frac{dy_2}{dt} &= f_2(t, y_1, y_2) \end{aligned} \tag{9.16}$$

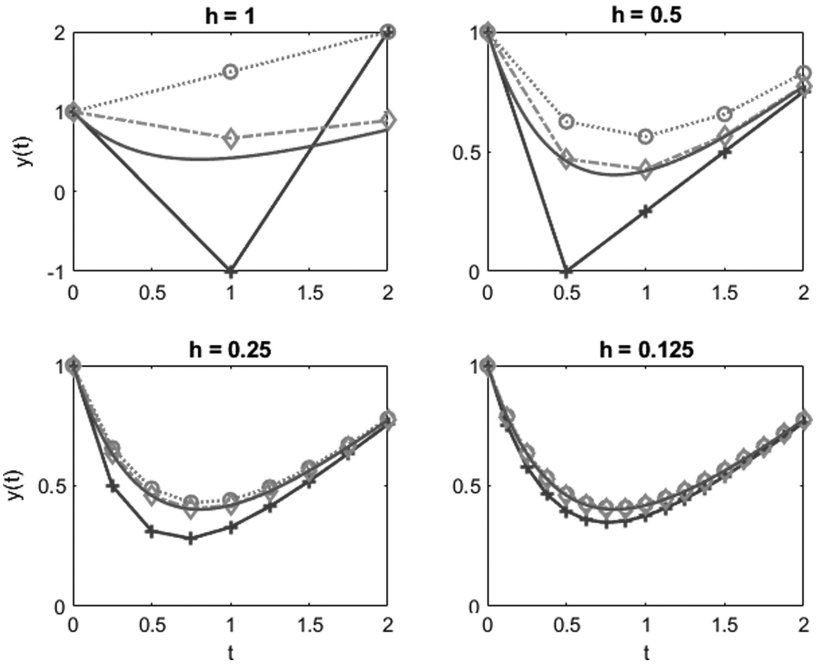


Figure 9.6 Comparison of the exact solution (solid line) with Euler (+), Heun (o), and fourth-order Runge–Kutta (◊) methods for various step sizes.

Since we now have two dependent variables, initial or starting conditions are required for both variables.

$$\begin{aligned} y_1 &= y_{1,0}, \quad t = 0 \\ y_2 &= y_{2,0}, \quad t = 0 \end{aligned} \tag{9.17}$$

A direct extension of Euler's method for a single ODE, given by Equation 9.3, leads to the following discretized solution of Equations 9.16:

$$\begin{aligned} y_{1,i+1} &= y_{1,i} + h \cdot f_1\left(t_i, y_{1,i}, y_{2,i}\right) \\ y_{2,i+1} &= y_{2,i} + h \cdot f_2\left(t_i, y_{1,i}, y_{2,i}\right) \end{aligned} \tag{9.18}$$

All the Runge–Kutta schemes summarized in Figure 9.5 can be extended to coupled systems of first-order equations. Any number of simultaneous equations can be solved in this manner.

9.4 SECOND-ORDER INITIAL VALUE PROBLEMS

Second-order equations can be solved by decomposing them into a system of two, coupled, first-order equations. For instance, consider the classical mass-spring-damper system

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = F\left(t\right) \tag{9.19}$$

This equation is often expressed as

$$\frac{d^2x}{dt^2} + 2\zeta\omega_n \frac{dx}{dt} + \omega_n^2 \cdot x = \frac{F(t)}{m} \tag{9.20}$$

$$\omega = \sqrt{k/m} = \text{natural frequency}$$

$$\zeta = \frac{c}{2\sqrt{km}} = \text{damping ratio}$$

Using the definition of velocity, $dx/dt = v$, the original second-order ODE can be written as an equivalent set of coupled first-order equations in the form

$$\frac{dx}{dt} = v \tag{9.21}$$

$$\frac{dv}{dt} = \left(-cv - kx + F(t)\right)/m \tag{9.22}$$

In matrix form, the previous equations can be written as

$$\frac{d}{dt}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} v \\ \left(-cv - kx + F(t)\right)/m \end{bmatrix} \tag{9.23}$$

The two initial conditions required to complete the mathematical model are

$$x = x_0, \quad \frac{dx}{dt} = v_0, \quad t = 0 \tag{9.24}$$

Numerical results for some important cases are shown in Figures 9.7 and 9.8.

9.5 IMPLICIT SCHEMES

The Euler scheme and other versions of the Runge–Kutta method are plagued by stability problems—that is, for a time step that is too large, nonphysical oscillations occur in the solutions. The implicit method is often used to avoid these problems.

The basic idea is to estimate the derivative at the new time level. Our first-order differential equation is

$$\frac{dy}{dt} = f(t, y) \tag{9.25}$$

Discretize using a *backwards difference* scheme. That is, evaluate the function $f(t,y)$ at the new time level $i+1$.

$$\left(\frac{dy}{dt}\right)_{i+1} \cong \frac{y_{i+1} - y_i}{h} = f\left(t_{i+1}, y_{i+1}\right) \tag{9.26}$$

Solve for the current value y_{i+1}.

$$y_{i+1} = y_i + h \cdot f\left(t_{i+1}, y_{i+1}\right) \tag{9.27}$$

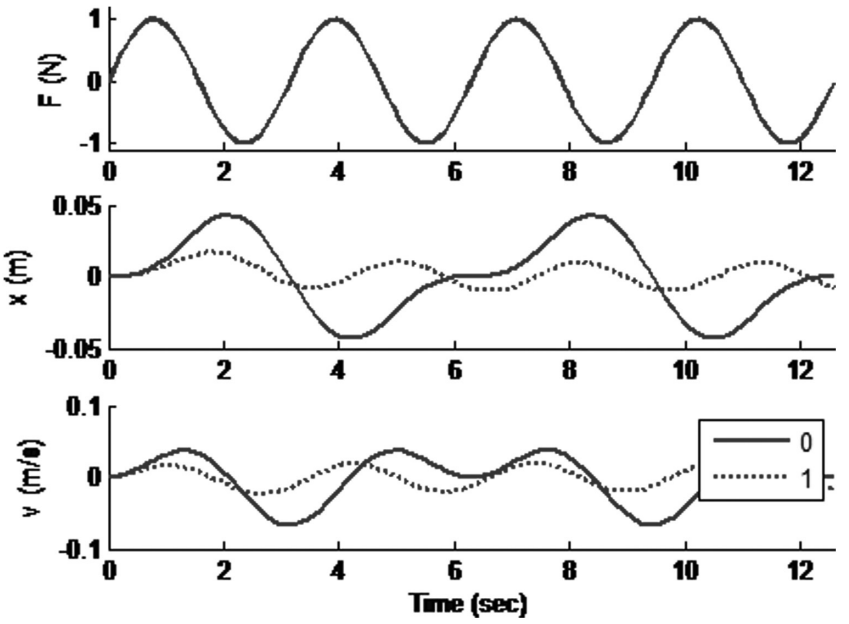


Figure 9.7 Solutions of the classical spring-mass system with harmonic forcing for $\zeta = 0$ (zero damping) and $\zeta = 1$ (critically damped), $m = 20$ kg, $k = 20$ N/m, and $x_0 = v_0 = 0$.

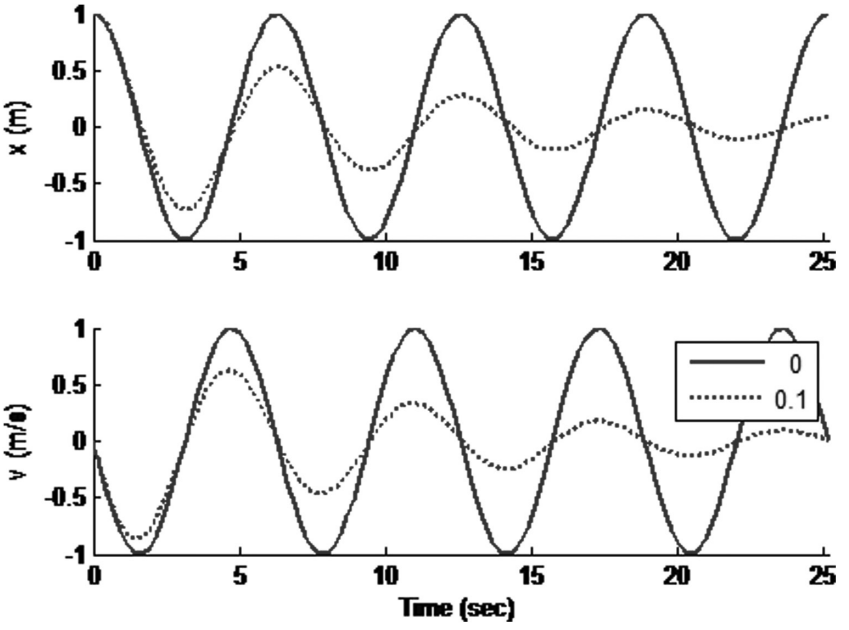


Figure 9.8 Solutions of the classical spring-mass system with initial displacement $x_0 = 1$, zero forcing, $\zeta = 0$ (zero damping) and $\zeta = 0.1$ (underdamped), $m = 20$ kg, $k = 20$ N/m, and $v_0 = 0$.

The solution at the new time level $i+1$ is not an *explicit* function of the solution at the old time level i; thus, we call this the *implicit* method. If the function $f(t,y)$ is linear in y, we can solve directly for y_{i+1}; otherwise, iteration is required.

As an example, consider the classical linear first-order system

$$\frac{dy}{dt} = -a \cdot y \tag{9.28}$$

Discretize this equation in a fully implicit fashion.

$$y_{i+1} = y_i - h \cdot a \cdot y_{i+1} \tag{9.29}$$

Solve for y_{i+1}.

$$y_{i+1} = y_i \left(\frac{1}{1 + h \cdot a} \right) \tag{9.30}$$

This numerical solution scheme is unconditionally stable.

9.6 SECOND-ORDER BOUNDARY VALUE PROBLEMS: THE SHOOTING METHOD

The idea behind the shooting method is to convert a boundary value problem into an initial value problem. An iterative scheme using standard initial value solvers, such as the Runge–Kutta method, is then used to obtain a solution.

To illustrate the method, the fin equation from heat transfer with specified boundary temperatures is considered.

$$\frac{d^2\theta}{dx^2} - m^2\theta = 0 \tag{9.31}$$

$$\begin{aligned} \theta &= \theta_0, \quad x = 0 \\ \theta &= \theta_L, \quad x = L \end{aligned} \tag{9.32}$$

The variable $\theta(x)$ is the temperature rise above the ambient temperature. This boundary value problem is converted to an initial value problem by defining the derivative as a new variable.

$$y_1 = \theta \tag{9.33}$$

$$\frac{dy_1}{dx} = y_2 \tag{9.34}$$

Substituting the variables defined by Equations 9.33 and 9.34 into Equation 9.31 transforms the original second-order ODE into a first-order equation. The resulting equation is

$$\frac{dy_2}{dx} = m^2 y_1 \tag{9.35}$$

Equations 9.34 and 9.35 form a system of first-order ODEs that can be solved for $y_1(x)$ and $y_2(x)$. If we knew the values of both y_1 and y_2 at $x = 0$, the problem could be solved as

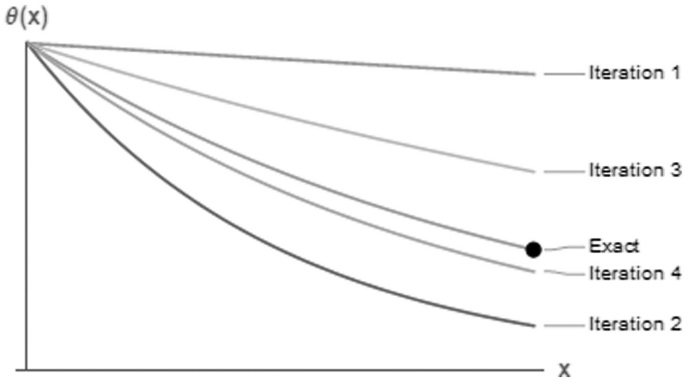


Figure 9.9 The shooting method.

an initial value problem by the techniques described in previous chapters. However, we only have the value for one of the variables at $x = 0$. Thus, we guess the other and perform the computations. This guess is iteratively improved until the boundary condition at $x = L$ is satisfied.

A typical example is shown in Figure 9.9. The terminology *shooting method* is now clear. Just as you would adjust a cannon to hit a target, you adjust the initial guess to hit the boundary condition at $x = L$. This technique can be applied to any linear or nonlinear differential equation and to any combination of the boundary conditions.

PROBLEMS

Problem 9.1: Runge–Kutta Basics

Solve the following initial value problem over the interval from $t = 0$ to 2 with $a = 1.1$ and $y_0 = 1$. Display all your results on a single graph.

$$\frac{dy}{dt} = y \cdot t^2 - a \cdot y$$

a) Analytically.
b) Using Euler's method with $h = 0.5$ and 0.25.
c) Using Heun's method with $h = 0.5$.

Problem 9.2: Euler's Method

Consider the differential equation

$$\frac{dy}{dt} = -y + 2$$

a) Sketch the phase diagram (dy/dt vs. y) and the anticipated solution ($y(t)$ vs. t).
b) Starting with the initial condition $y_0 = 0$, manually perform three steps of Euler's method using a step size of $h = 0.5$.
c) Using Euler's method, compute and plot the solution starting from initial conditions $y_0 = -2$, 0, 2, 4, and 6. Put all the curves on a single graph. Use a time range $0 \le t \le 5$. Experiment with the required step size in order to get accurate and smooth-looking plots. Report the step size you decided upon.

Problem 9.3: Lumped Thermal Mass

Consider the cooling of a hot object initially at temperature T_0 in cold air at T_∞ with heat transfer coefficient h. Assume that the lumped thermal capacity model, as described in the previous chapter, is valid. The differential equation is

$$\rho c V \frac{dT}{dt} = -hA_s\left(T - T_\infty\right) + gV$$

a) Draw the cause–effect diagram for this physical problem.
b) Write a function to numerically solve for temperature as a function of time. The system parameters and forcing functions should be inputs, and the time and temperature vectors should be outputs.
c) Explore the effect of the heat transfer coefficient. Plot temperature versus time for $h = 10$ W/m^2 °C (still day), 25 W/m^2 °C (typical day), and 75 W/m^2 °C (hurricane). Put all curves on a single graph. Carry out time long enough that the temperature begins to level out at steady state. The other parameters are $V = 10^{-6}$ m^3, $A_s = 10^{-4}$ m^2, $\rho = 1000$ kg/m^3, $c = 500$ J/kg °C, $T_\infty = 25$ °C, $T_0 = 400$ °C, $g = 0$.
d) Explore the effect of the heat source strength. Plot the solution for $g = 0$, 10^5, 2×10^5, and 3×10^5 W/m^3. Put all curves on a single graph. Use parameter values from part (b) with $h = 25$ W/m^2 °C.

Problem 9.4: Population Model

The earth's population can be estimated using the following simplified model:

$$\frac{dP}{dt} = aP - bP^2$$

where a is the birth rate parameter and b is the death rate parameter.

a) Determine the steady-state population.
b) Sketch the phase plot, dP/dt versus P. Using the phase plot, sketch the anticipated population history, $P(t)$ versus time. Draw a single graph with several curves corresponding to different starting populations.
c) When the time variable t is measured in years, experimental evidence suggests that the parameters a and b in the population model are approximately $a = 0.028$ 1/year and $b = 2.9 \times 10^{-12}$ 1/(people*year). Starting from a population of approximately $P_0 = 100$ million people in the year 1800, compute and plot the earth's population as a function of time in years. Run your simulation until the population appears to level off. Using these parameters, how many people will we eventually have on this earth? At what year will we have reached 99% of our maximum population?
d) Estimating population is an inexact science. Examine the effect of the birth rate parameter by plotting the earth's population for $a = 0.025$, 0.03, and 0.035 with $b = 2.9 \times 10^{-12}$. Put all curves on a single graph.
e) Examine the effect of the death rate parameter by plotting the earth's population for $b = 2.5 \times 10^{-12}$, 3×10^{-12}, and 3.5×10^{-12} with $a = 0.029$. Put all curves on a single graph.

Problem 9.5: Linear Oscillator, Effect of Damping

Consider the spring-mass-damper system with zero forcing, described by the following differential equation and initial conditions:

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = 0$$

$$\left.\begin{aligned} x &= x_0 \\ \frac{dx}{dt} &= v_0 \end{aligned}\right\} \quad t = 0$$

a) Draw the cause–effect diagram for this system.
b) Write a function to numerically solve this differential equation.
c) Use your function to solve and plot $x(t)$ and $v(t)$ versus t. Use the following parameters: $x_0 = 1$ m, $v_0 = 0$, $m = 20$ kg, $k = 20$ N/m, $c = 0$, 8, and 40 N·s/m. Create separate plots for $x(t)$ and $v(t)$ versus t. Put all three curves for different c values on a single graph (see Figure 9.7 for an example).

Problem 9.6: Linear Oscillator with Forcing

Consider the forced spring-mass-damper system, described by the following differential equation and initial conditions:

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = F(t)$$

$$x = x_0,\ v = v_0, \quad \text{when } t = 0$$

a) Draw the cause–effect diagram for this system.
b) Write a function to numerically solve this differential equation.
c) Use your function to compute the displacement and velocity for the following special cases. For each case, create a three-part graph containing $F(t)$, $x(t)$, and $v(t)$ versus t. For all cases, $m = 20$ kg, $c = 4$ N·s/m, $k = 20$ N/m, and $x_0 = v_0 = 0$. See Figure 9.6 for an example.

	Forcing: F(t) (N)
Constant forcing	$F(t) = F_0$ $F_0 = 10$ N
Harmonic forcing	$F(t) = F_0 \sin(\omega t)$
	$F_0 = 5$ N/kg, $\omega = 1$ rad/s
Pulsed forcing	$F(t) = F_0\left(H(t) - H(t - t_{on})\right)$
	$F_0 = 10$ N, $t_{on} = 40$ s

Problem 9.7: Pendulum

The oscillations of an undamped pendulum can be simulated with the following nonlinear equation:

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\sin(\theta) = 0$$

where

θ = angle of displacement
g = gravitational constant = 9.81 m²/s
L = pendulum length

For small angular displacements, $\sin(\theta) \approx \theta$, and the model can be linearized as

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\theta = 0$$

a) Reformulate these equations as two first-order equations for θ and $v = d\theta/dt$.
b) Draw the cause–effect diagram.
c) Create a function to compute the solution as a function of the input parameters.
d) Using the function you developed, solve for θ and v as functions of time for both the linear and nonlinear models with $L = 0.6$ m and initial conditions $\theta_0 = \pi/8$ and $v_0 = 0$. Plot the linear and nonlinear solutions on the same graph. Also, plot results for $\theta_0 = \pi/2$, $\theta_0 = 0.99\pi$, and $\theta_0 = \pi$.

Problem 9.8: Van der Pol Equation

The van der Pol equation is a model of a nonlinear circuit that arose back in the days of vacuum tubes:

$$\frac{d^2y}{dt^2} + \mu\left(1 - y^2\right)\frac{dy}{dt} + y = 0$$

a) Reformulate these equations as two first-order equations for $y_1 = y$ and $y_2 = dy/dt$.
b) Create a function to compute the solution as a function of the input parameters.
c) Using your function from part (b), solve for y_1 and y_2 as functions of time with $\mu = 1$ and initial conditions $y_1 = 0.1$ and $y_2 = 0$. Create the following graphs:
   - y_1 versus t and y_2 versus t
   - phase plot: y_1 versus y_2
d) Repeat part (c) with $\mu = 0$ and 10.

Problem 9.9: Sky Diver

The free-fall velocity of a parachutist can be estimated using the force balance

$$\frac{dv}{dt} = g - \frac{c_d}{m}v^2$$

where

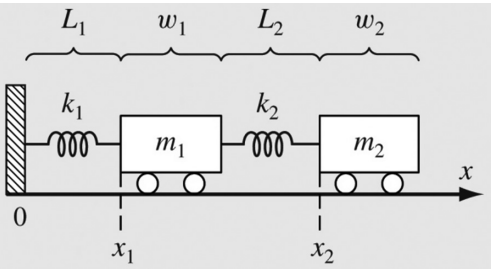
v = velocity (m/s)
t = time (s)
g = 9.81 m/s^2 = acceleration due to gravity
c_d = drag coefficient (kg/m)
m = mass (kg)

For an 80 kg parachutist, we wish to solve this differential equation given that $v = 0$ at $t = 0$. During free fall, $c_d = 0.25$ kg/m. However, at $t = t_{\text{open}}$, the chute opens, whereupon $c_d = 5$ kg/m.

a) What would be the steady-state velocity if the chute was not opened? What is the steady-state velocity after the chute opens?
b) Sketch the phase diagram (dv/dt versus v) and the anticipated solution for $v(t)$ versus t based on the phase diagram.
c) Write a function to solve this differential equation numerically from $t = 0$ to 30 s. Create a plot of $v(t)$ versus t for $t_{open} = 0$, 5, 10, and 20 s, putting all curves on a single plot.

Problem 9.10: Coupled Oscillators

Two masses are attached to a wall by linear springs.



Force balances based on Newton's second law can be written as

$$m_1 \frac{d^2x_1}{dt^2} = -k_1\left(x_1 - L_1\right) + k_2\left(x_2 - x_1 - w_1 - L_2\right)$$

$$m_2 \frac{d^2x_2}{dt^2} = -k_2\left(x_2 - x_1 - w_1 - L_2\right)$$

where

k_i = the spring constants (N/m)
m_i = masses (kg)
L_i = the length of the unstretched springs (m)
w_i = the widths of the masses (m)

Initially, the masses are at locations $x_{1,0}$ and $x_{2,0}$.

a) Draw the cause–effect diagram for this system.
b) Determine the steady-state or fixed points.
c) Write a function to numerically solve this system of differential equations.
d) Compute the positions of the masses as a function of time over the range $t = 0$ to 20 using the following parameters: $k_1 = k_2 = 5$, $m_1 = m_2 = 2$, $w_1 = w_2 = 5$, and $L_1 = L_2 = 2$. Set the initial conditions as $x_{1,0} = L_1$ and $x_{2,0} = L_1 + w_1 + L_2 + 6$ with zero velocity. Construct time-series plots of both the displacements and the velocities. Indicate the fixed points on your graph. In addition, produce a phase-plane plot of x_1 versus x_2.

Problem 9.11: An Epidemic

The following ODEs have been proposed as a model of an epidemic:

$$\frac{dS}{dt} = -aSI$$

$$\frac{dI}{dt} = aSI - rI$$

$$\frac{dR}{dt} = rI$$

where

S = the number of susceptible individuals
I = the number of infected individuals
R = the number of recovered individuals
a = infection rate
r = recovery rate

A city initially has 10,000 people, all of whom are susceptible. Then, a single infectious individual enters the city at $t = 0$. Use the following estimates for the parameters: $a = 0.002$/(person week) and $r = 0.15$/day.

a) Compute the progression of the epidemic. At what time does the number of infected individuals drop back to 10? Create time-series plots of all the state variables over a time range from 0 until the number of infected individuals falls below 10. Also, create a three-dimensional phase plot of S versus I versus R.
b) Suppose that after recovery, there is a loss of immunity that causes recovered individuals to become susceptible. This reinfection mechanism can be modeled as ρR, where

ρ = the reinfection rate = 0.03/day. Modify the model to include this mechanism and repeat the computations in (a).

c) Suppose that some of the infected people recover and some die. F is the number of dead people. Also consider that there is an influx of susceptible individuals moving to the city at the rate of Q people/day. Modify the model to include these possibilities and repeat the computations in (a).

Chapter 10

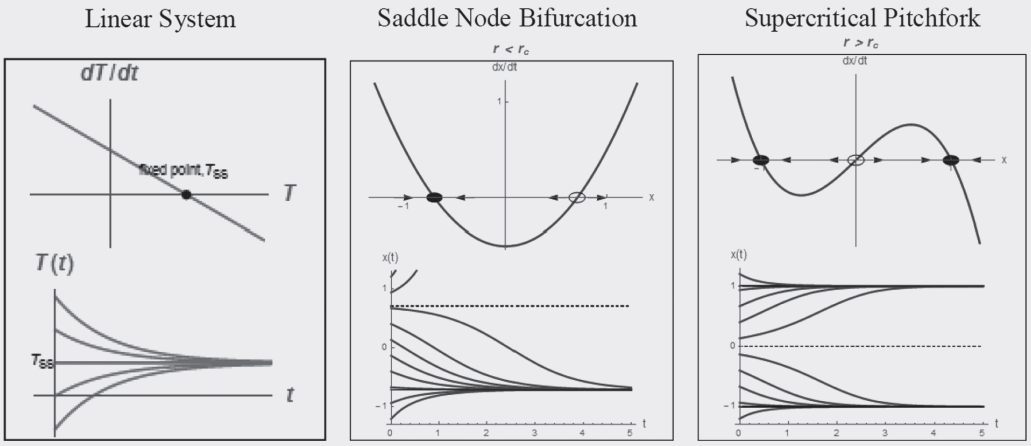
First-Order Ordinary Differential Equations

CHAPTER OBJECTIVES

The primary objective of this chapter is to analyze the behavior of first-order ordinary differential equations.

Specific objectives and topics covered are

- Stability analysis of the fixed points
- Characteristics of linear systems
- Linear first-order systems using proportionality and superposition
- Integrating factors for linear equations
- Nonlinear first-order systems
- Saddle-node, transcritical, supercritical pitchfork, and subcritical pitchfork bifurcations



10.1 STABILITY OF THE FIXED POINTS

In Section 7.2, we deduced the stability of the fixed points using graphical methods. How about a quantitative measure of stability? We can gain this sort of information by linearizing a small perturbation about a fixed point. This is the mathematical equivalent of nudging a physical system slightly away from a steady or fixed point and observing the resulting behavior.

Consider the first-order autonomous differential equation

$$\frac{d\theta}{dt} = f(\theta) \tag{10.1}$$

The fixed points are the state where the system comes to equilibrium and are defined by $f(\theta^*) = 0$. Consider a small perturbation away from θ^* defined as

$$\eta = \theta(t) - \theta^* \tag{10.2}$$

To see whether the perturbation grows or decays, we derive a differential equation for the perturbation η. The derivative of Equation 10.2 is

$$\frac{d\eta}{dt} = \underbrace{\frac{d}{dt}(\theta - \theta^*)}_{\theta^* \text{ is a constant}} = \frac{d\theta}{dt} = f(\theta) = f(\theta^* + \eta) \tag{10.3}$$

Now use a Taylor series expansion, retaining only the first-order term.

$$f(\theta^* + \eta) \cong \underbrace{f(\theta^*)}_{=0} + \eta \frac{df(\theta^*)}{d\theta} = \eta \frac{df(\theta^*)}{d\theta} \tag{10.4}$$

Using Equation 10.4 in Equation 10.3 gives

$$\frac{d\eta}{dt} = \eta \frac{df(\theta^*)}{d\theta} \tag{10.5}$$

This is a linear ordinary differential equation (ODE) for η. The solution subject to a starting perturbation of η_0 at $t = 0$ is

$$\eta(t) = \eta_0 \cdot \exp\left(\frac{df(\theta^*)}{d\theta} t\right) \tag{10.6}$$

Based on Equation 10.6, the following conclusions concerning stability can be made.

1. If $\frac{df(\theta^*)}{d\theta} > 0$, then η grows exponentially and θ^* is *unstable*.
2. If $\frac{df(\theta^*)}{d\theta} < 0$, then η decays exponentially and θ^* is *stable*.
3. If $\frac{df(\theta^*)}{d\theta} = 0$, then the $O(\eta^2)$ terms in the Taylor series given by Equation 10.4 are not negligible, and a nonlinear analysis is needed to determine stability.

The conclusion is that the sign of $df(\theta^*)/d\theta$ determines stability, as we saw graphically from the phase portraits in Section 7.2. This analysis provides a quantitative measure of stability for fixed points.

10.1.1 *RC* Electrical Circuit

Reconsider the series *RC* circuit with a voltage source *V* shown in Figure 7.8. The only fixed point is $Q^* = CV$. Taking the derivative of Equation 7.13 gives

$$\frac{df}{dQ} = -\frac{1}{RC} < 0 \tag{10.7}$$

The physical properties *R* and *C* are always positive, thus $df/dQ < 0$. Thus, stability analysis indicates that Q^* is a *stable* point, as expected from the phase portrait in Figure 7.8 as well as from the physics of this circuit.

10.1.2 Population Model

Returning to the logistic model of population growth of Section 7.2.4, the phase portrait in Figure 7.9 shows that the fixed point $P^* = 0$ is unstable, while $P^* = K$ is stable. Let's see if linear stability theory produces the same conclusions. Taking the derivative of Equation 7.18 produces

$$\frac{dP}{dt} = f(P) = r\left(1 - \frac{P}{K}\right)P, \quad \frac{df}{dP} = r\left(1 - 2\frac{P}{K}\right) \tag{10.8}$$

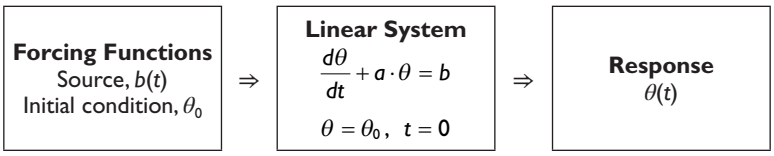
Analysis of the fixed points leads to the following conclusions:

- $P^* = 0$: $df/dP = r > 0 \rightarrow P^*$ is unstable
- $P^* = \mathrm{K}$: $df/dP = -r < 0 \rightarrow P^*$ is stable

Although we reached these same conclusions using the phase portrait shown in Figure 7.9, the graphical approach is often not possible for higher-order systems, and the linearized stability analysis presented in this section is required.

10.2 CHARACTERISTICS OF LINEAR SYSTEMS

A differential equation of the form $d\theta/dt = f(t,\theta)$ is **linear** if the function $f(t,\theta)$ is a linear function of θ. Linear, first-order systems are thus restricted to the special form $d\theta/dt + a \cdot \theta = b$. Here, $a(t)$ is a coefficient that could vary with t, and $b(t)$ is a general time-dependent source term. The cause and effect diagram for this system is



Linear mathematical systems have two important properties: *proportionality* and *superposition*.

Proportionality means that the response or output of the system is a linear combination of the applied forcing functions. That is,

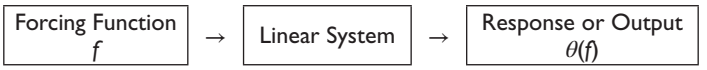
$$\text{Response} = \text{Constant} \times \text{Forcing Function}$$

Superposition means that the total response or output of the system is the sum or superposition of the outputs due to individual forcing functions acting alone.

$$\text{Total Response} = \sum \text{Responses due to each forcing function acting alone}$$

Linearity means that the rule that determines what a piece of a system is going to do next is not influenced by what it is doing now.

Consider a linear system with forcing function or stimulus f and output or response $\theta(f)$, shown in the following cause–effect diagram.

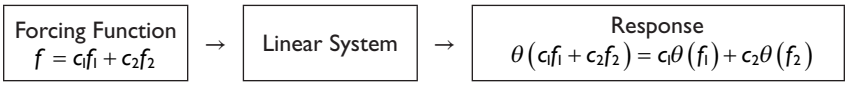


If the input to a linear system is

$$f = c_1 f_1 + c_2 f_2 \tag{10.9}$$

then, by virtue of the proportionality and superposition properties, the response is

$$\theta\left(c_1 f_1 + c_2 f_2\right) = c_1\theta\left(f_1\right) + c_2\theta\left(f_2\right) \tag{10.10}$$



An important consequence is that we can construct solutions to "hard" problems in terms of "easy" problems using *superposition* and *proportionality*.

10.3 SOLUTION USING INTEGRATING FACTORS

Consider the following linear first-order ODE, where the coefficient a is a constant.

$$\frac{d\theta}{dt} + a \cdot \theta = b(t) \tag{10.11}$$

The required initial condition is

$$\theta = \theta_0, \quad t = 0 \tag{10.12}$$

where a is constant but $b(t)$ is a general time-varying source. The solution is obtained by multiplying Equation 10.11 by the integrating factor, e^{at}, and rewriting it as

$$\frac{d}{dt}\left(e^{at}\theta\right) = e^{at}b(t) \tag{10.13}$$

Change the independent variable from t to t_o and integrate from $t_o = 0$ to t to get

$$\underbrace{\int_{t_o=0}^{t} \frac{d}{dt_o}\left(e^{at_o}\theta\left(t_o\right)\right)dt_o}_{e^{at}\theta(t)-e^{0}\theta_0} = \int_{t_o=0}^{t} e^{at_o} b\left(t_o\right)dt_o \tag{10.14}$$

Multiply by e^{-at} and rearrange to obtain the solution as

$$\theta\left(t\right) = \theta_0 e^{-at} + \int_{t_o=0}^{t} e^{a\left(t-t_o\right)} b\left(t_o\right)dt_o \tag{10.15}$$

For the special case of a constant source term, $b(t) = b$, Equation 10.15 reduces to

$$\theta\left(t\right) = \theta_0 e^{-at} + \frac{b}{a}\left(1 - e^{-at}\right) \tag{10.16}$$

Equation 10.16 is plotted in Figure 10.1, demonstrating the effects of the initial condition, the source, and the decay rate. These are three classical and important physical effects with a multitude of important applications.

Next, consider Equation 10.11, where both the coefficient a and the source b can vary with t. The mathematical model is

$$\frac{d\theta}{dt} + a\left(t\right)\cdot\theta = b\left(t\right) \tag{10.17}$$

This solution is obtained by multiplying the previous differential equation by the integrating factor, $\exp\left(\int_{t^*=0}^{t} a(t^*)\,dt^*\right)$, and rewriting it as

$$\frac{d}{dt}\left(\exp\left(\int_{t^*=0}^{t} a\left(t^*\right)dt^*\right)\theta\left(t\right)\right) = \exp\left(\int_{t^*=0}^{t} a\left(t^*\right)dt^*\right)b\left(t\right) \tag{10.18}$$

Changing the independent variable from t to and integrating from $t_o = 0$ to t gives the solution as

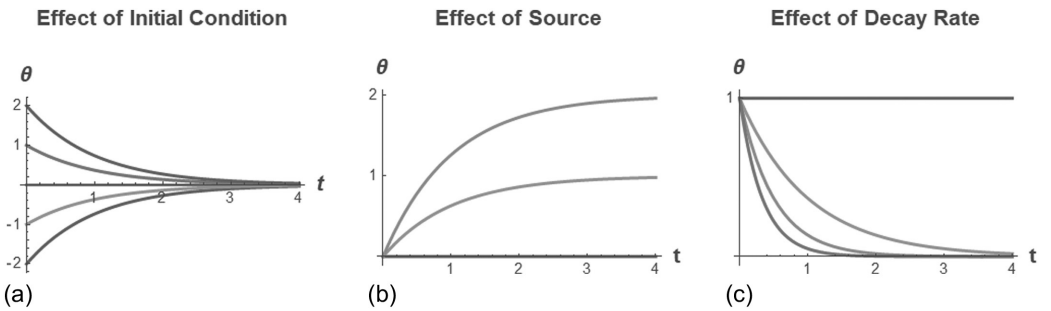


Figure 10.1 (a) Solution starting at various initial conditions with $b = 0$ and $a = 1$. (b) Solution for various sources: $a = 1$, $b = 0, 1, 2$. (c) Solution for various decay rates: $a = 0, 1, 2, 3$, $b = 0$.

$$\theta(t) = \theta_0 \exp\left(-\int_{t^*=0}^{t} a(t^*)\,dt^*\right) + \exp\left(-\int_{t^*=0}^{t} a(t^*)\,dt^*\right)\int_{t_o=0}^{t} \exp\left(\int_{t^*=0}^{t_o} a(t^*)\,dt^*\right) b(t_o)\,dt_o \tag{10.19}$$

10.4 FIRST-ORDER NONLINEAR SYSTEMS AND BIFURCATIONS

In dynamical systems, the stability as well as the number of fixed points can change as parameters change. This is known as a *bifurcation* and is accompanied by a qualitative change in the solution. *Bifurcation theory* is the study of stability changes in nonlinear problems as system parameters are changed. *Bifurcation points* or *critical values* are values of parameters at which the qualitative or topological nature of the dynamics changes.

Bifurcation literally means "splitting into two branches." A bifurcation occurs as a system parameter crosses a bifurcation point or critical threshold. The fixed points and character of the dynamic response depend on whether the parameter is above or below the bifurcation point.

For example, consider the buckling of a beam shown in Figure 10.2. If a small force is applied to the top of the beam, the beam can support the load and remain vertical. But as more force is applied, the load crosses a critical threshold beyond which the vertical position then becomes unstable and the beam may buckle. This situation is the proverbial "straw that breaks the camel's back."

In the following sections, the classical saddle-node, transcritical, supercritical pitchfork, and subcritical pitchfork bifurcations are examined.

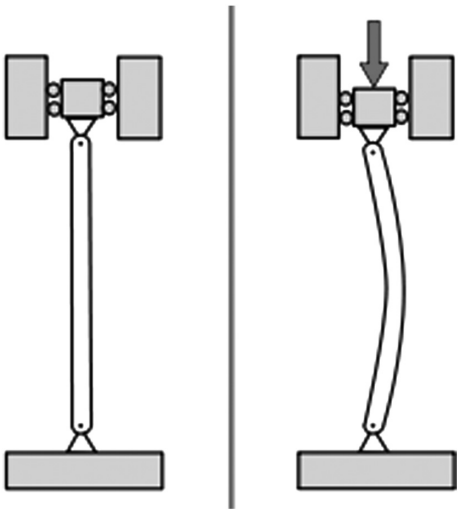


Figure 10.2 Buckling of a beam.

10.4.1 Saddle-Node Bifurcation

The saddle-node bifurcation is the basic mechanism by which fixed points are created and destroyed. As a parameter is varied, two fixed points move toward each other, collide, and mutually annihilate. The prototypical example of a saddle-node bifurcation is the following first-order system:

$$\frac{dx}{dt} = f(x) = r + x^2 \tag{10.20}$$

Setting $dx/dt = 0$ in Equation 10.20 indicates that the fixed points are at $x^* = \pm\sqrt{-r}$. Real-valued fixed points exist only for $r \leq 0$, while no real-valued fixed points exist for $r > 0$. Thus, the bifurcation point or critical threshold for this classical saddle-node is $r_c = 0$.

The phase portraits and numerical solutions are shown in Figure 10.3 for various values of the parameter r relative to the bifurcation point, $r_c = 0$.

As derived in Section 10.1, stability of the fixed points is determined by the sign of $f(x^*)/dx = 2x^*$. Thus, the conclusions regarding stability are:

- $r < r_c$: two fixed points, stable at $x^* = -\sqrt{-r}$, unstable at $x^* = \sqrt{-r}$
- $r = r_c$: one fixed point, half-stable at $x^* = 0$
- $r > r_c$: no fixed points

The fixed points and dynamical response depend on whether the parameter r is above, below, or at the bifurcation point.

The same conclusions can be made by examining the phase portraits in Figure 10.3. The top figures show the phase portrait, dx/dt versus x, with stable fixed points indicated by solid disks and unstable fixed points as open circles. A first-order ODE can be envisioned as a vector field on a line, and the arrows on the phase portraits indicate the flow direction.

The bottom panels in Figure 10.3 show the solutions, $x(t)$ versus t, starting from a number of initial states. The solutions were computed using numerical procedures (Runge–Kutta methods), as described in Section 9.2. Also shown are stable fixed points, indicated by solid lines, and unstable fixed points, indicated by dashed lines. Note how the fixed points and

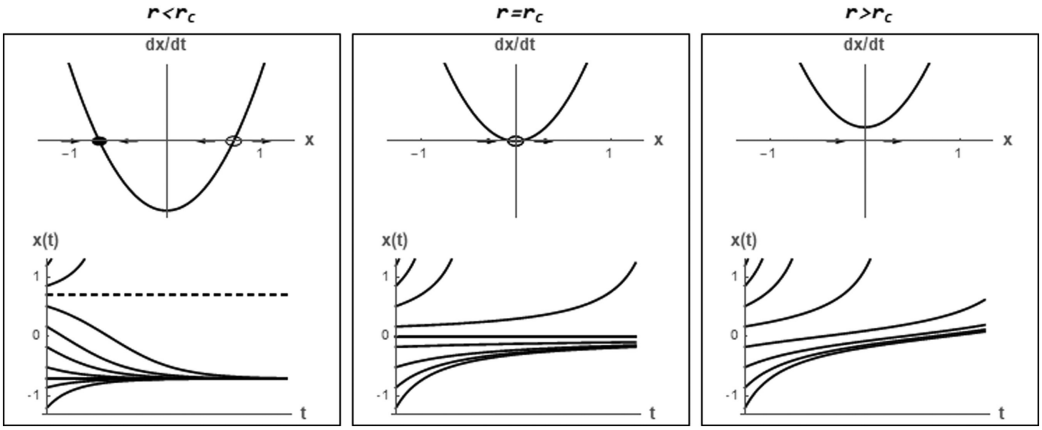


Figure 10.3 Phase portraits and numerical solutions for the typical saddle-node bifurcation.

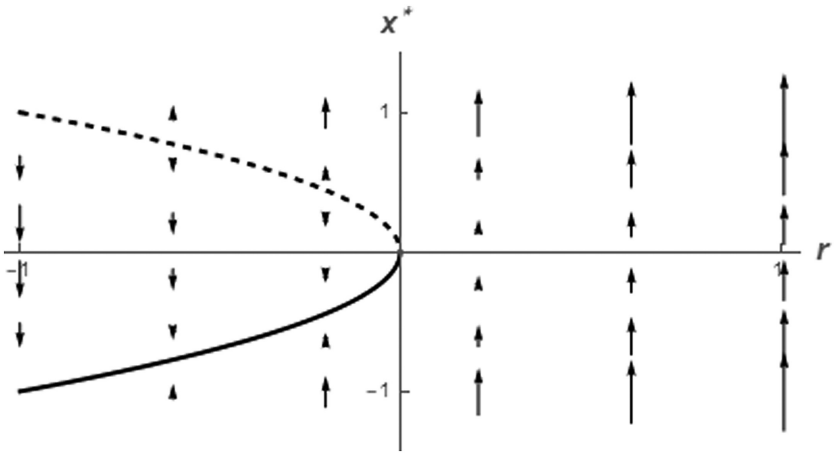


Figure 10.4 Bifurcation diagram for the typical saddle-node bifurcation.

solutions change as bifurcations occur when the parameter r crosses the bifurcation point at $r = r_c = 0$.

The *bifurcation diagram* shown in Figure 10.4 can be deduced from the phase portraits. The solid line represents the stable fixed point at $x^* = -\sqrt{-r}$, while the dashed line represents the unstable fixed point at $x^* = \sqrt{-r}$. The arrows in the bifurcation diagram indicate the direction and magnitude of the velocity of the state variable x for any combination of x and r. At any given r, the arrows represent the vector field for this first-order autonomous system.

10.4.2 Transcritical Bifurcation

There are certain scientific applications where a fixed point exists for all values of a parameter and can never be destroyed. However, it may *change its stability* as a parameter is varied. The transcritical bifurcation is the standard mechanism for such changes in stability. The normal form for a transcritical bifurcation is

$$\frac{dx}{dt} = f(x) = r \cdot x - x^2 \tag{10.21}$$

Setting $dx/dt = 0$ in Equation 10.21 indicates that there are two fixed points at $x^* = 0$ and $x^* = r$. Examining $df(x^*)/dx = r - 2x^*$ indicates that these two fixed points exchange stability at $r = r_c = 0$.

The phase portraits and numerical solutions are shown in Figure 10.5 for the various ranges of the parameter r. As in Figure 10.3, the top panels show the phase portraits with the stable fixed point as a solid disk and the unstable point as an open circle. The bottom panels show the dynamical solution starting from various initial conditions. The phase portraits indicate that for $r < r_c$, $x^* = r$ is unstable, while $x^* = 0$ is stable. However, when $r = r_c$, the fixed points collapse into a single half-stable point. Then, for $r > r_c$, the fixed points exchange stability, with $x^* = r$ stable and $x^* = 0$ unstable.

The bifurcation diagram in Figure 10.6 shows this exchange as r crosses $r_c = 0$. The arrows represent the vector field with flow toward the solid stable fixed points and away from the dashed unstable points.

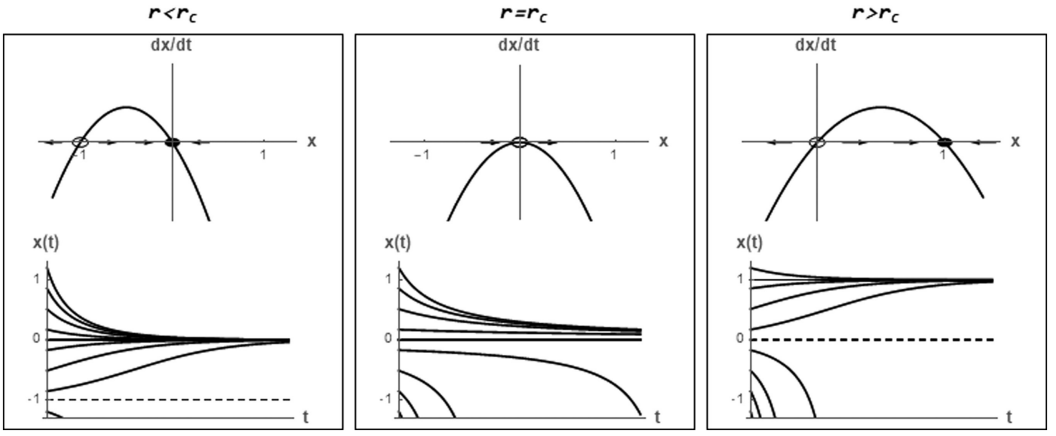


Figure 10.5 Phase portraits and numerical solutions for the typical transcritical bifurcation.

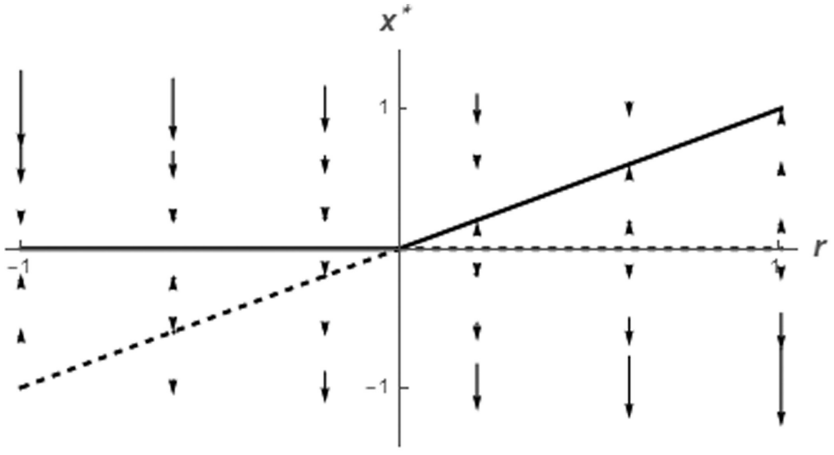


Figure 10.6 Bifurcation diagram for the typical transcritical bifurcation.

10.4.3 Example of a Transcritical Bifurcation: Laser Threshold

Consider a solid-state laser, consisting of a collection of special "laser-active" atoms embedded in a solid-state matrix, bounded by partially reflecting mirrors. An external energy source is used to excite or "pump" the atoms out of their ground state. A schematic is shown in Figure 10.7.

The growth rate of photons is physically modeled as

$$\frac{dn}{dt} = \text{gain} - \text{loss} = G \cdot n \cdot N - k \cdot n \tag{10.22}$$

where

$n(t)$ is the number of photons in the laser field
$N(t)$ is the number of excited atoms

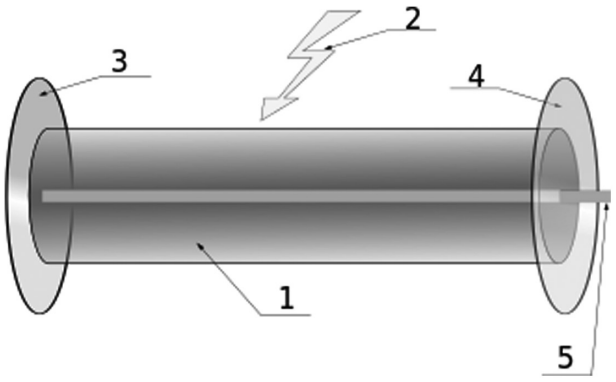


Figure 10.7 Components of a typical laser: (1) solid-state matrix or gain medium; (2) laser pumping energy; (3) high reflector; (4) output coupler; (5) laser beam.

Gain is due to stimulated emission: photons stimulate atoms to emit more photons with a gain coefficient G. Loss is due to escape of photons through the end faces with a rate constant k. The typical lifetime of a photon in the laser is thus $1/k$.

The key idea needed in this model is that after an excited atom emits a photon, it drops down to a lower energy level and is no longer excited. To capture this effect, we assume

$$N(t) = N_0 - \alpha \cdot n \tag{10.23}$$

where

N_0 is the number of exited atoms in the absence of laser action
α is the rate at which atoms drop back to ground state

Substitute Equation 10.23 into Equation 10.22 to get

$$\frac{dn}{dt} = G \cdot n(N_0 - \alpha \cdot n) - k \cdot n = (G \cdot N_0 - k)n - \alpha \cdot G \cdot n^2 \tag{10.24}$$

We have a *transcritical bifurcation* with $(G \cdot N_0 - k)$ playing the role of the parameter r in the normal transcritical form defined by Equation 10.21. The phase portrait and bifurcation diagram for Equation 10.24 are displayed in Figure 10.8.

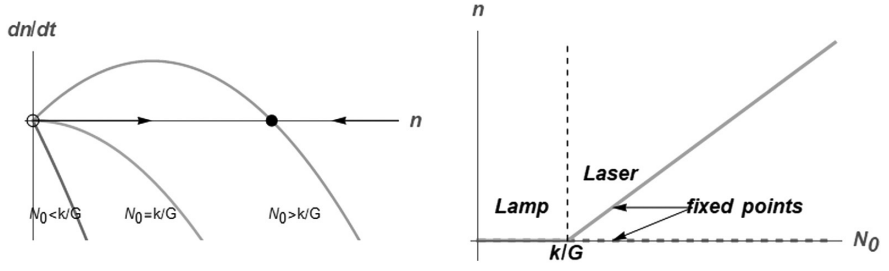


Figure 10.8 Phase portraits and bifurcation diagram for the laser model.

Note the possible dynamics of this system:

- $N_0 < k/G$
  $n^* = 0$ is a stable fixed point. There is no stimulated emission, and the laser acts like a lamp.
- $N_0 = \dfrac{k}{G}$ = laser threshold
  System undergoes a transcritical bifurcation.
- $N_0 > k/G$
  $n^* = 0$ loses stability, and a new fixed point appears at $n^* = \dfrac{(GN_0 - k)}{\alpha} > 0$.

Below the laser threshold, the laser acts like an ordinary lamp, and the atoms oscillate independently and emit random phased light waves. Above the laser threshold, the atoms begin to oscillate in phase, and the lamp has turned into a laser, producing a beam of radiation much more coherent and intense than that produced below the laser threshold. The process is *self-organizing*!

Although many real-world physical effects have been omitted, the model correctly predicts the existence of a threshold.

10.4.4 Supercritical Pitchfork Bifurcation

The pitchfork bifurcation is common in problems that have symmetry. For instance, a beam is stable if the load is small. But if the load exceeds the buckling threshold, the beam may buckle either left or right. The vertical position has become unstable, and two new symmetric fixed points have been born. There are two very different types of pitchfork bifurcations: the *supercritical* and *subcritical* types.

The normal form of the supercritical pitchfork bifurcation is

$$\frac{dx}{dt} = f(x) = r \cdot x - x^3 \tag{10.25}$$

This equation is *invariant* under the change of variable $x \to -x$. Invariance is the mathematical expression of physical symmetry.

The fixed points $x^* = 0$ and $\pm\sqrt{r}$ are found from setting $dx/dt = 0$. Stability is ascertained from the sign of $\dfrac{df(x^*)}{dx} = r - 3x^{*2}$. The bifurcation point is $r_c = 0$, since real fixed points $x^* = \pm\sqrt{r}$ exist only for $r \geq 0$. The conclusions about stability are:

- $r < r_c$: one fixed point, stable at $x^* = 0$
- $r = r_c = 0$: one fixed point, stable at $x^* = 0$, but weakly stable
- $r > r_c$: three fixed points, unstable at $x^* = 0$, stable at $x^* = -\sqrt{r}$, and stable at $x^* = \sqrt{r}$

The phase portraits and numerical solutions are shown in Figure 10.9 for the various ranges of the parameter r. As in Figure 10.4, the top panels show the phase portraits, while the bottom panels show the dynamical solution for a collection of initial conditions.

Figure 10.10 shows the corresponding bifurcation diagram with velocity vectors. The reason for the name "pitchfork" bifurcation is now evident. The phase portraits, solutions, and bifurcation diagram reveal the dynamics of this system.

10.4.5 Subcritical Pitchfork Bifurcation

The normal form of the subcritical pitchfork bifurcation is

$$\frac{dx}{dt} = f(x) = r \cdot x + x^3 \tag{10.26}$$

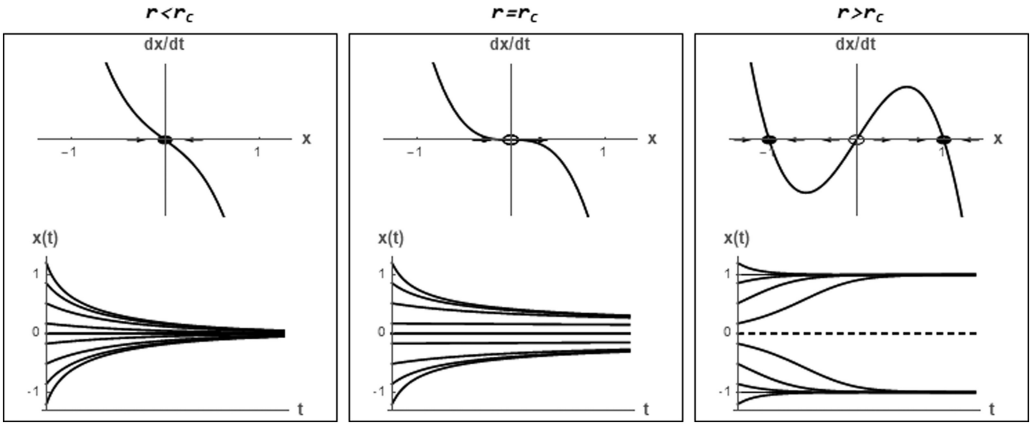


Figure 10.9 Phase portraits and numerical solutions for a typical supercritical pitchfork bifurcation.

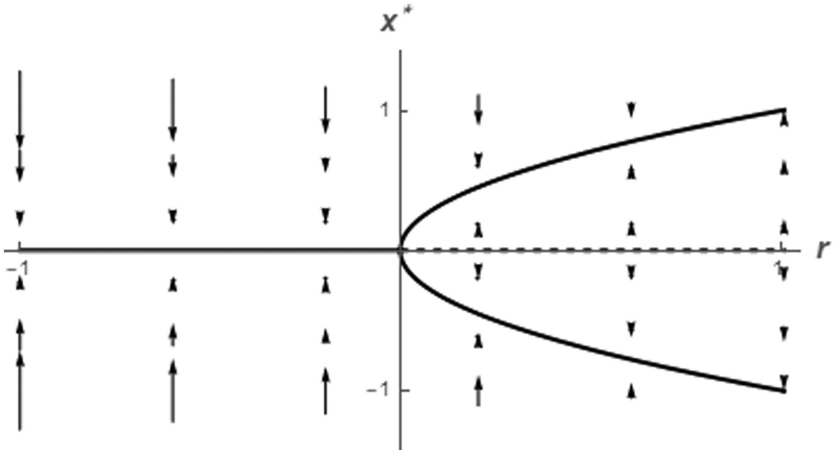


Figure 10.10 Bifurcation diagram for a typical supercritical pitchfork bifurcation.

Note that in the supercritical case, the cubic term is stabilizing; that is, it pulls x back toward 0. On the other hand, the cubic term in the subcritical bifurcation is destabilizing, since it repels x away from 0. The phase portraits are plotted in Figure 10.11, while the bifurcation diagram is shown in Figure 10.12.

In real physical systems, the explosive instability caused by the cubic term is usually opposed by the stabilizing influence of higher-order terms. Assuming that the system is still symmetric under the transformation $x \to -x$, the first stabilizing term is x^5. Thus, consider the system

$$\frac{dx}{dt} = f(x) = r \cdot x + x^3 - x^5 \tag{10.27}$$

The phase portraits are shown in Figure 10.13 and bifurcation diagram is shown in Figure 10.14.

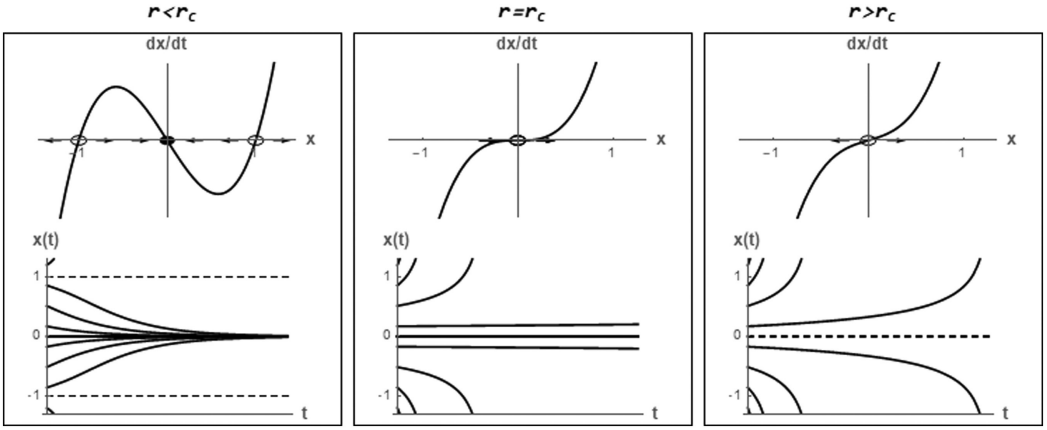


Figure 10.11 Phase portraits and numerical solutions showing a subcritical pitchfork bifurcation

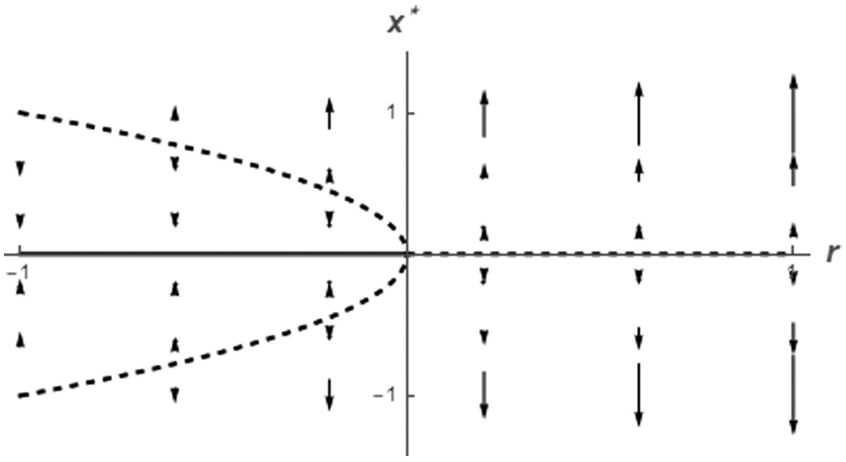


Figure 10.12 Bifurcation diagram for a subcritical pitchfork bifurcation.

There are five fixed points: $x^* = 0$ and $x^* = \pm\sqrt{\frac{1+\sqrt{1+4r}}{2}}$. The system has two bifurcation points: a saddle-node bifurcation at $r_{c1} = -0.25$, where two pairs of fixed points are born, and a subcritical pitchfork at $r_{c1} = 0$.

The range $r_{c1} < r < r_{c2}$ is particularly interesting, since two qualitatively different stable states coexist: the origin and the large-amplitude fixed points at $x^* = \pm\sqrt{\frac{1+\sqrt{1+4r}}{2}}$. The initial condition will determine which of these stable states is approached as time becomes large. In $r_{c1} < r < r_{c2}$, the origin is locally stable to small disturbances but not globally stable, since large perturbations can send the system to one of the large-amplitude fixed points.

Another interesting feature is the existence of *jumps* and *hysteresis* due to the existence of multiple stable fixed points. As the parameter r is varied across the range $r_{c1} < r < r_{c2}$, the system will exhibit hysteresis and jump to a new steady state as the bifurcation points are crossed.

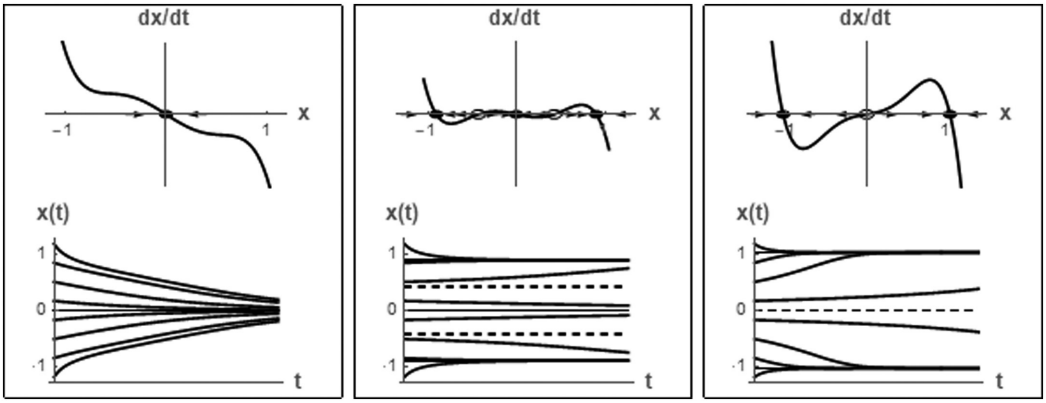


Figure 10.13 Phase portraits and numerical solutions for a subcritical pitchfork bifurcation with a fifth-order stabilizing term.

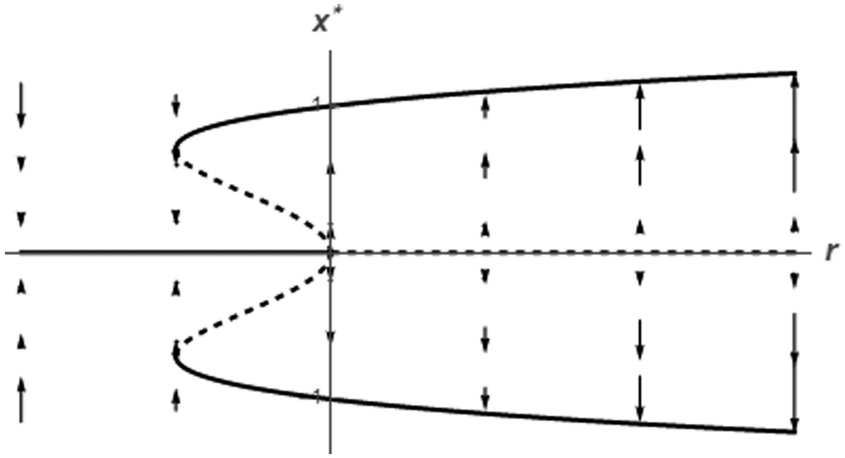


Figure 10.14 Bifurcation diagram for a subcritical pitchfork bifurcation with a fifth-order stabilizing term.

PROBLEMS

Problem 10.1

The functions $\psi_1(t)$and $\psi_2(t)$ are the solutions of the following simple, elementary problems.

$\psi_1(t)$	$\psi_2(t)$
$\frac{d\psi_1}{dt} = -\frac{1}{\tau}\psi_1 + 1$	$\frac{d\psi_2}{dt} = -\frac{1}{\tau}\psi_2$
$\psi_1 = 0, \quad t = 0$	$\psi_2 = 1, \quad t = 0$

Express the solutions to the following problems in terms of the ψ_1 and ψ_2 functions. Be specific about the arguments. Do not actually solve for ψ_1 and ψ_2. A few cases are given as examples in order to clarify the objectives.

$$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S(t)$$

$$\theta = \theta_0, \quad t = 0$$

Case	*Source, S(t)*	*Initial condition, θ_0*	*Solution in terms of $\psi_1(t)$ and $\psi_2(t)$*
Ex. 1	0	4	$4\psi_2(t)$
Ex. 2	$8H(t-t_o)$	0	$8\psi_1(t-t_o)$
1	S_c	0	
2	0	θ_0	
3	3	7	
4	S_c	θ_0	
5	$I(t, t_o, \Delta t)$	0	
6	$\text{RepPulse}(t, t_{\text{cycle}}, t_{\text{on}})$	0	
7	Arbitrary $S(t)$	0	

$H(t)$ is the step function.

$$I(t, t_o, \Delta t) = \text{impulse function} = \frac{H(t-t_o) - H(t-t_o-\Delta t)}{\Delta t}$$

$$\text{RepPulse}(t, t_{\text{cycle}}, t_{\text{on}}) = \text{repetitive pulse}$$

Problem 10.2

Consider the nonlinear problem

$$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta^2$$

$$\theta = \theta_0, \quad t = 0$$

Can the solution be expressed in terms of the ψ_1 and ψ_2 functions defined in the previous problem? If not, what is the difficulty?

Problem 10.3

A basic population dynamics model is

$$\frac{dP}{dt} = aP - bP^2 + S$$

$$P = P_0, \quad t = 0$$

Can the general solution be expressed as the superposition of the simpler solutions in the form

$$P(t) = P_1(t) + P_2(t)$$

where
$P_1(t)$ is due to P_0 only with $S = 0$
$P_2(t)$ is due to S only with $P_0 = 0$.

List the governing equations for $P_1(t)$ and $P_2(t)$ if this superposition attempt is successful.

Problem 10.4

Consider the thermal system described by the equations

$$\rho c V \frac{dT}{dt} = -hA_s\left(T - T_\infty\right) + gV$$

$$T = T_0, \quad t = 0$$

Show that the general solution can be expressed as the superposition of the simpler solutions

$$T(t) = T_1(t) + T_2(t)$$

where
$T_1(t)$ is due to T_0 only with $g = 0$
$T_2(t)$ is due to g only with $T_0 = 0$

List the governing equations for $T_1(t)$ and $T_2(t)$.

Problem 10.5

A lumped thermal system with radiation and heat generation is described by the equations

$$\rho c V \frac{dT}{dt} = -A_s \varepsilon \sigma \left(T^4 - T_{\text{sur}}^4\right) + g(t)V$$

$$T = T_0, \quad t = 0$$

Can the general solution be decomposed as the superposition of the simpler solutions

$$T(t) = T_1(t) + T_2(t)$$

where
$T_1(t)$ is due to T_0 only with $g = 0$
$T_2(t)$ is due to g only with $T_0 = 0$?

List the governing equations for $T_1(t)$ and $T_2(t)$ if this superposition attempt is successful.

Problem 10.6

Consider the linear first-order ODE where τ is constant.

$$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S(t)$$

$$\theta = \theta_0, \quad t = 0$$

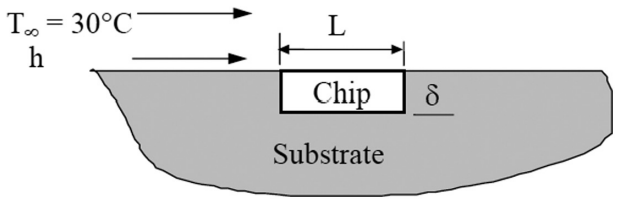
a) Determine the analytical solution for $S(t) = 0$. Create a meaningful plot of this solution.
b) Determine the analytical solution for $S(t) = S_c$ = constant with $\theta_0 = 0$. Create a meaningful plot of this solution.
c) Show that the general solution to this linear problem can be expressed as the superposition of effects due to the initial condition θ_0 and source $S(t)$.
d) Determine the analytical solution for $S(t) = S_c$ = constant and initial condition, θ_0, using the superposition principle demonstrated in part (c) and the results found in parts (a) and (b).
e) Consider the case where $\theta_0 = 0$ with a finite pulse source, defined by

$$S(t) = \frac{S_c}{\Delta t}\left(H(t - t_1) - H(t - t_1 - \Delta t)\right)$$

Determine the analytical solution. Create a meaningful plot of this solution. Consider the effect of Δt for a fixed S_c. Examine the limit as $\Delta t \Rightarrow 0$.

Problem 10.7

A square silicon chip of length $L = 10$ mm on a side and thickness $\delta = 5$ mm is embedded in a well-insulated substrate.



The chip draws $P = 0.5$ W of electrical power and is cooled by convection from the top surface to air at $T_\infty = 30$ °C with a heat transfer coefficient $h = 40$ W/m^2 K. Assume that radiation effects are negligible and that the *lumped capacity* approximation is valid. The chip has the following thermal properties:

$\rho = 2000$ kg/m^3, $c = 700$ J/kg K, $k = 150$ W/m K

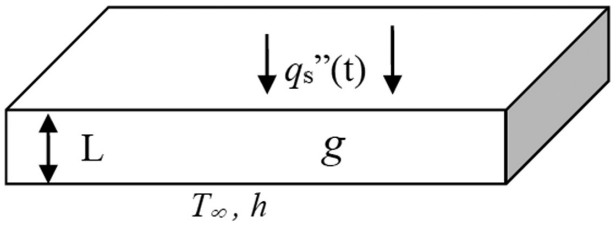
1. The chip is initially at the ambient temperature of T_∞ with the power off. At $t = 0$, the power is switched on.
   a. Sketch the temperature as a function of time for this process.
   b. Derive the energy equation for this case (symbolic form).
   c. Determine the eventual *steady-state* temperature of the chip (numerical value).
2. After the steady condition in Part 1 has been reached, the power is shut off.
   a. Sketch the temperature history as a function of time for this process, starting from the time the power is initially switched on, all the way to a new steady state after the power is switched off.

b. What does the energy equation reduce to for this case (symbolic form)?
c. Determine how long it takes for the chip temperature to drop to 35 °C after the power is shut off (numerical value).

3. After the steady condition in Part 1 has been reached, the system suffers a loss of coolant, causing the heat transfer from the surface due to convection to become negligibly small.
   a. Sketch the temperature history as a function of time for the entire process, starting from the time the power is initially switched on, all the way to the time the chip reaches its failure temperature.
   b. What does the energy equation reduce to for this case (symbolic form)?
   c. Determine how long it takes for the chip to reach its failure temperature of 300 °C after loss of coolant.

Problem 10.8

Consider a thin plate of area A and thickness L with material properties k, c, and ρ. The plate is exposed to convection on the bottom side and to a specified heat flux, $q_s''(t)$, on the top side. In addition, the plate is subjected to a volumetric heat source, $g(t)$ (W/m^3). The initial temperature is T_0. Assume that the lumped capacity approximation is valid.



a) Derive the energy equation. Formulate the complete mathematical model of the system.
b) Sketch the solution assuming g and q_s'' are constant. Consider $T_0 = T_\infty$. On a single graph, put curves for a zero, medium, and large value of h. On another graph, show the effect of the volumetric heat source, g. Try to create a meaningful graph that highlights the effect of g only.
c) Consider the mathematical model from part (a) in the form

$$\frac{d\theta}{dt} = -\frac{1}{\tau}\theta + S$$
$$\theta = \theta_0, \quad t = 0$$

1. What are the temperature rise, $\theta(t)$, the initial temperature, θ_0, the time constant, τ, and the source term, S?

d) Using the model from part (c), solve for the transient temperature solution, $\theta(t)$ when $S(t)$ is a constant. Plot your solution.
e) Using the model from part (c), solve for the transient temperature solution, $\theta(t)$ when $S(t)$ is a pulse:

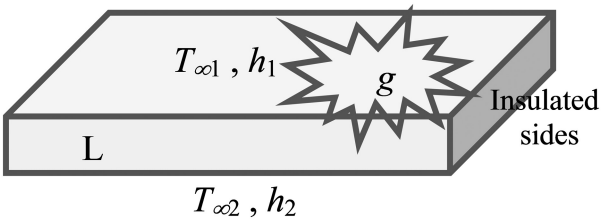
$$S(t) = S_0\left(H(t) - H(t - \Delta t)\right)$$

where
S_0 is a constant
Δt is the pulse time
Plot your solution.

f) Show that due to the linear nature of this problem, the solution can be constructed as a superposition or summation of two simpler solutions, one due to the heat source $S(t)$ only and another due to the initial condition θ_0 only. Do not simply examine the mathematical solution to verify that this works (zero credit for this approach). Instead, try the superposition $\theta(t) = \theta_s(t) + \theta_{ic}(t)$ directly in the governing equations. What equations govern the functions θ_s and θ_{ic}?

Problem 10.9

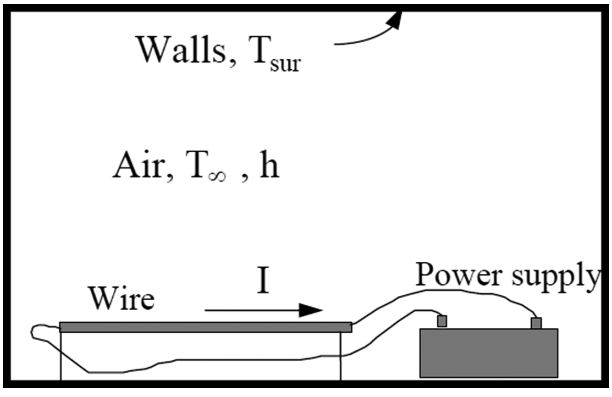
Consider a large, thin plate of area A and thickness L with material properties k, c, and ρ. The plate is exposed to convection at different environments on the top and the bottom. In addition, the plate is subjected to a constant volumetric heat source, g W/m^3. The initial temperature is T_0. Assume that the lumped capacity approximation is valid.



a) Derive the energy equation.
b) What is the steady-state temperature?
c) Solve for the transient temperature solution, $T(t)$.
d) Transform the energy equation into dimensionless form.
e) Sketch the anticipated solution for $h_1 >> h_2$, $h_1 = h_2$, and $h_1 << h_2$ with $g = 0$, $T_0 = (T_1 + T_2)/2$, and $T_1 > T_2$. Put all curves on a single graph.
f) On another graph, show the effect of the volumetric heat source, g. Try to create a meaningful graph that highlights the effect of g only.

Problem 10.10

A long, thin copper wire of diameter D and length L has an electrical resistance per length of wire R'_e (Ω/m), density ρ, specific heat c, and total emissivity ε. The wire is initially at steady state at temperature T_0. At time $t = 0$, an electric current I (amps) is passed through the wire, causing electrical resistance heating. As the wire temperature rises, heat is dissipated by convection to the air at temperature T_∞ with a heat transfer coefficient h and by radiation to the walls at temperature T_{sur}.



Assumptions

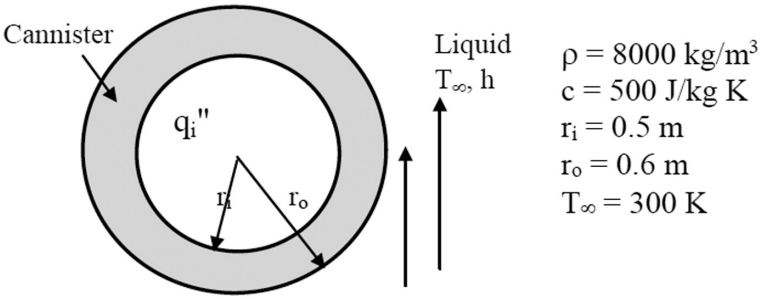
- Constant thermal and electrical properties
- Lumped capacity, valid for low Bi. That is, spatially uniform temperature at any time, $T = T(t)$.
- Negligible conduction from the ends or stand

a) Derive the mathematical model governing the transient temperature of the wire.
b) For the case where $T_0 = T_\infty = T_{sur}$, sketch the anticipated temperature versus time behavior of the wire. On a single graph, sketch curves corresponding to free convection, mild forced convection, and strong forced convection. Clearly label your graph.
c) Neglecting radiation effects, derive an analytical solution for the transient temperature of the wire. What is the steady temperature of the wire for this case?
d) Create a function to evaluate the solution numerically. Plot temperature versus time using the following parameters. Put all three curves on a single graph.
h values of 5, 25, and 100 W/m^2 K
$T_0 = T_\infty = T_{sur} = 300$ K

D = diameter = 0.001 m, L = length = 0.4 m
ρ = density = 8933 kg/m^3, k = 400 W/m K
c = specific heat = 385 J/kg K
ε = emissivity = 0.9
I = 6 amps, ρ_e = electrical resistivity = 80 μΩ-cm

Problem 10.11: Canister Wall

A spherical, stainless steel (AISI 302) canister is used to store reacting chemicals that provide for a uniform heat flux q_i'' to its inner surface. The canister is suddenly submerged in a liquid bath of temperature $T_\infty < T_i$, where T_i is the initial temperature of the canister wall.



a) Assuming negligible temperature gradients in the canister wall and a constant heat flux q_i'', develop an equation that governs the variation of the wall temperature with time during the transient process.
b) Develop an expression for the steady temperature of the wall.
c) The convection coefficient depends on the velocity of the fluid and whether or not the wall temperature is large enough to induce boiling in the liquid bath. For the parameters listed in the following, compute and plot the steady-state temperature as a function of h for the range $100 < h < 10{,}000$ W/m^2 K. Include curves for $q_i'' = 10^5$, 2×10^5, and 3×10^5 W/m^2. Put all three curves on a single graph.
d) Is there any value of h below which operation would be unacceptable?

Problem 10.12

Consider the system $dx/dt = r \cdot x - x^3$. For $r < 0$, $r = 0$, and $r > 0$:

a) Use linear stability analysis to classify the fixed points.
b) Plot the phase portraits and the anticipated solutions, $x(t)$ versus t.

Problem 10.13: Bifurcations

Consider the following first-order ODEs:

$$\frac{dx}{dt} = x\left(1 + \frac{r}{1+x^2}\right)$$

$$\frac{dx}{dt} = -rx + \frac{x}{1+x}$$

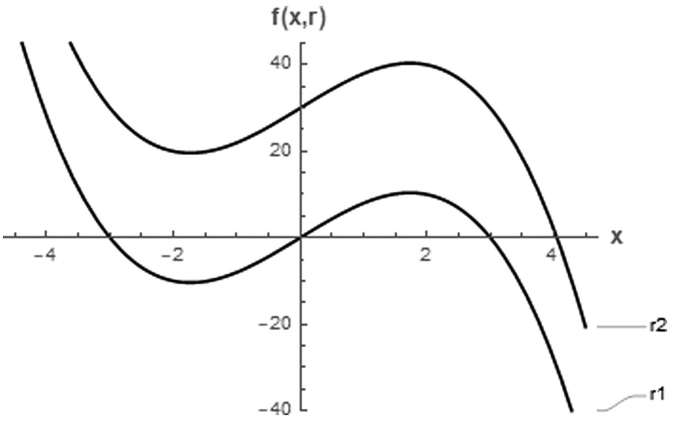
a) Find all the fixed points.
b) Find any critical values r_c at which bifurcations occur. What type of bifurcations occur (saddle-node, transcritical, supercritical pitchfork, or subcritical pitchfork)?
c) Classify the stability of the fixed points. Collect your results in a table.
d) Sketch the bifurcation diagram: x^* versus r.
e) Sketch the phase portraits and corresponding solutions for each qualitatively different behavior.

Problem 10.14: Transient Parameter

Consider a first-order ODE of the form

$$\frac{dx}{dt} = f(x,r)$$

where r is a parameter. The phase diagrams for two different values of the parameter r are shown.



Consider a process where the parameter r is at a value r_1 for quite some time and then changes to a new value r_2. Sketch the resulting dynamic response of the system ($x(t)$ versus t) starting from initial states $x_0 = -4$, -2, 0, and 2.

Problem 10.15: Improved Laser

An improved model of a laser is

$$\frac{dn}{dt} = GnN - kn$$

$$\frac{dN}{dt} = -GnN - fN + p$$

a) Assuming $\frac{dN}{dt} \approx 0$, derive a differential equation for n.
b) Show that $n^* = 0$ becomes unstable for $p > p_c$. Determine p_c.
c) What type of bifurcation occurs at p_c?
d) For what range of parameters is it valid to assume that $\frac{dN}{dt} \approx 0$?

Problem 10.16: Biochemical Switch

Zebra stripes and butterfly wing patterns are two of the most spectacular examples of biological pattern formation. Explaining the development of these patterns is one of the outstanding problems of biology.

As one ingredient in a model of pattern formation, consider a simple example of a biochemical switch, in which a gene G is activated by a biochemical signal substance S. For example, the gene may normally be inactive but can be "switched on" to produce a pigment or other gene product when the concentration of S exceeds a certain threshold. Let $g(t)$ denote the concentration of the gene product, and assume that the concentration s_o of S is fixed. The model is

$$\frac{dg}{dt} = k_1 s_o - k_2 g + \frac{k_3 g^2}{k_4 + g^2}$$

where the ks are positive constants. The production of g is stimulated by s_o at a rate k_1 and by an *autocatalytic* or positive feedback process (the nonlinear term). There is also a linear degradation of g at a rate k_2.

a) Show that the system can be put in the dimensionless form

$$\frac{dx}{d\tau} = s - rx + \frac{x^2}{1 + x^2}$$

where $r > 0$ and $s \geq 0$ are dimensionless groups. What are the dimensionless variables and parameters?
b) For $s = 0$, find and *classify* all the positive fixed points x^*. Is there any critical value r_c? Determine and plot the fixed points for all r and s.
c) Find the parametric equations for the bifurcation curves in (r, s) space and classify the bifurcations that occur. Plot the stability diagram in (r, s).
d) Assume that initially there is no gene product, that is, $g(0) = 0$, and suppose s is slowly increased from zero (the activating signal is turned on); what happens to $g(t)$? What happens if s then goes back to zero? Does the gene turn off again? Support your findings using *phase plots* and *numerical solutions* of the x versus τ behavior.

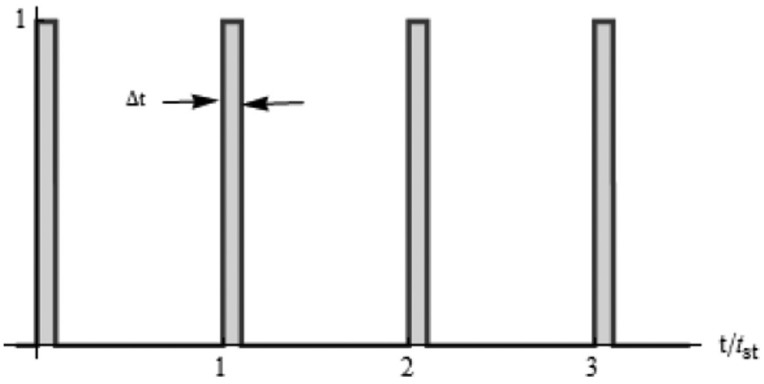
Problem 10.17: Trout Fishing in Stocked Streams

Fishing in stocked trout waters usually involves a sudden addition of fish on stocking days immediately followed by lots of fisherman feverishly trying to catch them. The rate of fish caught is usually proportional to the number of fish still available. Assuming no natural reproduction, the following model is proposed:

$$\frac{dN}{dt} = -rN + S(t)$$

The stocking function is modeled as a periodic sequence of regularly spaced pulse functions as shown where

N_{st} = number of fish per stocking
t_{st} = time between stockings
Δt = time required for each stocking



a) Express the stocking function $S(t)$ in mathematical form.
b) Determine the analytical solution for a single stocking event. Name this function $N_i(t, \Delta t)$ where $i = 1,2,3,\ldots$ is the ith stocking event.
c) Determine the general solution for a sequence of stocking events in terms of the function $N_i(t, \Delta t)$.
d) Determine the limit as $\Delta t \to 0$.
e) Plot your solution from part (d). Use parameter values that make sense.

Problem 10.18: Model of a Fishery

The equation

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) - H$$

provides a simple model of a fishery. Here, $N(t)$ is the number of fish. In the absence of fishing, the population is assumed to grow logistically. The effects of fishing are modeled by the term $-H$, which says that fish are caught or "harvested" at a constant rate $H > 0$, independently of the fish population N. This assumes that the fishermen are not worried about fishing the population dry; they simply catch the same number of fish every day.

a) Show that the system can be rewritten in dimensionless form as

$$\frac{dx}{dt} = x(1-x) - h$$

What are the dimensionless variables x, τ, and h?

b) Determine the fixed points. Classify the stability of the fixed points. Show that a bifurcation occurs for a certain critical value h_c. What type of bifurcation is this? Plot a bifurcation diagram (x^* versus h).

c) Plot separate phase portraits for $0 < h < h_c$, $h = h_c$, and $h > h_c$.

d) Plot the solutions $x(t)$ versus τ for $0 < h < h_c$, $h = h_c$, and $h > h_c$. Create three plots, one corresponding to each range of h. Include several initial conditions on each plot.

e) Discuss any unrealistic aspects of this model. Can you suggest an improved model? Sketch the anticipated solution for your improved model.

Problem 10.19: Improved Model of a Fishery

The equation

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) - H\frac{N}{A+N}$$

provides a model of a fishery. Here, $N(t)$ is the number of fish. In the absence of fishing, the population is assumed to grow logistically. The effects of fishing are modeled by the term $H\frac{N}{A+N}$, where $H > 0$ and $A > 0$, which says that fish are caught or "harvested" at a rate that decreases with the fish population N. This is plausible, since it gets harder to catch fish as the population decreases.

a) Show that the system can be rewritten in dimensionless form as

$$\frac{dx}{d\tau} = x(1-x) - h\frac{x}{a+x}$$

What are the dimensionless quantities x, τ, a, and h?

b) Determine the fixed points. Classify the stability of the fixed points using *linear stability analysis*. Collect your findings in an organized table. What types of bifurcations occur?

c) Plot the stability diagram in (a, h) space. Can hysteresis occur in any of the regions?

d) For each qualitatively different behavior, plot

- The phase diagram $\left(\frac{dx}{d\tau} \text{ vs. } x\right)$
- The transient responses (x vs. τ) starting from various initial conditions

Chapter 11

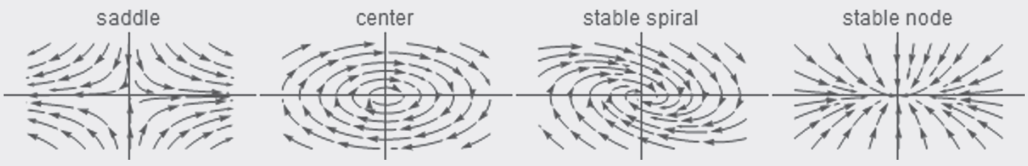
Second-Order Ordinary Differential Equations

CHAPTER OBJECTIVES

The primary objective of this chapter is to develop solutions for and understand the behavior of second-order initial and boundary value differential equations. In order to develop an intuitive feel, solutions will be visualized using *phase portraits*—graphical depictions of the trajectories in the phase plane. First, linear problems are studied in detail. Then, the insights gained from linear problems are extended to the rich world of nonlinear problems.

Specific objectives and topics covered are

- Solutions and classification of linear systems
- Behavior of linear mechanical oscillators
- Stability analysis of the fixed points
- The pendulum
- Competition models (rabbit versus sheep)
- Limit cycles
- Bifurcations
- Coupled oscillators



11.1 LINEAR SYSTEMS

The analytical solution to the second-order linear autonomous systems is described next. The system of differential equations and initial conditions are

$$\begin{aligned} \frac{dx}{dt} &= a \cdot x + b \cdot y \\ \frac{dy}{dt} &= c \cdot x + d \cdot y \end{aligned} \tag{11.1}$$

$$\left.\begin{aligned} x &= x_0 \\ y &= y_0 \end{aligned}\right\} \quad t = 0 \tag{11.2}$$

In matrix form,

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x}$$

$$\mathbf{x} = \mathbf{x}_0, \quad t = 0 \tag{11.3}$$

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

Note that $\mathbf{x}^* = [0,0]^T$ is the only fixed point for any **A**. Based on our success with exponential solutions for single, linear first-order ordinary differential equations (ODEs), we try to find a solution in the form

$$\mathbf{x} = e^{\lambda t}\mathbf{v}$$

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \text{eigenvector}, \quad \lambda = \text{eigenvalue} \tag{11.4}$$

Figure 11.1 depicts a solution in the form suggested by Equation 11.4.

This proposed solution is composed of a vector **v** in the phase plane that either grows, shrinks, or oscillates according to the nature of the eigenvalues, λ. If appropriate eigenvalues and eigenvectors can be determined, we have a valid solution. Note that a single eigenvector restricts the solution to one direction in the phase plane. Thus, we expect or hope that two eigenvalues with two eigenvectors in linearly independent directions will be found.

Substitute our proposed solution, Equation 11.4, into the original system of ODEs given by Equation 11.3 and cancel the exponent to get

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

$$(\mathbf{A} - \lambda \cdot \mathbf{I})\mathbf{v} = 0 \tag{11.5}$$

$$\mathbf{I} = \text{identity matrix}$$

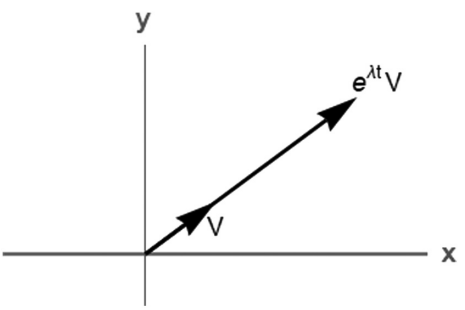


Figure 11.1 Solution composed of an eigenvector scaled by an exponent.

This is the traditional eigenvalue problem from linear algebra. The eigenvalues are the solutions of the *characteristic equation*

$$\begin{aligned}\det\left(\mathbf{A}-\lambda\cdot\mathbf{I}\right) &= \det\begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} \\ &= \lambda^2-\left(a+d\right)\lambda+\left(ad-bc\right)=\lambda^2-\tau\cdot\lambda+\Delta=0\end{aligned} \tag{11.6}$$

where the trace and determinant of the coefficient matrix **A** are defined as

$$\begin{aligned}\text{trace}\left(\mathbf{A}\right)&=\tau=a+d \\ \det\left(\mathbf{A}\right)&=\Delta=ad-bc\end{aligned} \tag{11.7}$$

The solution of this quadratic Equation 11.6 gives the desired eigenvalues

$$\lambda_1=\frac{\tau+\sqrt{\tau^2-4\Delta}}{2},\quad \lambda_2=\frac{\tau-\sqrt{\tau^2-4\Delta}}{2} \tag{11.8}$$

The system $\left(\mathbf{A}-\lambda\cdot\mathbf{I}\right)\mathbf{v}=0$ now allows us to determine the eigenvectors. The two equations represented by this system are not linearly independent; thus, the eigenvectors can be found only to within an undetermined constant. The general solution thus takes on the form

$$\mathbf{x}\left(t\right)=c_1e^{\lambda_1 t}\mathbf{v}_1+c_2e^{\lambda_2 t}\mathbf{v}_2 \tag{11.9}$$

To determine the constants c_1 and c_2, force the solution to satisfy the initial condition.

$$\mathbf{x}_0=c_1\mathbf{v}_1+c_2\mathbf{v}_2 \tag{11.10}$$

The general analytical solution is now complete for any linear system of two autonomous, first-order ODEs. Linear systems consisting of any number of coupled first-order ODEs can be solved in a similar manner. Note that the solution in general involves a combination of two linearly independent eigenvectors, as shown in Figure 11.2. Thus, we have the possibility of following the solution as it travels about the phase plane.

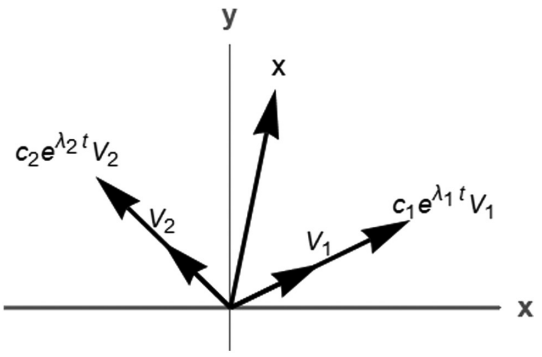


Figure 11.2 Visual representation of Equation 11.9.

11.2 CLASSIFICATION OF LINEAR SYSTEMS

All the information needed to classify second-order equations is contained in the eigenvalues given by Equations 11.8. The various types of behavior and corresponding stability of the fixed point $\mathbf{x}^* = [0,0]^T$ are summarized in Table 11.1.

Some observations are:

- $\Delta < 0$, eigenvalues are real and have opposite signs: hence, the fixed point is a saddle point.
- $\Delta > 0$, eigenvalues are either real with the same sign (nodes) or complex conjugates (spirals).
  - Nodes satisfy $\tau^2 - 4\Delta > 0$.
  - Spirals satisfy $\tau^2 - 4\Delta < 0$.
- $\tau^2 - 4\Delta = 0$ is the border between nodes and spirals. Star nodes and degenerate nodes are found here.
- Stability of the nodes and spirals is determined by τ.
  - $\tau < 0$, both eigenvalues have negative real parts and are stable.
  - $\tau > 0$, both eigenvalues have positive real parts and are unstable.
  - $\tau = 0$, eigenvalues are purely imaginary.

The eigenvalues are the key to determining the type of behavior to expect. All these types of behaviors are pictured on the τ versus Δ diagram in Figure 11.3.

11.3 CLASSICAL SPRING-MASS-DAMPER

Figure 11.4 shows an all-time classic—the damped spring-mass system. Applying Newton's second law produces Equation 7.26. For zero applied force, this equation reduces to

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = 0 \tag{11.11}$$

This equation can be expressed in terms of natural frequency and damping ratio as

$$\frac{d^2x}{dt^2} + 2\zeta\omega_n\frac{dx}{dt} + \omega_n^2 x = 0 \tag{11.12}$$

Table 11.1 Classification of linear second-order systems

Conditions			*Eigenvalues* λ_1 *and* λ_2	*Stability of fixed point* $x^* = [0,0]^T$
Δ	τ	$\tau^2 - 4\Delta$		
<0	all values	>0	real, one positive, one negative	Saddle
>0	<0	>0	real, negative	Stable node
		<0	complex conjugates, negative real parts	Stable spiral
	=0	<0	complex conjugates, purely imaginary	Center
	>0	<0	complex conjugates, positive real parts	Unstable spiral
		>0	real, positive	Unstable node

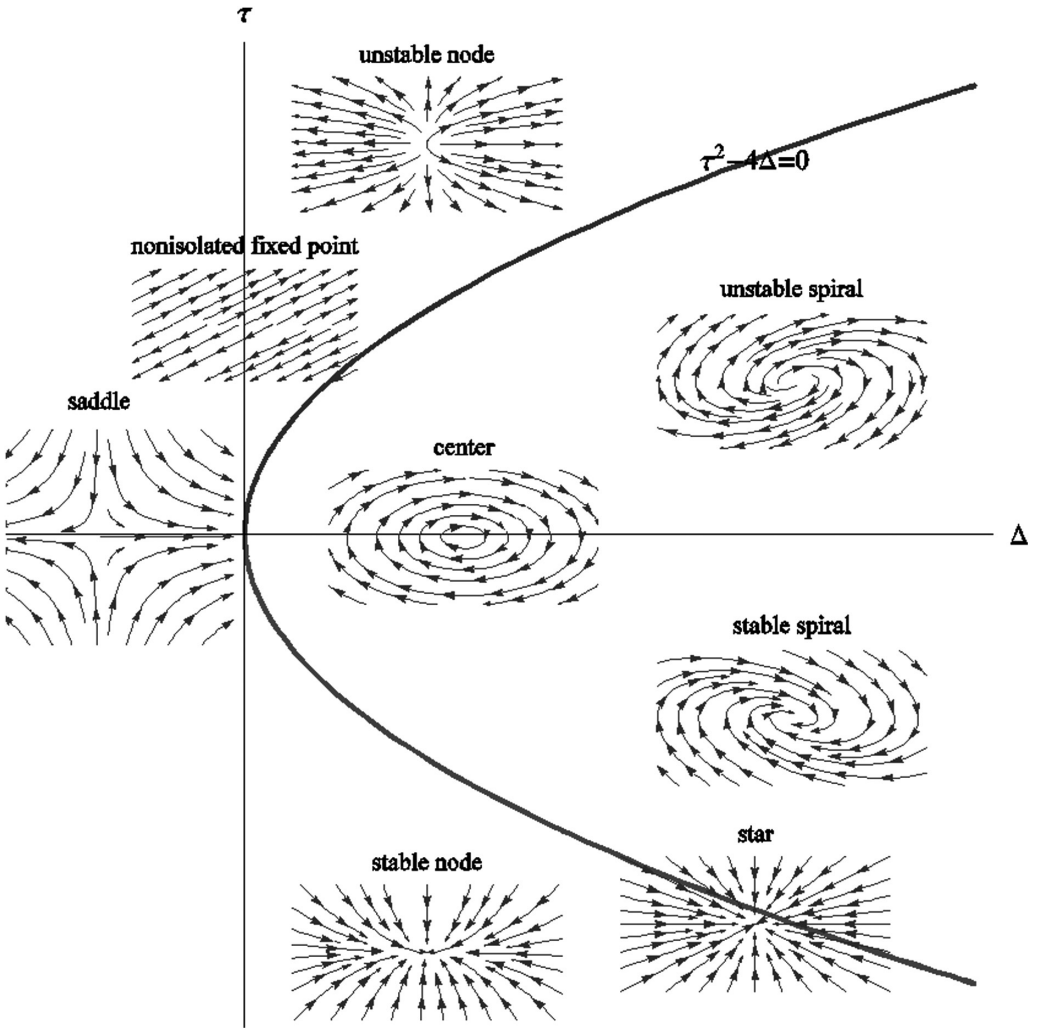


Figure 11.3 Classification of linear second-order systems.

$$\omega_n = \sqrt{k/m} = \text{natural frequency}$$

$$\zeta = \frac{c}{2\sqrt{km}} = \text{damping ratio}$$

We can use the definition of velocity to write Equation 11.11 as an equivalent coupled set of first-order equations.

$\dfrac{dx}{dt} = v$	Matrix Form
$\dfrac{dv}{dt} = -2\zeta\omega_n v - \omega_n^2 \cdot x$	$\begin{bmatrix} \dfrac{dx}{dt} \\ \dfrac{dv}{dt} \end{bmatrix} = \mathbf{A}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\zeta\omega_n \end{bmatrix}\begin{bmatrix} x \\ v \end{bmatrix}$

(11.13)

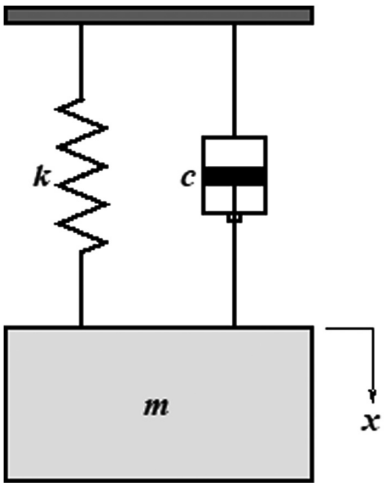


Figure 11.4 The spring-mass-damper system with variable definitions. $x(t)$ = displacement from static equilibrium (m), m = mass (kg), c = damping coefficient (N s/m), and k = spring constant (N/m).

The initial position and velocity need to be specified in order to obtain a unique solution.

$$\left.\begin{aligned} x &= x_0 \\ \frac{dx}{dt} &= v_0 \end{aligned}\right\} \quad t = 0 \tag{11.14}$$

We could obtain an exact solution for this system in terms of sines and cosines. However, as soon as we get to nonlinear equations, exact solutions are impossible to find. As a result, we want to first deduce the behavior without actually solving the equations.

The motion in the phase plane is determined by a vector field that comes directly from the differential equations given by Equations 11.13. These trajectories are the parametric representation of the solution (x,v). The nature of the solution is determined by the eigenvalues. From the matrix form of Equation 11.13, we have

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\zeta\omega_n \end{bmatrix} \tag{11.15}$$

$$\begin{aligned} \text{trace}(\mathbf{A}) &= \tau = -2\zeta\omega_n \\ \det(\mathbf{A}) &= \Delta = \omega_n^2 \end{aligned} \tag{11.16}$$

Then, from Equations 11.8, we get the eigenvalues

$$\begin{aligned} \lambda_1 &= \frac{\tau + \sqrt{\tau^2 - 4\Delta}}{2} = \omega_n\left(-\zeta + \sqrt{\zeta^2 - 1}\right) \\ \lambda_2 &= \frac{\tau - \sqrt{\tau^2 - 4\Delta}}{2} = \omega_n\left(-\zeta - \sqrt{\zeta^2 - 1}\right) \end{aligned} \tag{11.17}$$

The complete solution can be obtained by enforcing the initial conditions. The solution is

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

$$v(t) = c_1 \lambda_1 e^{\lambda_1 t} + c_2 \lambda_2 e^{\lambda_2 t} \tag{11.18}$$

$$c_1 = \frac{v_0 - \lambda_2 x_0}{\lambda_1 - \lambda_2}, \quad c_2 = \frac{\lambda_1 x_0 - v_0}{\lambda_1 - \lambda_2}$$

Some interesting and important special cases can be obtained, depending on the magnitude of the dimensionless damping coefficient.

(a) **$\zeta = 0$: Zero Damping, Center**

The eigenvalues from Equations 11.17 reduce to

$$\lambda_1 = i\omega_n, \quad \lambda_2 = -i\omega_n \tag{11.19}$$

This case is a *center* with pure imaginary eigenvalues. Figure 11.5 shows the phase portrait on the left with the specific trajectory starting from the initial condition $(x_0, v_0) = (0,1)$ highlighted with a thick line. The time trace of position and velocity starting from this same initial condition is shown on the right.

(b) **$\zeta < 1$: Underdamped, Stable Spiral**

The eigenvalues from Equations 11.17 reduce to

$$\lambda_1 = \omega_n\left(-\zeta + i\sqrt{1-\zeta^2}\right), \quad \lambda_2 = \omega_n\left(-\zeta - i\sqrt{1-\zeta^2}\right) \tag{11.20}$$

This case is a *stable spiral*, since the eigenvalues are complex conjugates with a negative real part. The phase portrait and a typical trajectory are shown in Figure 11.6.

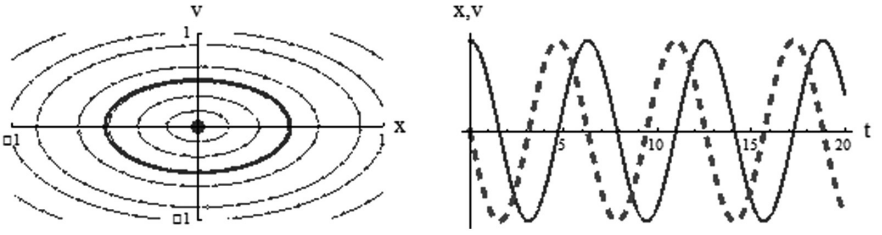


Figure 11.5 Phase portrait and typical solution for zero damping.

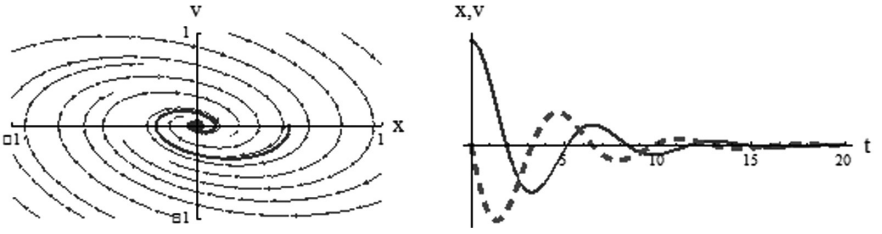


Figure 11.6 Phase portrait and typical solution for an underdamped system.

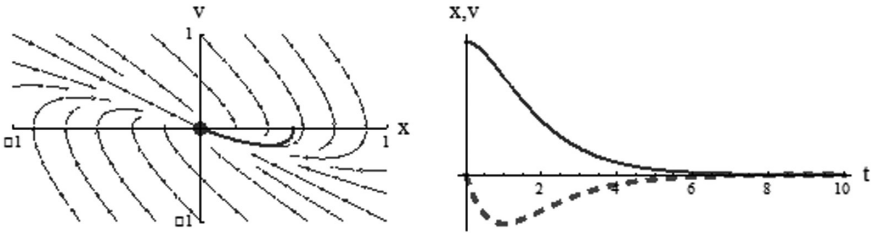


Figure 11.7 Phase portrait and typical solution for critical damping.

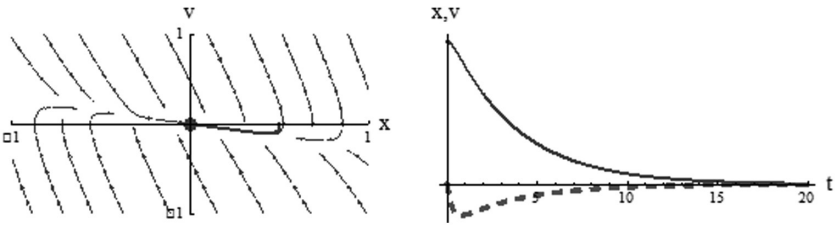


Figure 11.8 Phase portrait and typical solution for an overdamped system.

(c) **$\zeta = 1$: Critically Damped, Degenerate Node**

The eigenvalues from Equations 11.17 reduce to

$$\lambda_1 = \lambda_2 = -\omega_n \tag{11.21}$$

This case is a *degenerate node*. The eigenvalues are repeated negative real numbers. The behavior is demonstrated in Figure 11.7.

(d) **$\zeta > 1$: Overdamped, Stable Node**

The eigenvalues from Equations 11.17 reduce to

$$\lambda_1 = \omega_n\left(-\zeta + \sqrt{\zeta^2 - 1}\right), \quad \lambda_2 = \omega_n\left(-\zeta - \sqrt{\zeta^2 - 1}\right) \tag{11.22}$$

This case is a *stable node* or an overdamped system. Both eigenvalues are real and negative and the behavior is plotted in Figure 11.8.

Note that a negative damping coefficient would produce unstable spirals and unstable nodes. However, it has no meaning for classical linear oscillators and is not considered in this example.

11.4 STABILITY ANALYSIS OF THE FIXED POINTS

The linearized stability technique developed for one-dimensional systems in Section 10.1 is extended to two-dimensional systems. An autonomous system is one with no external driving forces and thus, no explicit time dependence. The general form for a second-order system is

$$\begin{aligned} \frac{dx}{dt} &= f(x,y) \\ \frac{dy}{dt} &= g(x,y) \end{aligned} \tag{11.23}$$

Suppose our system has a fixed point at (x^*, y^*). This means that

$$\begin{aligned} f\left(x^*, y^*\right) &= 0 \\ g\left(x^*, y^*\right) &= 0 \end{aligned} \tag{11.24}$$

Now, we place the system at the fixed point, perturb it slightly, and examine the subsequent behavior to deduce stability. This process is the mathematical counterpart of building the physical system, placing it carefully at equilibrium, giving it a slight nudge, and observing the behavior. If stable, the system will return to its equilibrium point. If unstable, the system will fly away to some other location.

Let small disturbances or perturbations from the fixed point be defined as

$$\begin{aligned} u &= x - x^* \\ v &= y - y^* \end{aligned} \tag{11.25}$$

To see whether the disturbance grows or decays, we derive the differential equations governing the dynamics of the perturbations by taking the derivatives:

$$\begin{aligned} \frac{du}{dt} &= \frac{d\left(x - x^*\right)}{dt} = \frac{dx}{dt} \Leftarrow \text{since } x^* \text{ is constant} \\ &= f\left(x^* + u, y^* + v\right) \Leftarrow \text{use a Taylor series} \\ &= \underbrace{f\left(x^*, y^*\right)}_{0} + u\frac{\partial f}{\partial x}\left(x^*, y^*\right) + v\frac{\partial f}{\partial y}\left(x^*, y^*\right) + O\left(u^2, v^2, uv\right) \\ &\cong u\frac{\partial f}{\partial x}\left(x^*, y^*\right) + v\frac{\partial f}{\partial y}\left(x^*, y^*\right) \end{aligned} \tag{11.26}$$

A similar relation can be derived for the y perturbation.

$$\frac{dv}{dt} \cong u\frac{\partial g}{\partial x}\left(x^*, y^*\right) + v\frac{\partial g}{\partial y}\left(x^*, y^*\right) \tag{11.27}$$

These two equations provide two linear ODEs for the dynamics of the perturbations, written in matrix form as

$$\begin{bmatrix} \frac{du}{dt} \\ \frac{dv}{dt} \end{bmatrix} = \mathbf{J}\begin{bmatrix} u \\ v \end{bmatrix}, \quad \text{where } \mathbf{J} = \text{Jacobian} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix}_{\left(x^*, y^*\right)} \tag{11.28}$$

The matrix $\mathbf{J}$ is called the Jacobian at the fixed point (x^*, y^*). This is the multivariable analog of the derivative $df(x^*)/dx$ for one-dimensional systems. The dynamics of this system can now be surmised using the Jacobian matrix at a fixed point. We could encounter all the types of fixed points shown in Figure 11.3.

Example

Consider the following system:

$$\begin{aligned}\frac{dx}{dt} &= x + e^{-y} = f(x,y)\\ \frac{dy}{dt} &= -y = g(x,y)\end{aligned} \tag{11.29}$$

This system has only one fixed point at $(x^*,y^*) = (-1,0)$. The Jacobian matrix is

$$\mathbf{J} = \begin{bmatrix}\frac{\partial f}{\partial x} & \frac{\partial f}{\partial y}\\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y}\end{bmatrix}_{(x^*,y^*)} = \begin{bmatrix}1 & -e^{-y}\\ 0 & -1\end{bmatrix}_{(-1,0)} = \begin{bmatrix}1 & -1\\ 0 & -1\end{bmatrix} \tag{11.30}$$

The eigenvalues of a 2-by-2 matrix are given by Equation 11.8. The Jacobian matrix given by Equation 11.30 thus has eigenvalues

$$\tau = 0, \quad \Delta = -1 \Rightarrow \begin{cases}\lambda_1 = 1\\ \lambda_2 = -1\end{cases} \Rightarrow \text{Saddle point}$$

Since the eigenvalues are real with one positive and one negative, linear stability analysis indicates that the fixed point is a saddle, as verified in the phase portrait shown in Figure 11.9.

11.5 PENDULUM

A damped pendulum with a constant applied torque is pictured in Figure 11.10.

The equation of motion of the pendulum is obtained from Newton's second law for a rotational body. Summing moments about the pivot point gives

$$I\alpha = \sum M = -b\frac{d\theta}{dt} - m \cdot g \cdot L \cdot \sin(\theta) + \Gamma \tag{11.31}$$

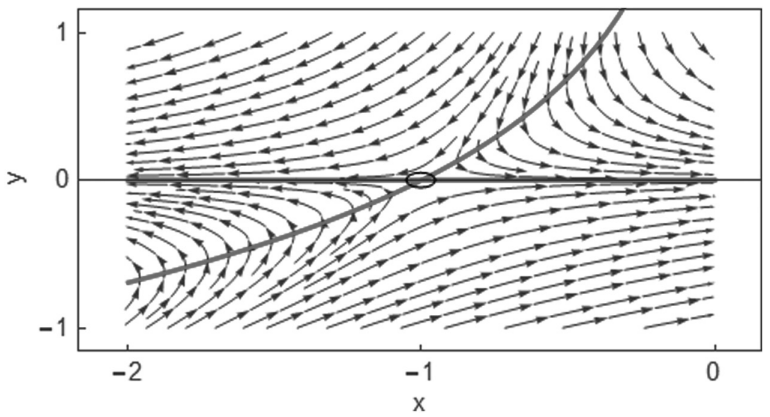


Figure 11.9 Phase portrait for Equations 11.29.

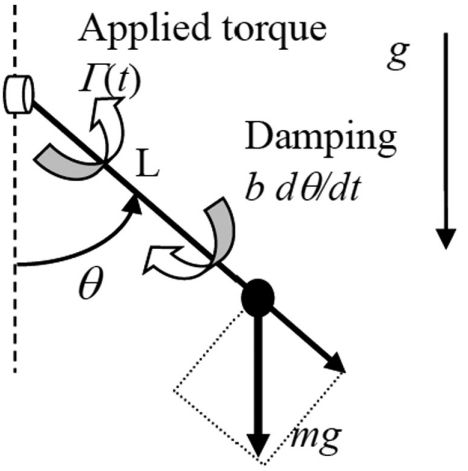


Figure 11.10 Schematic of a pendulum. m = mass (kg), $I = mL^2$ = mass moment of inertia, L = pendulum length (m), b = viscous damping coefficient (N·m·s), g = gravitational acceleration (m/s^2), and Γ = applied torque (N·m).

$$\alpha = \frac{d^2\theta}{dt^2} = \text{angular acceleration}$$

The resulting equation of motion is

$$L^2 m \frac{d^2\theta}{dt^2} + b\frac{d\theta}{dt} + m \cdot g \cdot L \cdot \sin(\theta) = \Gamma \tag{11.32}$$

Often, the small angle approximation, $\sin(\theta) \cong \theta$, is made, reducing the equation of motion to a linear system. How about the large angle behavior?

We next nondimensionalize Equation 11.32 using the procedure presented in Section 3.4. Choosing g, L, and m as reference quantities, the dimensionless equation of motion becomes

$$\frac{d^2\theta}{dt^{+2}} + B\frac{d\theta}{dt^{+}} + \sin(\theta) = \gamma \tag{11.33}$$

where the dimensionless parameters are

$$t^{+} = t\sqrt{\frac{g}{L}}, \quad B = \frac{b}{m \cdot g^{1/2} L^{3/2}}, \quad \gamma = \frac{\Gamma}{mgL} \tag{11.34}$$

The second-order system given by Equation 11.34 can be decomposed into the following equivalent system of first-order equations:

$$\begin{aligned} \frac{d\theta}{dt^{+}} &= \omega \\ \frac{dv}{dt^{+}} &= -B \cdot \omega - \sin(\theta) + \gamma \end{aligned} \tag{11.35}$$

11.5.1 Fixed Points: No Forcing, No Damping

For this restricted case, the fixed points of Equations 11.35 are at $(\theta^*,\omega^*)=(k\pi,0)$, where k is any integer. There's no difference between angles separated by multiples of 2π, so we can concentrate on the two fixed points (0,0) and $(\pi,0)$. The Jacobian is

$$\text{Jacobian} = \mathbf{J} = \begin{bmatrix} \dfrac{\partial f}{\partial \theta} & \dfrac{\partial f}{\partial \nu} \\ \dfrac{\partial g}{\partial \theta} & \dfrac{\partial g}{\partial \nu} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\cos(\theta) & 0 \end{bmatrix} \tag{11.36}$$

- At (0, 0), $\mathbf{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, $\tau = 0$, $\Delta = 1 > 0$: center
- At $(\pi, 0)$, $\mathbf{J} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\tau = 0$, $\Delta = -1 < 0$: saddle point

11.5.2 Fixed Points: General Case

The more general case has fixed points when

$$\begin{aligned} \nu^* &= 0 \\ \text{or } \arcsin(\gamma) &= \theta^* \end{aligned} \tag{11.37}$$

Note that the formal "*arcsin*" function on most $\sin(\theta^*)=\gamma$ computers is usually restricted between $-\pi/2$ and $\pi/2$. This equilibrium condition can be visualized graphically by plotting $\sin(\theta)$ and γ and noting the intersections. In Figure 11.10, we see that for $|\gamma|<1$, two fixed points exist. For $|\gamma|>1$, no fixed points exist, since the driving torque is so strong that it can never be balanced by gravity—the pendulum continually whirls over the top (Figure 11.11).

The fixed points are classified according to ranges of the driving torque parameter as summarized in Table 11.2.

Figure 11.12 shows a collection of phase portraits for different values of the dimensionless damping coefficient and driving torque. Figure 11.13 displays the solutions obtained with a numerical solver for some selected combinations of damping coefficient, driving torque, and initial conditions.

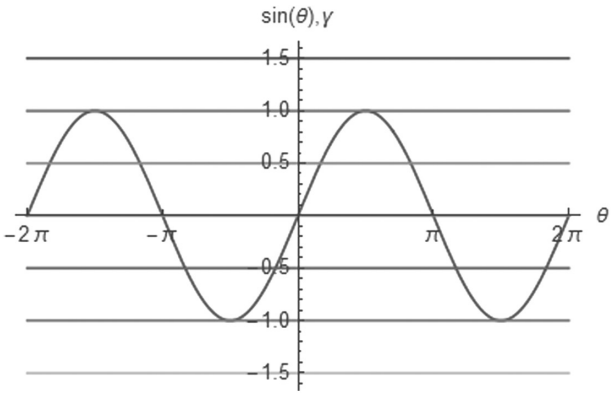


Figure 11.11 Graphical visualization of the fixed points of Equations 11.35.

Table 11.2 Classification of the stability of a pendulum

$\lvert\gamma\rvert < 1$		$\lvert\gamma\rvert = 1$	$\lvert\gamma\rvert > 1$
$\arcsin(\gamma) = \theta_1^*$	$\pi - \arcsin(\gamma) = \theta_2^*$	$\theta_1^* = \theta_1^* = \pi/2$	
$\tau = -B < 0$	$\tau = -B < 0$	Neutrally stable.	No fixed points. The driving torque is too strong.
$\Delta = \cos(\theta_1^*) > 0$	$\Delta = \cos(\theta_2^*) < 0$		
$B = 0 \rightarrow$ center $B < 2 \rightarrow$ stable spiral $B > 2 \rightarrow$ stable node	Saddle point		

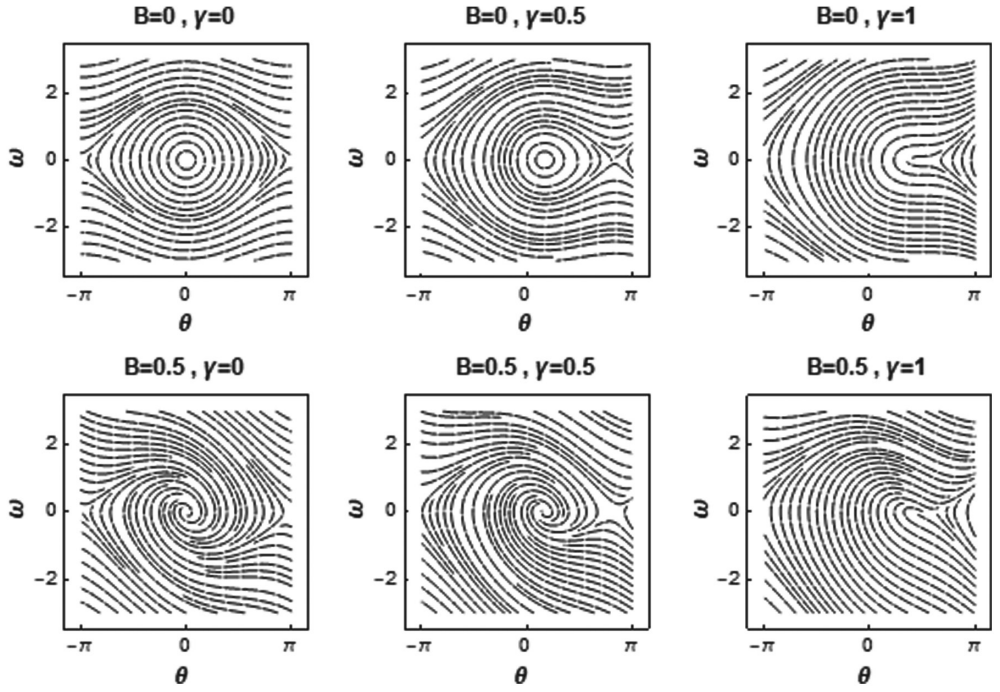


Figure 11.12 Phase portraits for the pendulum.

11.6 COMPETITION MODELS

Competition-type models of interacting populations are based on the idea that competing species do not directly kill each other but rather, hurt each other indirectly by competing for the same resources. The populations might be rabbits and sheep. One such model is the *Lotka–Volterra* competition model, which has the mathematical form

$$\begin{aligned} \frac{dx}{dt} &= a_1 x\left(1 - \frac{x}{K_1}\right) - b_1 \cdot x \cdot y \\ \frac{dy}{dt} &= a_2 y\left(1 - \frac{y}{K_2}\right) - b_2 \cdot x \cdot y \end{aligned} \tag{11.38}$$

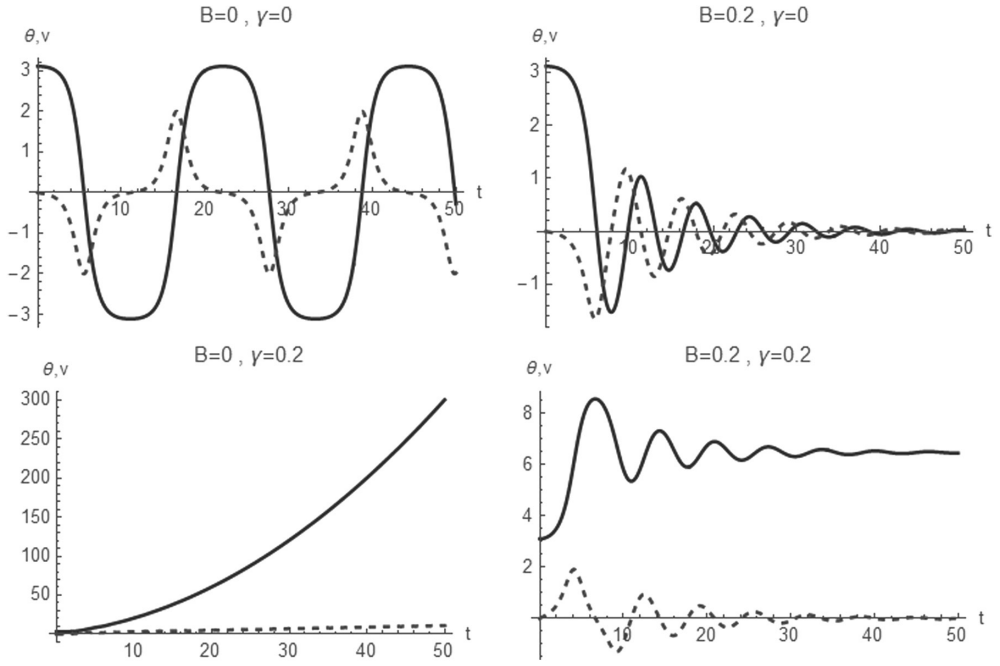


Figure 11.13 Dynamical responses of the pendulum. Position is the solid line, and velocity is the dashed line.

The following physical assumptions are incorporated into this model:

1. Each population in the absence of the other will grow according to the logistic model with growth rates a_1 and a_2 and carrying capacities K_1 and K_2.
2. The negative effect of competition for the same limited resources is modeled with a term proportional to the product of the populations.

If this model were applied to rabbits and sheep, the growth rate a and carrying capacity K for the rabbit population would be greater than the corresponding values for the sheep. Also, the competition parameter b would be larger for the rabbits than the sheep, since the sheep are larger and could simply push the rabbits out of their way.

11.6.1 Coexistence

Let's examine the dynamics of the following coupled system of equations:

$$\begin{aligned} \frac{dx}{dt} &= f(x,y) = x(3-2x-y) \\ \frac{dy}{dt} &= g(x,y) = y(2-x-y) \end{aligned} \tag{11.39}$$

In an attempt to understand the dynamics of this system, we will

- locate the fixed points
- investigate the stability of the fixed points
- draw the phase portrait

The fixed points are the (x^*,y^*) locations where

$$\begin{aligned} f\left(x^*,y^*\right) &= x^*\left(3-2x^*-y^*\right)=0 \Rightarrow x^*=0 \text{ or } \left(3-2x^*-y^*\right)=0 \\ g\left(x^*,y^*\right) &= y^*\left(2-x^*-y^*\right)=0 \Rightarrow y^*=0 \text{ or } \left(2-x^*-y^*\right)=0 \end{aligned} \tag{11.40}$$

Thus, there are four fixed points (x^*,y^*) at locations (0,0), (0,2), (1.5,0), and (1,1). Figure 11.14 clarifies the existence and locations of these four fixed points.

The two solid lines are the values where $dx/dt = f(x,y) = 0$, while the dashed lines are the locations where $dy/dt = g(x,y) = 0$. The intersections of the solid and dashed lines are the locations of the fixed points.

In order to analyze stability of the fixed points, we examine the Jacobian at each fixed point.

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \dfrac{\partial g}{\partial x} & \dfrac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} 3-4x-y & -x \\ -y & 2-x-2y \end{bmatrix} \tag{11.41}$$

Using linear stability analysis, the conclusions summarized in Table 11.3 can be drawn about the stability of this system.

The complete phase portrait, as shown in Figure 11.15, is beginning to emerge.

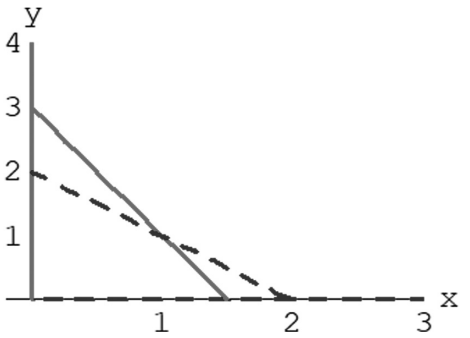


Figure 11.14 Visualization of the fixed points for Equations 11.39.

Table 11.3 Classification of the stability of the fixed points of Equations 11.39

$(x^*,y^*)=(0,0)$	$(x^*,y^*)=(3/2,0)$	$(x^*,y^*)=(0,2)$	$(x^*,y^*)=(1,1)$
$\mathbf{J}=\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$	$\mathbf{J}=\begin{bmatrix} 1 & 0 \\ -2 & -2 \end{bmatrix}$	$\mathbf{J}=\begin{bmatrix} -3 & -1.5 \\ 0 & 0.5 \end{bmatrix}$	$\mathbf{J}=\begin{bmatrix} -2 & -1 \\ -1 & -1 \end{bmatrix}$
$\lambda_1 = 3$	$\lambda_1 = 1$	$\lambda_1 = -3$	$\lambda_1 = -0.38$
$\lambda_2 = 2$	$\lambda_2 = -2$	$\lambda_2 = 0.5$	$\lambda_2 = -2.6$
Unstable node	**Saddle point**	**Saddle point**	**Stable node**

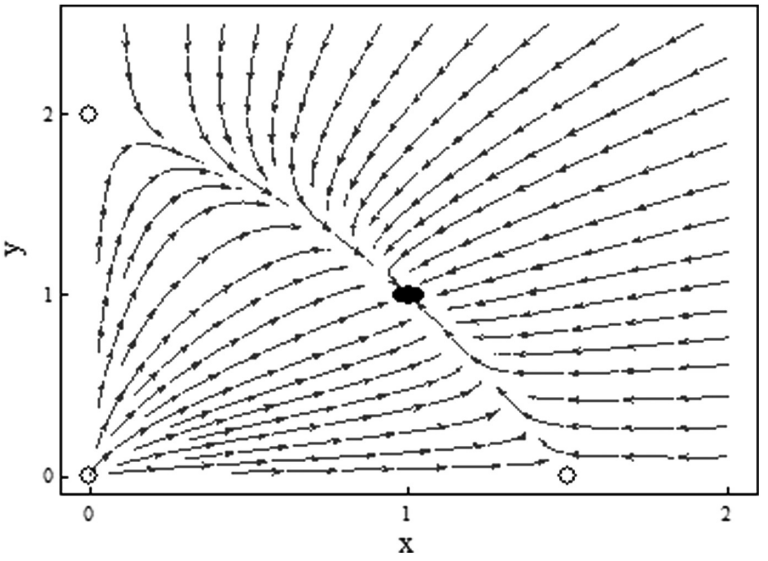


Figure 11.15 Phase portrait for Equations 11.39.

11.6.2 Extinction

Let's examine the dynamics of a system of coupled equations similar to the previous case but with slightly different coefficients.

$$\frac{dx}{dt} = x(3 - x - 2y), \quad \frac{dy}{dt} = y(2 - x - y) \tag{11.42}$$

A stability analysis on the fixed points yields

(x^*, y^*)	*Stability*
(0, 0)	Unstable node
(0, 2)	Stable node
(3, 0)	Stable node
(1, 1)	Saddle point

The phase portrait is shown in Figure 11.16.

For this type of competition model, one species drives the other to extinction. This is seen as *basins of attraction*. The basins for each species are separated by a ridge line or basin boundary. In the language of nonlinear dynamics, this boundary line is called the *stable manifold of the saddle point*.

11.7 LIMIT CYCLES

A limit cycle is an isolated closed trajectory. *Isolated* means that neighboring trajectories are not closed: they spiral either toward or away from the limit cycle. Figure 11.17 displays the notion of a stable, unstable, and half-stable limit cycle. They are characterized as:

- *Stable* or attracting limit cycle: all neighboring trajectories *approach* the limit cycle.
- *Unstable* limit cycle: all neighboring trajectories are repelled.
- *Half-stable* limit cycle.

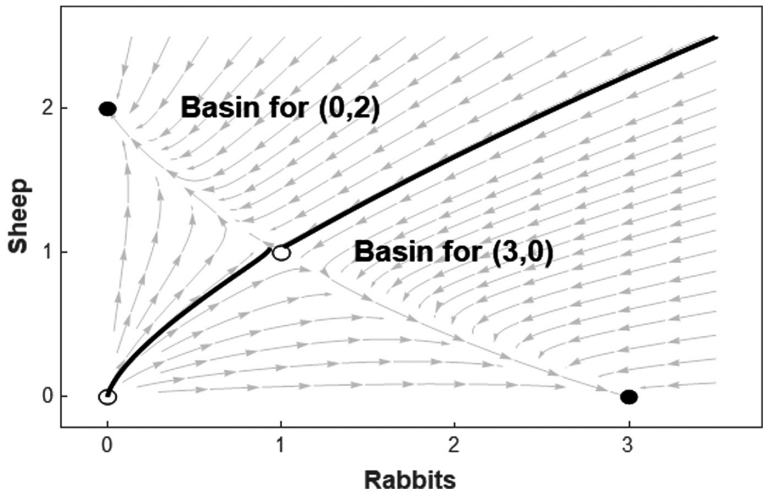


Figure 11.16 Phase portrait for Equations (11.42).

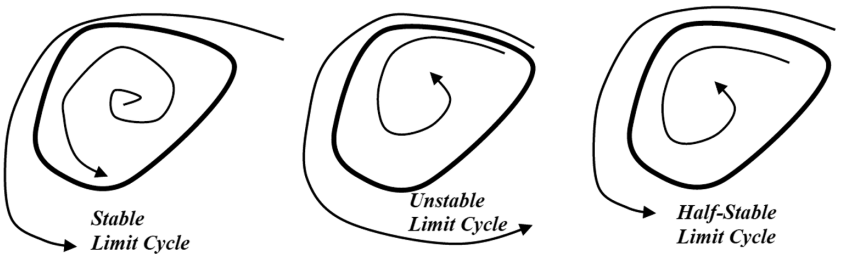


Figure 11.17 The idea of a limit cycle.

Applications of limit cycles include:

- Beating of a heart
- Periodic firing of a pacemaker neuron
- Daily rhythms in human body temperature and hormone secretion
- Chemical reactions that oscillate spontaneously
- Dangerous self-excitations in bridges and airplane wings

In each case, there is a standard oscillation of some period, amplitude, and waveform. If the system is perturbed, it returns to the standard oscillation. Limit cycles are inherently nonlinear phenomena: they cannot occur in a linear system. A linear system can have closed orbits, but they won't be isolated.

Using a simple example in cylindrical coordinates, it is easy to construct limit cycles. In the following system, the radial and angular dynamics are uncoupled and can be analyzed separately.

$$\begin{aligned} \frac{dr}{dt} &= r\left(1-r^2\right) \\ \frac{d\theta}{dt} &= 1 \end{aligned} \tag{11.43}$$

The angle is simply $\theta(t) = \theta_0 + t$. The radius has a stable fixed point at $r^* = 1$, as seen by the phase portrait in Figure 11.18. Thus, the dynamics are easy to predict—all trajectories except $r_0 = 0$ monotonically approach the circle $r^* = 1$. We verify this expectation by examining the solution in the phase plane. The conversion to rectangular coordinates is

$$\begin{aligned} x(t) &= r(t)\cos(\theta_0 + t) \\ y(t) &= r(t)\sin(\theta_0 + t) \end{aligned} \tag{11.44}$$

The following solutions are for several initial conditions. In all cases, the solutions tend toward a limit cycle on the circle $r^* = 1$ (Figure 11.19).

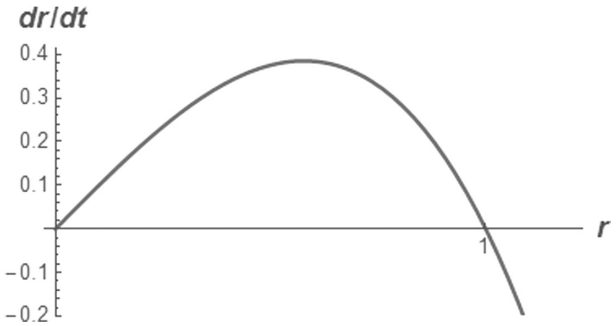


Figure 11.18 Phase portrait of Equation 11.43.

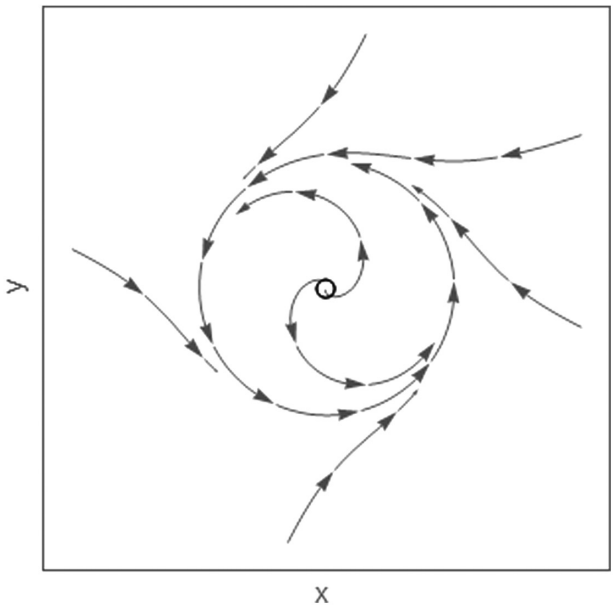


Figure 11.19 Phase portrait of Equation 11.44.

11.7.1 van der Pol Oscillator

A system that played a central role in the development of nonlinear dynamics is the van der Pol equation, given by

$$\frac{d^2x}{dt^2} + \mu\left(x^2 - 1\right)\frac{dx}{dt} + x = 0 \tag{11.45}$$

where $\mu > 0$ is a parameter. Historically, this equation arose in connection with the nonlinear electrical circuits used in the first radios.

This equation is similar to a simple harmonic oscillator but with a nonlinear damping term. That is,

- $|x| > 1$, normal positive damping
- $|x| < 1$, strange negative damping

Let's examine some solutions shown in Figure 11.20 for several values of the parameter μ.

For $\mu = 0$, the linear, undamped oscillator is recovered. For $\mu \gg 1$, the nonlinear damping term becomes strong. The limit cycle consists of an extremely slow buildup followed by a sudden discharge, followed by another slow buildup, and so on. Oscillations of this type are often called *relaxation oscillations*, because the "stress" accumulated during the slow buildup is "relaxed" during the sudden discharge.

11.7.2 Poincare–Bendixson Theorem

This theorem is one of only a few methods to establish that a closed orbit exists in a particular system. Suppose that:

1. $\mathfrak{R}$ is a closed, bounded subset of the plane.
2. $d\mathbf{x}/dt = \mathbf{f}(\mathbf{x})$ is continuously differentiable.

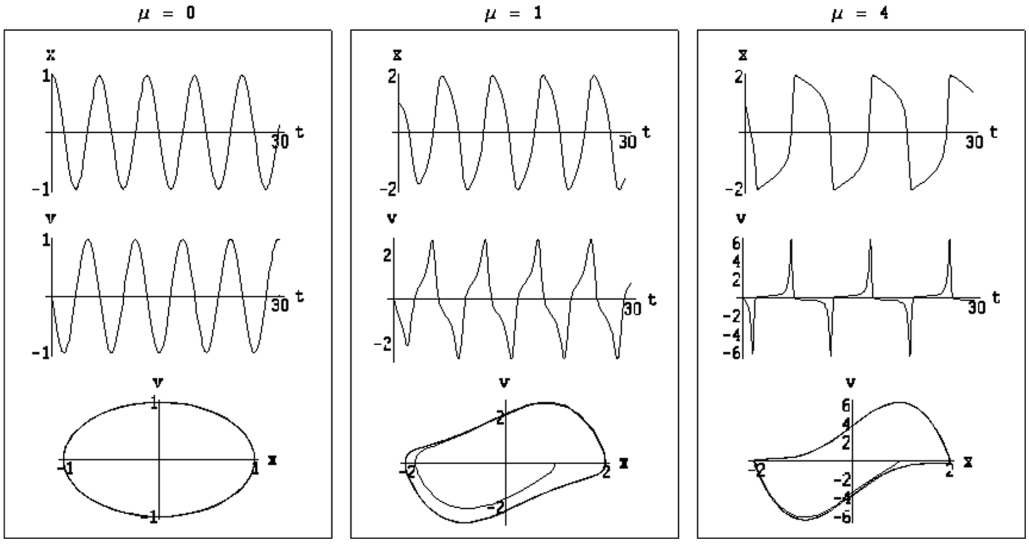


Figure 11.20 Solutions of the van der Pol equation.

3. $\mathfrak{R}$ contains no fixed points.
4. There exists a trajectory **C** that is confined in $\mathfrak{R}$.

Then, the theorem states that either **C** is a closed orbit or it spirals toward a closed orbit.

The Poincare–Bendixson theorem is one of the key results in nonlinear dynamics. It says that if a trajectory is confined to a closed, bounded region that contains no fixed points, then that trajectory must approach a limit cycle. Nothing more complex can occur—*chaos is not possible*.

11.8 BIFURCATIONS

The types of bifurcations found in first-order systems discussed in Section 10.4 have direct analogs in second-order systems as well as in all dimensions. Yet, it turns out that nothing really new happens when more dimensions are added:

- All the action is confined to a one-dimensional subspace along which bifurcations occur.
- In the extra dimensions, the flow is either simple attraction or repulsion from that subspace.

In the following sections, the various types of bifurcations are examined. You may explore further on your own using the Mathematic Demo: "A Tour of Second-Order Ordinary Differential Equations.nb." This demo is published in the Wolfram demo site at http://demonstrations.wolfram.com/ATourOfSecondOrderOrdinaryDifferentialEquations/.

This demo was developed to explore various types of behavior exhibited by second-order ordinary differential equations including linear systems, limit cycles, and bifurcations.

11.8.1 Saddle-Node Bifurcation

The saddle-node bifurcation is the basic mechanism for the creation and destruction of fixed points. The prototypical example is

$$\begin{aligned}\frac{dx}{dt} &= f(x,y) = \mu - x^2 \\ \frac{dy}{dt} &= g(x,y) = -y\end{aligned} \tag{11.46}$$

In the x-direction, one-dimensional bifurcation behavior occurs. In the y-direction, exponential decay occurs. Fixed points for exist when $\mu > 0$ at $\left(x^*, y^*\right) = \left(\pm\sqrt{\mu}, 0\right)$, while no real-valued fixed points exist when $\mu < 0$. Stability analysis of the fixed points:

$$\text{Jacobian} = \mathbf{J} = \begin{bmatrix} \partial f/\partial x & \partial f/\partial y \\ \partial g/\partial x & \partial g/\partial y \end{bmatrix}_{\left(x^*, y^*\right)} = \begin{bmatrix} -2x^* & 0 \\ 0 & -1 \end{bmatrix} \Rightarrow \begin{cases} \lambda_1 = -2x^* \\ \lambda_2 = -1 \end{cases} \tag{11.47}$$

The findings from our fixed point analysis are summarized in a logical manner in Table 11.4.

Table 11.4 Stability of the fixed points of a saddle-node bifurcation

	$\mu < \mu_c = 0$	$\mu > \mu_c = 0$
$\left(x^*, y^*\right) = \left(-\sqrt{\mu}, 0\right)$	Does not exist	Saddle point
$\left(x^*, y^*\right) = \left(\sqrt{\mu}, 0\right)$	Does not exist	Stable node

Note that $\mu = \mu_c = 0$ is a critical value of the parameter μ at which a bifurcation occurs. Phase portraits for the possible ranges of μ values are shown in Figure 11.21. The stable fixed point is shown as a solid disk, while the unstable saddle is shown as an open circle. Even after the fixed points have annihilated each other ($\mu < \mu_c$), they leave a ghost or bottleneck that sucks trajectories in and delays them before allowing passage out the other side.

11.8.2 Transcritical Bifurcation

The prototypical transcritical bifurcation example is

$$\frac{dx}{dt} = \mu \cdot x - x^2, \quad \frac{dy}{dt} = -y \tag{11.48}$$

As in the previous saddle-node case, the fixed points are identified, and a stability analysis is performed. The results are summarized in Table 11.5.

As the parameter μ crosses the critical value $\mu_c = 0$, a bifurcation occurs with the two fixed points switching stability (Figure 11.22).

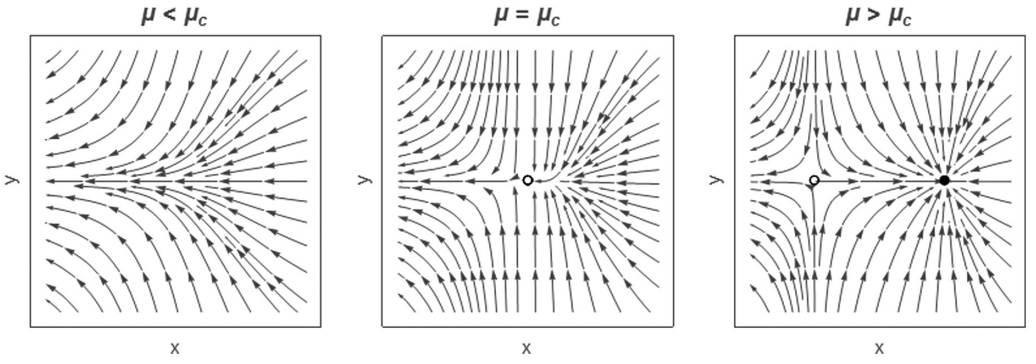


Figure 11.21 Phase portraits for the saddle-node bifurcation.

Table 11.5 Stability of the fixed points for a transcritical bifurcation

	$\mu < \mu_c = 0$	$\mu > \mu_c = 0$
$(x^*, y^*) = (0,0)$	Stable node	Saddle point
$(x^*, y^*) = (\mu, 0)$	Saddle point	Stable node

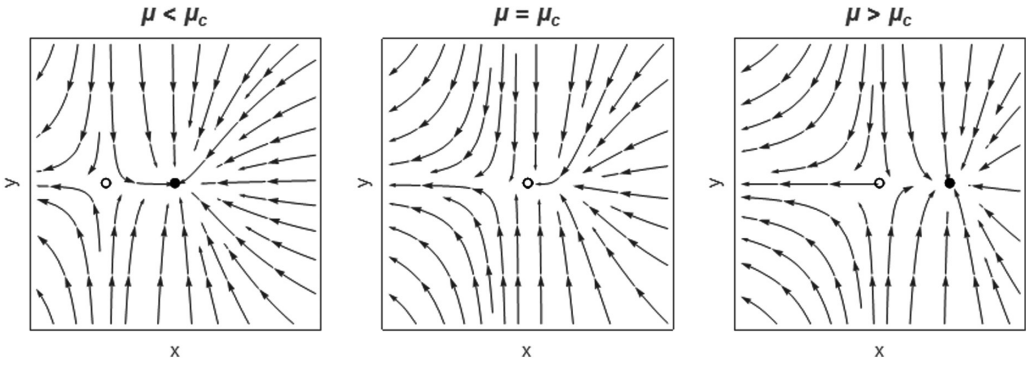


Figure 11.22 Phase portraits for the transcritical bifurcation.

11.8.3 Supercritical Pitchfork Bifurcation

The classical or prototypical supercritical pitchfork bifurcation example is

$$\frac{dx}{dt} = \mu \cdot x - x^3, \quad \frac{dy}{dt} = -y \tag{11.49}$$

The fixed points and their stability are summarized in Table 11.6 As the parameter μ drops below zero, a bifurcation occurs with the two stable fixed points at $\left(\pm\sqrt{\mu}, 0\right)$ coalescing into a stable fixed point at (0,0). This behavior is shown in Figure 11.23.

11.8.4 Subcritical Pitchfork Bifurcation

The classical or prototypical subcritical pitchfork bifurcation example is

$$\frac{dx}{dt} = \mu \cdot x + x^3, \quad \frac{dy}{dt} = -y \tag{11.50}$$

The fixed points and their stability are summarized in Table 11.7.

The bifurcation point is $\mu_c = 0$. For positive values of the parameter μ, the only fixed point is an unstable saddle at the origin. For negative values of μ, the origin becomes stable, and two unstable saddle points come into existence (Figure 11.24).

11.8.5 Hopf Bifurcations

For a second-order ODE, the linearized stability analysis from Section 11.4 reveals that the stability of a fixed point is determined by the eigenvalues of the Jacobian matrix. If the

Table 11.6 Stability of the fixed points for a supercritical pitchfork bifurcation

	$\mu < \mu_c = 0$	$\mu > \mu_c = 0$
$(x^*, y^*) = (0,0)$	Stable node	Saddle point
$\left(x^*, y^*\right) = \left(-\sqrt{\mu}, 0\right)$	Does not exist	Stable node
$\left(x^*, y^*\right) = \left(\sqrt{\mu}, 0\right)$	Does not exist	Stable node

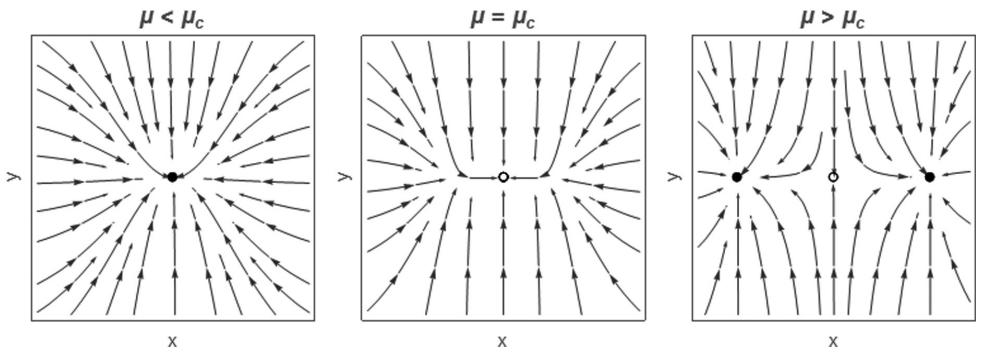


Figure 11.23 Phase portraits for the supercritical pitchfork bifurcation.

Table 11.7 Stability of the fixed points for a subcritical pitchfork bifurcation

	$\mu < \mu_c = 0$	$\mu > \mu_c = 0$
$(x^*, y^*) = (0,0)$	Stable node	Saddle point
$\left(x^*, y^*\right) = \left(-\sqrt{-\mu}, 0\right)$	Saddle point	Does not exist
$\left(x^*, y^*\right) = \left(\sqrt{-\mu}, 0\right)$	Saddle point	Does not exist

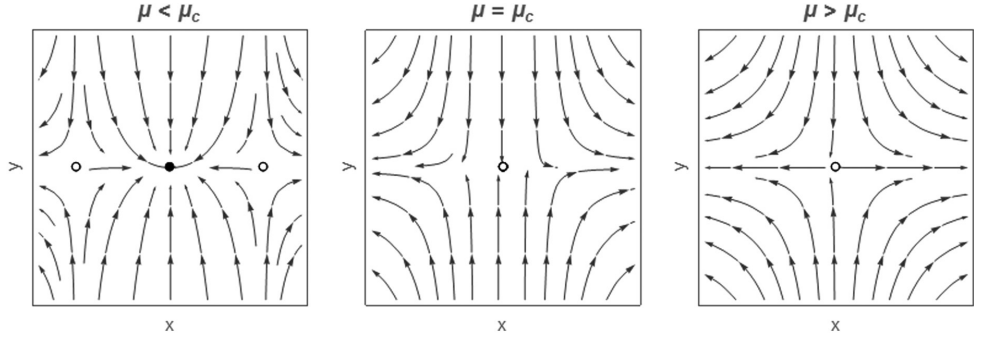


Figure 11.24 Phase portraits for the subcritical pitchfork bifurcation.

fixed point is stable, the real parts of both eigenvalues must be negative. Thus, the eigenvalues must both be negative real numbers or complex conjugates with negative real parts, as shown in Figure 11.25.

If the fixed point is to lose stability as a parameter changes, one or both of the eigenvalues must cross into the right half of the complex plane, where the real part is positive.

11.8.6 Supercritical Hopf Bifurcation

Here, a stable spiral changes into an unstable spiral surrounded by a small limit cycle. A mathematical example of such behavior is easiest to construct if we use polar coordinates. For instance, consider the following system (Figure 11.26).

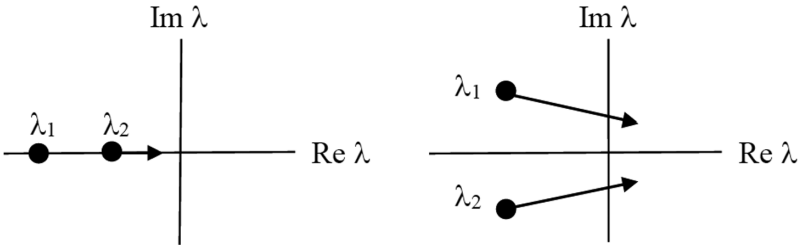


Figure 11.25 Real and imaginary parts of the eigenvalues.

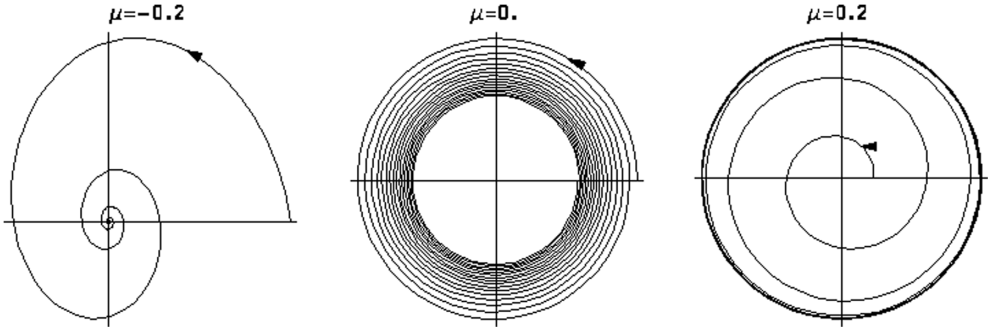


Figure 11.26 Supercritical Hopf bifurcation.

$$\frac{dr}{dt} = \mu \cdot r - r^3, \quad \frac{d\theta}{dt} = \omega + b \cdot r^2 \tag{11.51}$$

$\mu < 0$: Origin is a stable spiral.
$\mu = 0$: Origin is a half-stable spiral.
$\mu > 0$: Origin is an unstable spiral. A limit cycle exists at $r = \sqrt{\mu}$.

As an example, consider the system

$$\frac{dx}{dt} = \mu \cdot x - y + xy^2, \quad \frac{dy}{dt} = x + \mu \cdot y + y^3 \tag{11.52}$$

The origin is the only fixed point: $(x^*, y^*) = (0, 0)$. The Jacobian at the origin and eigenvalues are

$$\text{Jacobian} = \mathbf{J} = \begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix} \Rightarrow \begin{cases} \lambda_1 = \mu + i \\ \lambda_2 = \mu - i \end{cases} \tag{11.53}$$

These eigenvalues are complex conjugates. As μ crosses 0, a bifurcation occurs. The real part changes sign, characterizing the *Hopf* bifurcation. The origin changes from a stable spiral for $\mu < 0$ to an unstable spiral for $\mu > 0$, as shown in the accompanying phase portraits (Figure 11.27).

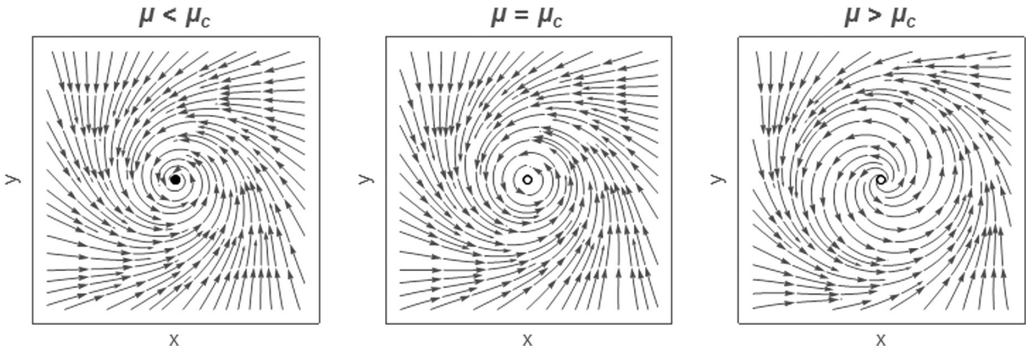


Figure 11.27 Phase portraits for the supercritical Hopf bifurcation.

Our other types of bifurcations—saddle, transcritical, and pitchfork—are characterized by real valued eigenvalues that change sign at some critical value of a system parameter.

11.8.7 Subcritical Hopf Bifurcation

The subcritical case is characterized by a jump to a distant attractor after bifurcation. This is much more dramatic than the supercritical Hopf bifurcation. An example is

$$\frac{dr}{dt} = \mu \cdot r + r^3 - r^5, \quad \frac{d\theta}{dt} = \omega + b \cdot r^2 \tag{11.54}$$

In this model, the cubic term is destabilizing, as it pushes trajectories away from the origin. The dynamics are restrained by the r^5 term.

11.9 COUPLED OSCILLATORS

Consider the coupled system

$$\begin{aligned} \frac{d\theta_1}{dt} &= f_1(\theta_1, \theta_2) \\ \frac{d\theta_2}{dt} &= f_2(\theta_1, \theta_2) \end{aligned} \tag{11.55}$$

The functions $f_1$and f_2 are periodic in (θ_1,θ_2). An example is

$$\begin{aligned} \frac{d\theta_1}{dt} &= \omega_1 - K_1 \sin(\theta_1 - \theta_2) \\ \frac{d\theta_2}{dt} &= \omega_2 + K_2 \sin(\theta_1 - \theta_2) \end{aligned} \tag{11.56}$$

where

θ_1,θ_2 are the *phases* of the oscillators
ω_1,ω_2 are the *natural frequencies* of the oscillators
K_1,K_2 are the *coupling coefficients* between the oscillators

Rather than work with each oscillator individually, we can derive a differential equation for the phase difference between the two oscillators:

$$\phi = \theta_1 - \theta_2 = \text{phase difference} \tag{11.57}$$

$$\frac{d\phi}{dt} = \frac{d\theta_1}{dt} - \frac{d\theta_2}{dt} = (\omega_1 - \omega_2) - (K_1 + K_2)\sin(\phi) \tag{11.58}$$

The fixed points are determined as follows.

$$\frac{d\phi}{dt} = 0 \Rightarrow \sin(\phi^*) = \frac{\omega_1 - \omega_2}{K_1 + K_2} \tag{11.59}$$

$$|\omega_1 - \omega_2| > K_1 + K_2 \Rightarrow \text{no fixed points}$$

$$|\omega_1 - \omega_2| \le K_1 + K_2 \Rightarrow \text{stable fixed point at } \phi^* = \text{ArcSine}\left(\frac{\omega_1 - \omega_2}{K_1 + K_2}\right)$$

Figure 11.28 shows the dynamics of this particular bifurcation.

Let's explore the conditions for the frequencies to become equal or "phase locked" at some value ω^*. Using Equations 11.56, we have

$$\begin{aligned} \frac{d\theta_1}{dt} &= \omega^* = \omega_1 - K_1 \sin(\phi^*) \\ \frac{d\theta_2}{dt} &= \omega^* = \omega_2 + K_2 \sin(\phi^*) \end{aligned} \tag{11.60}$$

Eliminate $\sin(\phi^*)$ from the previous equations to get

$$\omega^* = \frac{K_1\omega_2 + K_2\omega_1}{K_1 + K_2} \tag{11.61}$$

Thus, for $|\omega_1 - \omega_2| \le K_1 + K_2$, phase locking is possible.

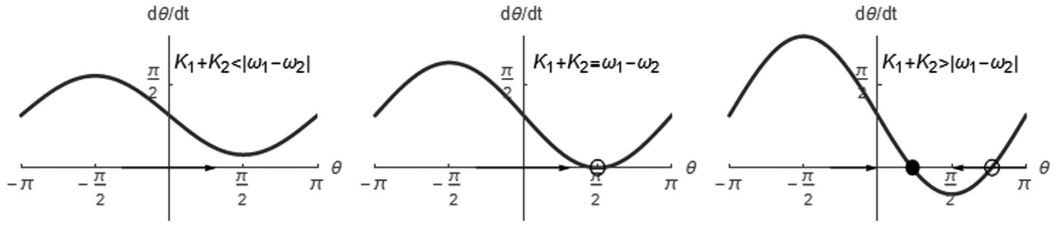


Figure 11.28 Phase portraits for coupled oscillators.

PROBLEMS: LINEAR SYSTEMS

Problem 11.1.1

The motion of a linear damped harmonic oscillator is described by

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + k \cdot x = 0$$

$$x = x_0, \quad \frac{dx}{dt} = v_0, \quad t = 0$$

where $c \geq 0$ is the damping coefficient.

a) Rewrite this equation as simultaneous first-order equations.
b) Classify the fixed point at the origin, and plot the phase portrait. Be sure to show all the various cases that can occur depending on the relative sizes of the parameters.
c) How do your results relate to standard notions of overdamped, critically damped, underdamped, and undamped vibrations?

Problem 11.1.2

Consider the following system of linear ODEs:

$$\frac{dx}{dt} = ay + b, \quad \frac{dy}{dt} = x$$

a) Determine the fixed points.
b) Determine the eigenvalues and the corresponding eigenvectors.
c) Classify the fixed point for various values of parameters a and b.
d) Sketch phase portraits showing all the qualitatively different behaviors.
e) Determine the solution $x(t)$ and $y(t)$ for $a = 0$, $x_0 = 0$, and $y_0 = 1$.

Problem 11.1.3

Consider the system of linear ODEs

$$\frac{dx}{dt} = a \cdot x + b \cdot y + p$$
$$\frac{dy}{dt} = c \cdot x + d \cdot y + q$$

This system is nonhomogeneous due to the presence of the terms p and q.

a) Determine the fixed points (x^*, y^*).
b) What variable change would transform this to a homogeneous system (one with p and q zero)?

c) Are the eigenvalues and stability of the system affected by the presence of the nonhomogeneous terms p and q? If your answer is yes, explain how.
d) Let's say that for a certain combination of parameters, the homogeneous problem is a center. With the presence of the nonhomogeneous terms p and q, the fixed point is determined to be $(x^*,y^*) = (1,0)$. Sketch the phase portrait.

Problem 11.1.4

Consider a series LRC circuit equation:

$$L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}\cdot I = 0$$

where L, $C > 0$ and $R \geq 0$.

a) Rewrite the equation as a two-dimensional linear system.
b) Show that the origin is asymptotically stable if $R > 0$ and neutrally stable if $R = 0$.
c) Classify the fixed point at the origin. Plot the phase portrait in all cases.

Problem 11.1.5

Consider the two-dimensional system of equations

$$\frac{dx}{dt} = -a(x-y), \quad \frac{dy}{dt} = a(x-y)$$

$$\left.\begin{aligned} x &= x_0 \\ y &= y_0 \end{aligned}\right\} \quad t = 0$$

a) Express the system in matrix form.
b) Determine the eigenvalues of this system.
c) Classify the behavior of this system.
d) Sketch the phase portrait.
e) Sketch a typical solution.

Problem 11.1.6

For each of the following cases:

a) Classify the stability of the fixed point based on the eigenvalues. Identify all the different possible types of behavior depending on the parameters a and b.
b) Determine the analytical solution using eigenvalues/eigenvectors. Carefully examine the case where sign(ab) = 0.
c) For each different type of behavior identified in part (a), plot the phase diagram. Also, plot the corresponding solution for the special case $R_0 = 1$, $J_0 = 0$.

Case 1: Jerry Springer

$$\frac{dR}{dt} = J, \quad \frac{dJ}{dt} = -R + J$$

Case 2: Out of Touch with Their Own Feelings

Suppose Romeo and Juliet are out of touch with their own feelings, such that they react to each other but not to themselves.

$$\frac{dR}{dt} = a \cdot J, \quad \frac{dJ}{dt} = b \cdot R$$

Case 3: Fire and Water

$$\frac{dR}{dt} = a \cdot R + b \cdot J, \quad \frac{dJ}{dt} = -b \cdot R - a \cdot J$$

Problem 11.1.7

Suppose we have a system of three linear, first-order ODEs. The eigenvalues and eigenvectors are

$\lambda_1 = 2 + i$	$\bar{V}_1 = [1, 0, 0]$
$\lambda_2 = 2 - i$	$\bar{V}_2 = [0, 1, 0]$
$\lambda_3 = -1$	$\bar{V}_3 = [0, 0, 1]$

Sketch the phase portraits in the x-y plane, the x-z plane, and the y-z plane.

Problem 11.1.8: Nonhomogeneous Linear System

Consider the linear system of linear ODEs

$$\frac{dx}{dt} = a \cdot x + b \cdot y + p$$
$$\frac{dy}{dt} = c \cdot x + d \cdot y + q$$

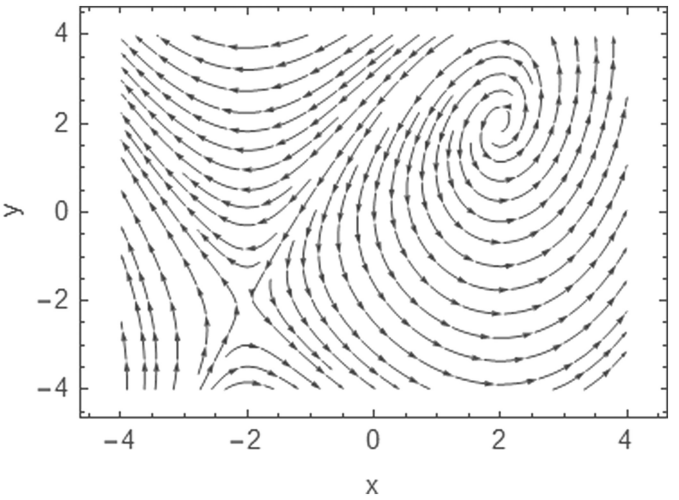
This system is nonhomogeneous due to the presence of the terms p and q.

a) Determine the fixed points (x^*, y^*).
b) What variable change would transform this to a homogeneous system (one with p and q zero)?
c) Are the eigenvalues and stability of the system affected by the presence of the nonhomogeneous terms p and q? If your answer is yes, explain how.
d) Let's say that for a certain combination of parameters, the homogeneous problem is a center. With the presence of the nonhomogeneous terms p and q, the fixed point is determined to be $(x^*, y^*) = (1.0)$. Sketch the phase portrait.

PROBLEMS: NONLINEAR SYSTEMS

Problem 11.2.1

The phase plot for a second-order system is shown.



a) Speculate on the location and stability of the fixed points.
b) Sketch solutions $x(t)$ and $y(t)$ starting from initial conditions $(x_0,y_0) = (1,1)$.

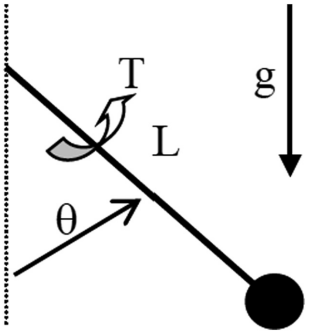
Problem 11.2.2

Consider the following nonlinear second-order systems:

$\frac{dx}{dt} = x - y$	$\frac{dx}{dt} = y^3 - 4x$
$\frac{dy}{dt} = x^2 - 4$	$\frac{dy}{dt} = y^3 - y - 3x$

a) Determine all the fixed points.
b) Classify the stability of the fixed points.
c) Sketch the phase portrait.
d) Sketch the solution starting from the initial conditions $x_0 = 1$, $y_0 = 0$ when $t = 0$.

Problem 11.2.3



Consider an undamped pendulum of mass m driven by a constant torque, T (N m).

a) Derive the equation of motion for the angular position, $\theta(t)$.
b) Change variables in order to express the equation of motion in the form

$$\frac{d^2\theta}{d\tau^2} + \sin\theta = \gamma$$

What are the variables τ and γ?

c) Find all the equilibrium points and classify them as γ varies.
d) Is the system conservative? If so, find a conserved quantity.
e) Using computer-generated solutions, plot phase portraits for various values of γ.
f) Find the approximate frequency of small oscillations about any centers in the phase portrait.

Problem 11.2.4

Consider the system

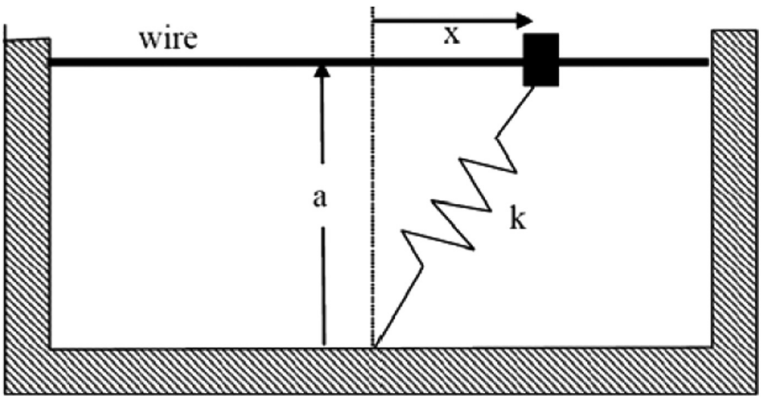
$$\frac{dx}{dt} = y - 2x$$

$$\frac{dy}{dt} = \mu + x^2 - y$$

a) Determine all the fixed points. What is the critical value μc at which a bifurcation occurs?
b) Classify the stability of the fixed points.
c) Sketch a phase portrait for $\mu < \mu c$ and one for $\mu > \mu c$.

Problem 11.2.5: Bead on a Wire

Consider a bead of mass m constrained to slide along a wire. Suppose that the motion is opposed by a viscous damping force bdx/dt.



x = coordinate along wire, measured from closest point to the spring support
m = mass
k = spring stiffness
L_0 = relaxed length
a = distance between support point and wire

a) Derive Newton's law for the motion of the bead.
b) Transform the equation of motion into the following dimensionless form:

$$\frac{d^2X}{d\tau^2} = -B \cdot \frac{dX}{d\tau} - X\left(1 - \frac{1}{\sqrt{A^2 + X^2}}\right)$$

Determine these dimensionless variables and parameters.
c) Find all possible equilibrium points (fixed points). Identify any bifurcations and determine critical parameter values.
d) Consider the highly damped case.
   1. Under what conditions can we neglect the acceleration term (second derivative term) compared with the damping term?
   2. Classify the stability of all the fixed points for this case.
   3. Plot the bifurcation diagram (X^* vs. A). What kind of bifurcation do we have?
   4. Plot phase portraits for $A > 1$, $A = 1$, and $A < 1$. Plot some solutions corresponding to each of these phase portraits.
e) Now consider the general case.
   1. Classify the stability of all the fixed points.
   2. Plot phase portraits $A > 1$, $A = 1$, and $A < 1$. Plot some solutions corresponding to each of these phase portraits. Create a set of plots for $B = 0$ and another set for $B = 0.1$.
f) Now consider an inclined wire.
   1. Derive Newton's law for the motion of the bead.
   2. Determine and classify all the fixed points for this case.

Problem 11.2.6

Consider the bead of mass m constrained to slide along a wire as described in Problem 11.2.5. A constant force f is applied, and the motion is opposed by a viscous damping force bdx/dt.

a) Derive Newton's law for the motion of the bead.
b) Transform the equation of motion into the following dimensionless form:

$$\frac{d^2X}{d\tau^2} = -B\frac{dX}{d\tau} - X\left(1 - \frac{1}{\sqrt{A^2 + X^2}}\right) + F$$

Determine these dimensionless variables and parameters.
c) For $F = 0$,
   1. Find and classify all possible fixed points.
   2. Plot the bifurcation diagram (X^* vs. A).
   3. Plot phase portraits for each qualitatively different behavior.

d) For any F,
   1. How do you find the fixed points? Create a function to determine them.
   2. Classify the fixed points.
   3. Plot bifurcation diagrams (X^* vs. A) for $F=0.01$, $F=0.1$, $F=0.5$, and $F=1$. Also, plot a three-dimensional bifurcation diagram showing X^* as a function of A and F.
   4. Plot phase portraits and some solutions for each qualitatively different behavior.

Problem 11.2.7

The Kermack–McKendrick (1927) model of an epidemic is

$$\frac{dx}{dt} = -kxy$$

$$\frac{dy}{dt} = kxy - ry$$

where

x = number of healthy population
y = number of sick population
r = death rate constant for sick people
k = infection rate constant

The equation for the deaths plays no role in the model and is omitted.

a) Find and classify all the fixed points.
b) Sketch the nullclines and vector field.
c) Find a conserved quantity for the system.
d) Plot the phase portrait. What happens as $t \to \infty$?
e) Let (x_o, y_o) be the initial condition. An epidemic is said to occur if $y(t)$ increases initially. Under what conditions does an epidemic occur?

Problem 11.2.8

Odell (1980) considered the system

$$\text{Prey:} \quad \frac{dx}{dt} = x\left(x(1-x) - y\right)$$

$$\text{Predator:} \quad \frac{dy}{dt} = y(x-a)$$

where

$x \geq 0$ is the dimensionless population of the prey
$y \geq 0$ is the dimensionless population of the predator
$a \geq 0$ is a control parameter

a) Plot the nullclines in the first quadrant $x, y \geq 0$.
b) Find the fixed points. Classify the stability of these fixed points.
c) Plot the phase portrait for $a > 1$, and show that the predators go extinct.

d) Show that a Hopf bifurcation occurs for at $a_c = 1/2$. Is it subcritical or supercritical?
e) Estimate the frequency of the limit cycle oscillations for a near the bifurcation.
f) Plot the topologically different phase portraits for $0 < a < 1$.

Problem 11.2.9

Consider the predator–prey model.

$$\text{Prey:} \quad \frac{dx}{dt} = x\left(b - x - \frac{y}{1+x}\right)$$

$$\text{Predator:} \quad \frac{dy}{dt} = y\left(\frac{x}{1+x} - ay\right)$$

where

$x, y \geq 0$ are the populations
$a, b > 0$ are parameters

a) Sketch the nullclines and discuss the bifurcations that occur as b varies.
b) Show that a positive fixed point $x^* > 0$, $y^* > 0$ exists for all $a, b > 0$. (Don't try to find the fixed point explicitly; use a graphical argument instead.)
c) Show that a Hopf bifurcation occurs at the positive fixed point if

$$a = a_c = \frac{4(b-2)}{b^2(b+2)}$$

and $b > 2$. (Hint: A necessary condition for a Hopf bifurcation to occur is $\tau = 0$, where τ is the trace of the Jacobian matrix at the fixed point. Show that $\tau = 0$ if and only if $2x^* = b - 2$. Then, use the fixed point conditions to express a in terms of x^*. Finally, substitute $x^* = (b - 2)/2$ into the expression for a, and you're done.)
d) Using a computer, check the validity of the expression in (c) and determine whether the bifurcation is subcritical or supercritical. Plot typical phase portraits above and below the Hopf bifurcation.

Problem 11.2.10

In Problem 10.15, an improved model of a laser was introduced.

$$\frac{dn}{dt} = GnN - kn$$

$$\frac{dN}{dt} = -GnN - fN + p$$

where

$N(t)$ = number of excited atoms
$n(t)$ = number of photons in laser field
G = gain coefficient for stimulated emission
k = decay rate of the photons
f = decay rate for spontaneous emission
p = pump strength

All parameters are positive.

a) Nondimensionalize the system. Use photon parameters as reference quantities.
b) Find the fixed points. Classify the fixed points using *linear stability* analysis.
c) Plot the stability diagram for the system. What types of bifurcation occur?
d) Plot all the qualitatively different phase portraits and transient responses.

Problem 11.2.11

A two-mode laser produces two different kinds of photons with numbers n_1 and n_2. By analogy with the simple laser model, the rate equations are

$$\frac{dn_1}{dt} = G_1 N n_1 - k_1 n_1$$

$$\frac{dn_2}{dt} = G_2 N n_2 - k_2 n_2$$

where
$N(t) = N_0 - \alpha_1 n_1 - \alpha_2 n_2$ is the number of excited atoms
n_1 and n_2 are the number of photons. The parameters G_1, G_2, k_1, Gk_2, α_1, α_2, and N_0 are all positive.

a) Find and classify all the fixed points.
b) Depending on the values of the parameters, how many qualitatively different phase portraits can occur? For each of the qualitatively different phase portraits, what does the model predict about the long-term behavior?

Problem 11.2.12

S = Average size of trees
E = Energy reserve; a measure of health
B = constant budworm population

$$\text{Trees:} \quad \frac{dS}{dt} = r_s S\left(1 - \frac{S}{K_s}\frac{K_E}{E}\right)$$

$$\text{Energy:} \quad \frac{dE}{dt} = r_E E\left(1 - \frac{E}{K_E}\right) - P\frac{B}{S}$$

a) Interpret the terms in the model biologically.
b) Nondimensionalize the system.
c) Plot the nullclines. Show fixed points. Identify the type of bifurcation that occurs.
d) Plot the phase portraits.

Problem 11.2.13

A (dumb) dog, instead of seeking to head off the duck, swims at constant speed in a circular pond straight at the (even dumber) duck. The duck, deciding not to fly, makes no other

attempt at escape from the dog beyond swimming (at constant speed) in a circle the radius of the pond.

Parametrically determine the path of the (previously admittedly dumb) dog. Develop a nondimensionalization that minimizes the number of parameters necessary for characterization of this solution.

Parameters:

- Pond radius, R, m
- Angular velocity of (again, very dumb) duck, ω, rad/s
- Speed of (dumb) dog, V_{dog}, m/s

Index