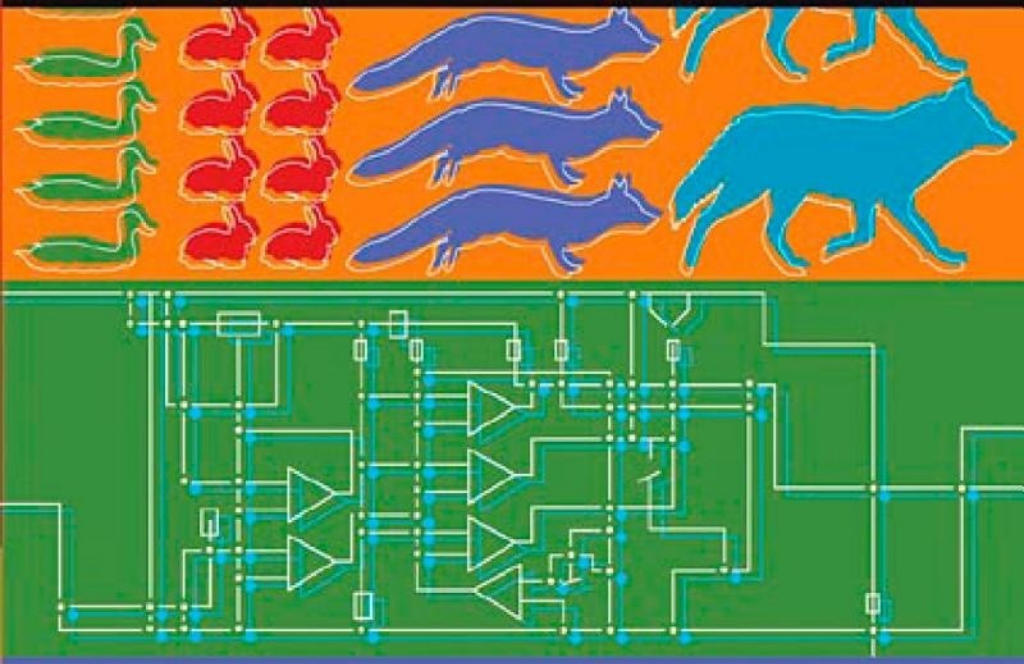
ELSEVIER
ACADEMIC
PRESS

Principles of Mathematical Modeling

SECOND EDITION



Clive L. Dym

Preface

My Goals For This Book

Science and engineering students depend heavily on concepts of mathematical modeling. In an age where almost everything is done on a computer, *it is my conviction that students of engineering and science are better served if they understand and "own" the underlying mathematics that the computers are doing on their behalf.* Mathematics is a *necessary* language for doing engineering and science. This will remain true no matter how good computation becomes. I repeatedly tell students that it is risky to accept computer calculations without having done some parallel closed-form modeling to benchmark the computer results. *Without such benchmarking and validation, how do we know that the computer isn't talking nonsense?* Finally, I find it satisfying and fun to do mathematical manipulations that explain how or why something happens, and to use mathematics to obtain corresponding numerical data or predictions.

Thus, as it was for the first edition, *my primary goal for this second edition remains to engage the reader in developing a foundation for mathematical modeling.* Further, knowing that mathematical models are built in a range of disciplines—including physics, biology, ecology, economics, sociology, military strategy, as well as all of the many branches of engineering—and knowing that mathematical modeling is comprised of a very diverse set of skills and tools, *I focused on techniques of particular interest to engineers, scientists, and others who model continuous systems.*

Features of This Edition

Aided by a variety of reviewers' comments and suggestions, this second edition features:

- A more formal statement of a principled approach to mathematical modeling (in Chapter 1). Ten principles are articulated and invoked as applications are developed, and each of them is identified by a key word (see below).
- Some 360 problems, many of which are designed to reinforce skills in mathematical manipulation. Many could be done with a computer algebra system (CAS), and there are others for which numerical programs could be used. However, given my goals for this book, I would ask students do the problems in "the old-fashioned way."
- A reordering and expansion of the applications chapters that reflects some sense of increasing complexity.
- Expanded figure captions that are intended to be more informative.

How This Book Is Organized

The book is organized into two parts: foundations and applications. The first part lays out the fundamental mathematical ideas of interest to the model builder: dimensional analysis, scaling, and elementary approximations of curves and functions. The applications part of the book develops a series of models and discusses their origins, their validity, and their meaning. These models include a host of exponential models, traffic flow models, free and forced vibration of linear (and occasionally nonlinear) oscillators, and optimization as done both with calculus and with elementary operations research techniques.

In the applications discussions, reference to the modeling principles is made by highlighting appropriate key words in the margin immediately adjacent to the appropriate text, as in:

Why? Find? "Lanchester wanted to describe the attrition of opposing forces at war. This required modeling the changes of two army populations whose respective rates of attrition depend on the size of the opposing army."

The foundations and applications parts of the book are connected only loosely. The following matrix indicates roughly how the chapters in each part relate to each other. In fact, the reader—and the teacher—can easily start with Chapter 5 and work through the applications models, referring back to corresponding discussions of the foundations as needed.

The problems distributed throughout and at the end of each chapter (save Chapter 1) are an integral part of the book. Like bike riding and dancing and designing, mathematical modeling cannot be learned simply by reading. Skills are developed and honed by doing problems, both elementary and difficult. Thus, there are problems that provide drills in basic skills, and there are problems that either develop new models or expand on models developed earlier in the text. For example, in problems at the end of Chapter 3 we show how dimensional groups are used to interpret experimental results. The problems in Chapter 5 demonstrate how dimensional analysis interacts with other approaches to deriving the governing equations for the oscillating pendulum, and the problems in Chapter 7 include data on resonance and impedance for a variety of forced oscillators.

		Models				
		5	6	7	8	9
Tools		Exponential Growth and Decay	Traffic Flow Models	Modeling Free Vibration	Applying Vibration Models	What Is the Best?
2	Dimensional Analysis	•	•	•		•
3	Scaling	•	•	•		
4	Approximation	•	•	•	•	•

As noted earlier, many of the problems could be done with a computer, whether a symbolic manipulator, a spreadsheet, or an algorithmic number cruncher. However, in order to learn to do mathematical modeling, the problems should be done in "closed form," with pencil and paper, with access only to a simple electronic calculator. This will both reinforce skills and provide a basis for benchmarking future computer calculations.

Three appendices from the first edition have been moved closer to their use in the book. A brief review of elementary transcendental functions is now appended to Chapter 4; the mathematics of the first-order equation, $dN/dt - \lambda N = 0$, is outlined in Section 5.2.2; and the mathematics of the second-order oscillator equation, $md^2x/dt^2 + kx = F(t)$, is detailed in Sections 7.2.2. and 8.6.

Lastly, the book can be used in several ways. The first edition was developed for new courses in mathematical modeling that were offered to first-year *engineering students* at Carnegie Mellon University and at the University of Massachusetts at Amherst. The book could also serve as a first course in applied mathematics for *mathematics majors*, or as a "technical elective" for various science and engineering majors, or conceivably as a supplementary text in basic calculus courses. In hopes of extending

its audience, I have tried to enhance both the book's accessibility and its flexibility.

I Presume That You, the Reader, Have. . .

. . . taken courses in elementary algebra, trigonometry, and first-year calculus. I further presume that you recognize what a differential equation is and what it means for $y(x)$ or $y(x, t)$ to be a solution of a differential equation. While you won't be asked to "solve" a differential equation, you will be asked to confirm and manipulate some of the solutions that are given. Finally, I do assume some basic understanding of first-year physics, mainly mechanics.

Acknowledgments

This book is the second edition of a text originally published in 1980 and written by me and Elizabeth S. Ivey; Betty was then both a professor of physics at Smith College and an adjunct professor of mechanical engineering at the University of Massachusetts at Amherst. When approached in the summer of 2000 by Academic Press to do a second edition, Betty and I decided that I would do the second edition alone, but I cannot view the second edition as complete without acknowledging the wonderful nature of our original collaboration and the long-standing friendship that resulted.

Many people deserve much credit for the good in this new version, although the responsibility for the bad (and the ugly) is entirely mine. Professors Robert L. Borrelli (Harvey Mudd College), Edward A. Connors (University of Massachusetts at Amherst), Ricardo Diaz (University of North Colorado), Michael Kirby (Colorado State University), Mark S. Korlie (Montclair State University), Thomas Seidman (University of Maryland Baltimore County), Caroline Smith (James Madison University), and William H. Wood (University of Maryland Baltimore County) were kind enough to provide pre-publication reviews of this second edition that were very helpful and supportive.

Professor Ewart Carson (University of London) provided a solid and helpful review of the first edition that prompted several changes in the second.

The artist Miriam Dym once again agreed to design the cover for one of her father's books, for which I am very grateful.

Drs. Ashkay Gupta and Ali Reza (Exponent FAA) provided a thoughtful review of the discussion of the vibration of a tall, slender building (Section 8.2).

Several Harvey Mudd colleagues provided helpful comments on various sections: Patrick Little (Engineering) of Chapter 9 on optimization; R. Erik Spjut (Engineering) of Section 9.5.1 on nucleation; Lisa M. Sullivan (Humanities and Social Sciences) of Section 5.5 on exponential modeling of money matters; and Harry E. Williams (Engineering), who commented on much of the early material in painstaking detail.

Dr. John H. McMasters (The Boeing Company) allowed me to use parts of an unpublished manuscript on geometric programming in constructing Sections 9.5.2 and 9.5.3, and graciously reviewed the final product.

Professor David Powers (Clarkson University) served once again as a reviewer of this book. In fact, Dr. Powers has also reviewed the prepublication prospectus of the first edition, provided a very useful re-review of that first edition that helped define the second edition, and then reviewed this manuscript during its preparation.

Professor Wilfred W. Recker (University of California, Irvine), a good friend for almost forty years, carefully reviewed the discussion of traffic flow modeling (Chapter 6) and provided some new materials.

Robert Ross, formerly of Academic Press, instigated this second edition, assisted ably by Mary Spencer. Subsequently, over the extended writing life of the project, Barbara Holland has been an encouraging and supportive editor, and she was assisted very ably by Tom Singer.

Professors Michael J. Scott (University of Illinois, Chicago) and William H. Wood (University of Maryland Baltimore County) commented insightfully on my adaptation (for Section 9.4) of our joint paper on choosing among alternatives.

Gautam Thatte, Harvey Mudd College '03, did a great job of researching and rendering illustrations.

Finally, Joan Dym has provided her usual friendship, love, and oft-needed distraction while this second edition was being written. It is a true pleasure to acknowledge what this has meant to me.

Clive L. Dym
Claremont, California
May 2004

תושלבע

1
What Is Mathematical Modeling?

We begin this book with a dictionary definition of the word *model*:

model *(n): a miniature representation of something; a pattern of something to be made; an example for imitation or emulation; a description or analogy used to help visualize something (e.g., an atom) that cannot be directly observed; a system of postulates, data and inferences presented as a mathematical description of an entity or state of affairs*

This definition suggests that *modeling* is an activity, a *cognitive activity* in which we think about and make models to describe how devices or objects of interest behave.

There are many ways in which devices and behaviors can be described. We can use words, drawings or sketches, physical models, computer programs, or mathematical formulas. In other words, the modeling activity can be done in several languages, often simultaneously. Since we are particularly interested in using the language of mathematics to make models,

we will refine the definition just given:

mathematical model *(n): a representation in mathematical terms of the behavior of real devices and objects*

We want to know how to make or generate mathematical representations or models, how to validate them, how to use them, and how and when their use is limited. But before delving into these important issues, it is worth talking about why we do mathematical modeling.

1.1 Why Do We Do Mathematical Modeling?

Since the modeling of devices and phenomena is essential to both engineering and science, engineers and scientists have very practical reasons for doing mathematical modeling. In addition, engineers, scientists, and mathematicians want to experience the sheer joy of formulating and solving mathematical problems.

1.1.1 Mathematical Modeling and the Scientific Method

In an elementary picture of the *scientific method* (see Figure 1.1), we identify a "real world" and a "conceptual world." The external world is the one we call real; here we observe various phenomena and behaviors, whether natural in origin or produced by artifacts. The conceptual world is the world of the mind—where we live when we try to understand what is going on in that real, external world. The conceptual world can be viewed as having three stages: observation, modeling, and prediction.

In the *observation* part of the scientific method we measure what is happening in the real world. Here we gather empirical evidence and "facts on the ground." Observations may be direct, as when we use our senses, or indirect, in which case some measurements are taken to indicate through some other reading that an event has taken place. For example, we often know a chemical reaction has taken place only by measuring the product of that reaction.

In this elementary view of how science is done, the *modeling* part is concerned with analyzing the above observations for one of (at least) three reasons. These rationales are about developing: *models that describe* the

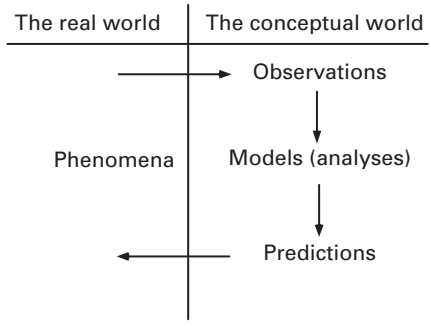


Figure 1.1 An elementary depiction of the *scientific method* that shows how our conceptual models of the world are related to observations made within that real world (Dym and Ivey, 1980).

behavior or results observed; *models that explain why* that behavior and results occurred as they did; or *models that allow us to predict* future behaviors or results that are as yet unseen or unmeasured.

In the *prediction* part of the scientific method we exercise our models to tell us what will happen in a yet-to-be-conducted experiment or in an anticipated set of events in the real world. These predictions are then followed by observations that serve either to validate the model or to suggest reasons that the model is inadequate.

The last point clearly points to the looping, iterative structure apparent in Figure 1.1. It also suggests that modeling is central to all of the conceptual phases in the elementary model of the scientific method. We build models and use them to predict events that can confirm or deny the models. In addition, we can also improve our gathering of empirical data when we use a model to obtain guidance about where to look.

1.1.2 Mathematical Modeling and the Practice of Engineering

Engineers are interested in *designing* devices and processes and systems. That is, beyond observing how the world works, engineers are interested in creating artifacts that have not yet come to life. As noted by Herbert A. Simon (in *The Sciences of the Artificial*), "Design is the distinguishing activity of engineering." Thus, engineers must be able to describe and analyze objects and devices into order to predict their behavior to see if

that behavior is what the engineers want. In short, engineers need to model devices and processes if they are going to design those devices and processes.

While the scientific method and engineering design have much in common, there are differences in motivation and approach that are worth mentioning. In the practices of science and of engineering design, models are often applied to predict what will happen in a future situation. In engineering design, however, the predictions are used in ways that have far different consequences than simply anticipating the outcome of an experiment. Every new building or airplane, for example, represents a model-based prediction that the building will stand or the airplane will fly without dire, unanticipated consequences. Thus, beyond simply validating a model, prediction in engineering design assumes that resources of time, imagination, and money can be invested with confidence because the predicted outcome will be a good one.

1.2 Principles of Mathematical Modeling

Mathematical modeling is a *principled* activity that has both principles behind it and methods that can be successfully applied. The principles are over-arching or *meta*-principles phrased as questions about the intentions and purposes of mathematical modeling. These meta-principles are almost philosophical in nature. We will now outline the principles, and in the next section we will briefly review some of the methods.

A visual portrayal of the basic philosophical approach is shown in Figure 1.2. These methodological modeling principles are also captured in the following list of questions and answers:

- **Why?** What are we looking for? Identify the need for the model.
- **Find?** What do we want to know? List the data we are seeking.
- **Given?** What do we know? Identify the available relevant data.
- **Assume?** What can we assume? Identify the circumstances that apply.
- **How?** How should we look at this model? Identify the governing physical principles.
- **Predict?** What will our model predict? Identify the equations that will be used, the calculations that will be made, and the answers that will result.
- **Valid?** Are the predictions valid? Identify tests that can be made to *validate* the model, i.e., is it consistent with its principles and assumptions?
- **Verified?** Are the predictions good? Identify tests that can be made to *verify* the model, i.e., is it useful in terms of the initial reason it was done?

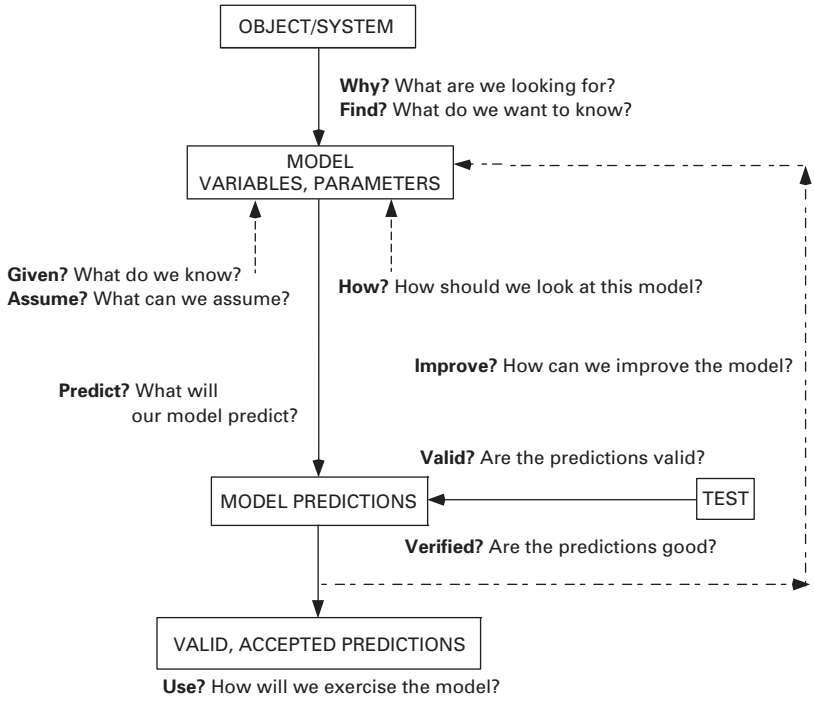


Figure 1.2 A first-order view of *mathematical modeling* that shows how the questions asked in a principled approach to building a model relate to the development of that model (inspired by Carson and Cobelli, 2001).

- **Improve?** Can we improve the model? Identify parameter values that are not adequately known, variables that should have been included, and/or assumptions/restrictions that could be lifted. Implement the iterative loop that we can call "model-validate-verify-improve-predict."
- **Use?** How will we exercise the model? What will we do with the model?

This list of questions and instructions is *not* an algorithm for building a good mathematical model. However, the underlying ideas are key to mathematical modeling, as they are key to problem formulation generally. Thus, we should expect the individual questions to recur often during the modeling process, and we should regard this list as a fairly general approach to *ways of thinking* about mathematical modeling.

Having a clear picture of why the model is wanted or needed is of prime importance to the model-building enterprise. Suppose we want to estimate how much power could be generated by a dam on a large river, say a dam located at The Three Gorges on the Yangtze River in Hubei Province in the People's Republic of China. For a first estimate of the available power, we

wouldn't need to model the dam's thickness or the strength of its foundation. Its height, on the other hand, would be an essential parameter of a power model, as would some model and estimates of river flow quantities. If, on the other hand, we want to design the actual dam, we would need a model that incorporates all of the dam's physical characteristics (e.g., dimensions, materials, foundations) and relates them to the dam site and the river flow conditions. Thus, defining the task is the first essential step in model formulation.

We then should list what we know—for example, river flow quantities and desired power levels—as a basis for listing the variables or parameters that are as yet unknown. We should also list any relevant assumptions. For example, levels of desired power may be linked to demographic or economic data, so any assumptions made about population and economic growth should be spelled out. Assumptions about the consistency of river flows and the statistics of flooding should also be spelled out.

Which physical principles apply to this model? The mass of the river's water must be conserved, as must its momentum, as the river flows, and energy is both dissipated and redirected as water is allowed to flow through turbines in the dam (and hopefully not spill over the top!). And mass must be conserved, within some undefined system boundary, because dams do accumulate water mass from flowing rivers. There are well-known equations that correspond to these physical principles. They could be used to develop an estimate of dam height as a function of power desired. We can validate the model by ensuring that our equations and calculated results have the proper dimensions, and we can exercise the model against data from existing hydroelectric dams to get empirical data and validation.

If we find that our model is inadequate or that it fails in some way, we then enter an *iterative loop* in which we cycle back to an earlier stage of the model building and re-examine our assumptions, our known parameter values, the principles chosen, the equations used, the means of calculation, and so on. This iterative process is essential because it is the only way that models can be improved, corrected, and validated.

1.3 Some Methods of Mathematical Modeling

Now we will review some of the mathematical techniques we can use to help answer the philosophical questions posed in Section 1.2. These mathematical principles include: dimensional homogeneity, abstraction and scaling,

conservation and balance principles, and consequences of linearity. We will expand these themes more extensively in the first part of this book.

1.3.1 Dimensional Homogeneity and Consistency

There is a basic, yet very powerful idea that is central to mathematical modeling, namely, that every equation we use must be *dimensionally homogeneous* or *dimensionally consistent*. It is quite logical that every term in an energy equation has total dimensions of energy, and that every term in a balance of mass should have the dimensions of mass. This statement provides the basis for a technique called *dimensional analysis* that we will discuss in greater detail in Chapter 2.

In that discussion we will also review the important distinction between physical *dimensions* that relate a (derived) quantity to fundamental physical quantities and *units* that are numerical expressions of a quantity's dimensions expressed in terms of a given physical standard.

1.3.2 Abstraction and Scaling

An important decision in modeling is choosing an appropriate level of detail for the problem at hand, and thus knowing what level of detail is prescribed for the attendant model. This process is called *abstraction* and it typically requires a thoughtful approach to identifying those phenomena on which we want to focus, that is, to answering the fundamental question about why a model is being sought or developed.

For example, a linear elastic spring can be used to model more than just the relation between force and relative extension of a simple coiled spring, as in an old-fashioned butcher's scale or an automobile spring. It can also be used to model the static and dynamic behavior of a tall building, perhaps to model wind loading, perhaps as part of analyzing how the building would respond to an earthquake. In these examples, we can use a very abstract model by subsuming various details within the parameters of that model. We will explore these issues further in Chapter 3.

In addition, as we talk about finding the right level of abstraction or the right level of detail, we are simultaneously talking about finding the right *scale* for the model we are developing. For example, the spring can be used at a much smaller, *micro* scale to model atomic bonds, in contrast with the *macro* level for buildings. The notion of scaling includes several ideas, including the effects of geometry on scale, the relationship of function to scale, and the role of size in determining limits—all of which are needed to choose the right scale for a model in relation to the "reality" we want to capture.

1.3.3 Conservation and Balance Principles

When we develop mathematical models, we often start with statements that indicate that some property of an object or system is being conserved. For example, we could analyze the motion of a body moving on an ideal, frictionless path by noting that its energy is *conserved*. Sometimes, as when we model the population of an animal colony or the volume of a river flow, we must *balance* quantities, of individual animals or water volumes, that cross a defined boundary. We will apply *balance* or *conservation principles* to assess the effect of maintaining or conserving levels of important physical properties. Conservation and balance equations are related—in fact, conservation laws are special cases of balance laws.

The mathematics of balance and conservation laws are straightforward at this level of abstraction. Denoting the physical property being monitored as $Q(t)$ and the independent variable time as t, we can write a balance law for the *temporal* or time rate of change of that property within the system boundary depicted in Figure 1.3 as:

$$\frac{dQ(t)}{dt} = q_{in}(t) + g(t) - q_{out}(t) - c(t), \tag{1.1}$$

where $q_{in}(t)$ and $q_{out}(t)$ represent the flow rates of $Q(t)$ into (the *influx*) and out of (the *efflux*) the system boundary, $g(t)$ is the rate at which Q is generated within the boundary, and $c(t)$ is the rate at which Q is consumed within that boundary. Note that eq. (1.1) is also called a *rate equation* because each term has both the meaning and dimensions of the rate of change with time of the quantity $Q(t)$.

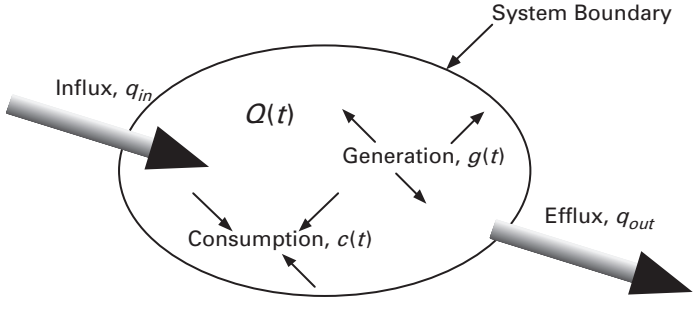


Figure 1.3 A system boundary surrounding the object or system being modeled. The influx $q_{in}(t)$, efflux $q_{out}(t)$, generation $g(t)$, and consumption $c(t)$, affect the rate at which the property of interest, $Q(t)$, accumulates within the boundary (after Cha, Rosenberg, and Dym, 2000).

In those cases where there is no generation and no consumption within the system boundary (i.e., when $g = c = 0$), the balance law in eq. (1.1) becomes a conservation law:

$$\frac{dQ(t)}{dt} = q_{in}(t) - q_{out}(t). \tag{1.2}$$

Here, then, the rate at which $Q(t)$ accumulates within the boundary is equal to the difference between the influx, $q_{in}(t)$, and the efflux, $q_{out}(t)$.

1.3.4 Constructing Linear Models

Linearity is one of the most important concepts in mathematical modeling. Models of devices or systems are said to be *linear* when their basic equations—whether algebraic, differential, or integral—are such that the magnitude of their behavior or response produced is *directly proportional* to the excitation or input that drives them. Even when devices like the pendulum discussed in Chapter 7 are more fully described by nonlinear models, their behavior can often be approximated by linearized or perturbed models, in which cases the mathematics of linear systems can be successfully applied.

We apply linearity when we model the behavior of a device or system that is forced or pushed by a complex set of inputs or excitations. We obtain the response of that device or system to the sum of the individual inputs by adding or *superposing* the separate responses of the system to each individual input. This important result is called the *principle of superposition.* Engineers use this principle to predict the response of a system to a complicated input by decomposing or breaking down that input into a set of simpler inputs that produce known system responses or behaviors.

1.4 Summary

In this chapter we have provided an overview of the foundational material we will cover in this book. In so doing, we have defined mathematical modeling, provided motivation for its use in engineering and science, and set out a principled approach to doing mathematical modeling. We have also outlined some of the important tools that will be covered in greater detail later: dimensional analysis, abstraction and scaling, balance laws, and linearity.

It is most important to remember that mathematical models are representations or descriptions of reality—by their very nature they *depict reality.* Thus, we close with a quote from a noted linguist (and former senator from

California) to remind ourselves that we are dealing with models that, we hope, represent something that seems real and relevant to us. However, they are abstractions and models, they are themselves real only as models, and they should never be confused with the reality we are trying to model. Thus, if the behavior predicted by our models does not reflect what we see or measure in the real world, it is the models that need to be fixed—and not the world:

"*The symbol is NOT the thing symbolized; the word is NOT the thing; the map is NOT the territory it stands for.*"

—S. I. Hayakawa, *Language in Thought and Action*

1.5 References

E. Carson and C. Cobelli (Eds.), *Modelling Methodology for Physiology and Medicine*, Academic Press, San Diego, CA, 2001.

P. D. Cha, J. J. Rosenberg, and C. L. Dym, *Fundamentals of Modeling and Analyzing Engineering Systems*, Cambridge University Press, New York, 2000.

C. L. Dym, *Engineering Design: A Synthesis of Views*, Cambridge University Press, New York, 1994.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

S. I. Hayakawa, *Language in Thought and Action*, Harcourt, Brace, New York, 1949.

G. Kemeny, *A Philosopher Looks at Science*, Van Nostrand-Reinhold, New York, 1959.

H. J. Miser, "Introducing Operational Research," *Operational Research Quarterly*, **27**(3), 665–670, 1976.

M. F. Rubinstein, *Patterns of Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

H. A. Simon, *The Sciences of the Artificial*, 3rd Edition, MIT Press, Cambridge, MA, 1999.

2
Dimensional Analysis

We begin this chapter, the first of three dealing with the *tools* or techniques for mathematical modeling, with H. L. Langhaar's definition of *dimensional analysis*:

Dimensional analysis *(n): a method by which we deduce information about a phenomenon from the single premise that the phenomenon can be described by a dimensionally correct equation among certain variables.*

This quote expresses the simple, yet powerful idea that we introduced in Section 1.3.1: all of the terms in our equations must be *dimensionally consistent*, that is, each separate term in those equations must have the same net physical dimensions. For example, when summing forces to ensure equilibrium, every term in the equation must have the physical dimension of force. (Equations that are dimensionally consistent are sometimes called *rational.*) This idea is particularly useful for validating newly developed mathematical models or for confirming formulas and equations before doing calculations with them. However, it is also a weak statement because the available tools of dimensional analysis are rather limited, and applying them does not always produce desirable results.

2.1 Dimensions and Units

The physical quantities we use to model objects or systems represent *concepts*, such as time, length, and mass, to which we also attach *numerical* values or measurements. Thus, we could describe the width of a soccer field by saying that it is 60 meters wide. The concept or abstraction invoked is length or distance, and the numerical measure is 60 meters. The numerical measure implies a comparison with a standard that enables both communication about and comparison of objects or phenomena without their being in the same place. In other words, common measures provide a frame of reference for making comparisons. Thus, soccer fields are wider than American football fields since the latter are only 49 meters wide.

The physical quantities used to describe or model a problem come in two varieties. They are either *fundamental* or *primary* quantities, or they are *derived* quantities. Taking a quantity as fundamental means only that we can assign it a standard of measurement independent of that chosen for the other fundamental quantities. In mechanical problems, for example, mass, length, and time are generally taken as the fundamental mechanical variables, while force is derived from Newton's law of motion. It is equally correct to take force, length, and time as fundamental, and to derive mass from Newton's law. For any given problem, of course, we need enough fundamental quantities to express each derived quantity in terms of these primary quantities.

While we relate primary quantities to standards, we also note that they are chosen arbitrarily, while derived quantities are chosen to satisfy physical laws or relevant definitions. For example, length and time are fundamental quantities in mechanics problems, and speed is a derived quantity expressed as length per unit time. If we chose time and speed as primary quantities, the derived quantity of length would be (speed $\times$ time), and the derived quantity of area would be (speed $\times$ time)^2.

The word *dimension* is used to relate a derived quantity to the fundamental quantities selected for a particular model. If mass, length, and time are chosen as primary quantities, then the dimensions of area are (length)^2, of mass density are mass/(length)^3, and of force are (mass $\times$ length)/(time)^2. We also introduce the notation of brackets [] to read as "the dimensions of." If M, L, and T stand for mass, length, and time, respectively, then:

$$[A = \text{area}] = (\mathsf{L})^2, \tag{2.1a}$$

$$[\rho = \text{density}] = \mathsf{M}/(\mathsf{L})^3, \tag{2.1b}$$

$$[F = \text{force}] = (\mathsf{M} \times \mathsf{L})/(\mathsf{T})^2. \tag{2.1c}$$

The *units* of a quantity are the numerical aspects of a quantity's dimensions expressed in terms of a given physical standard. By definition, then, a unit is an arbitrary multiple or fraction of that standard. The most widely accepted international standard for measuring length is the meter (m), but it can also be measured in units of centimeters (1 cm = 0.01 m) or of feet (0.3048 m). The magnitude or size of the attached number obviously depends on the unit chosen, and this dependence often suggests a choice of units to facilitate calculation or communication. The soccer field width can be said to be 60 m, 6000 cm, or (approximately) 197 feet.

Dimensions and units are related by the fact that identifying a quantity's dimensions allows us to compute its numerical measures in different sets of units, as we just did for the soccer field width. Since the physical dimensions of a quantity are the same, there must exist numerical relationships between the different systems of units used to measure the amounts of that quantity. For example,

$$1 \text{ foot (ft)} \cong 30.48 \text{ centimeters (cm)},$$

$$1 \text{ centimeter (cm)} \cong 0.000006214 \text{ miles (mi)},$$

$$1 \text{ hour (hr)} = 60 \text{ minutes (min)} = 3600 \text{ seconds (sec } \textit{or}\text{ s)}.$$

This equality of units for a given dimension allows us to change or convert units with a straightforward calculation. For a speed of 65 miles per hour (mph), for example, we can calculate the following equivalent:

$$65\frac{\text{mi}}{\text{hr}} = 65\frac{\text{mi}}{\text{hr}} \times 5280\frac{\text{ft}}{\text{mi}} \times 0.3048\frac{\text{m}}{\text{ft}} \times 0.001\frac{\text{km}}{\text{m}} \cong 104.6\frac{\text{km}}{\text{hr}}.$$

Each of the multipliers in this conversion equation has an effective value of unity because of the equivalencies of the various units, that is, 1 mi = 5280 ft, and so on. This, in turn, follows from the fact that the numerator and denominator of each of the above multipliers have the same physical dimensions. We will discuss *systems of units* and provide some conversion data in Section 2.4.

2.2 Dimensional Homogeneity

We had previously defined a *rational equation* as an equation in which each independent term has the same net dimensions. Then, taken in its entirety, the equation is *dimensionally homogeneous*. Simply put, we cannot add length to area in the same equation, or mass to time, or charge to stiffness—although we can add (and with great care) quantities having the same dimensions but expressed in different units, e.g., length in meters

and length in feet. The fact that equations must be rational in terms of their dimensions is central to modeling because it is one of the best—and easiest—checks to make to determine whether a model makes sense, has been correctly derived, or even correctly copied!

We should remember that a dimensionally homogeneous equation is independent of the units of measurement being used. However, we can create unit-dependent versions of such equations because they may be more convenient for doing repeated calculations or as a memory aid. In an example familiar from mechanics, the period (or cycle time), T_0, of a pendulum undergoing small-angle oscillations can be written in terms of the pendulum's length, l, and the acceleration of gravity, g:

$$T_0 = 2\pi\sqrt{l/g}. \tag{2.2}$$

This dimensionally homogeneous equation is independent of the system of units chosen to measure length and time. On the other hand, we may find it convenient to work in the metric system, in which case $g = 9.8\,\mathrm{m/s^2}$, from which it follows that

$$T_0(s) = 2\pi\sqrt{l/9.8} \cong 2\sqrt{l}. \tag{2.3}$$

Equation (2.3) is valid *only* when the pendulum's length is measured in meters. In the so-called British system,[1] where $g = 32.17\,\mathrm{ft/sec^2}$,

$$T_0(\mathrm{sec}) = 2\pi\sqrt{l/32.17} \cong 1.1\sqrt{l}. \tag{2.4}$$

Why not?

Equations (2.3) and (2.4) are not dimensionally homogeneous. So, while these formulas may be appealing or elegant, we have to remember their limited ranges of validity, as we should whenever we use or create similar formulas for whatever modeling we are doing.

2.3 Why Do We Do Dimensional Analysis?

We presented a definition of dimensional analysis at the beginning of this chapter, where we also noted that the "method" so defined has both powerful implications—rational equations and dimensional consistency—and severe limitations—the limited nature of the available tools. Given this limitation, why has this method or technique developed, and why has it persisted?

[1] One of my Harvey Mudd colleagues puckishly suggests that we should call this the American system of units as we are, apparently, the only country still so attached to feet and pounds.

Figure 2.1 A picture of a "precision mix batch mixer" that would be used to mix large quantities of foods such as peanut butter and other mixes of substances that have relatively high values of density ρ and viscosity μ (courtesy of H. C. Davis Sons Manufacturing Company, Inc.).

Dimensional analysis developed as an attempt to perform extended, costly experiments in a more organized, more efficient fashion. The underlying idea was to see whether the number of variables could be grouped together so that fewer trial runs or fewer measurements would be needed. Dimensional analysis produces a more compact set of outputs or data, with perhaps fewer charts and graphs, which in turn might better clarify what is being observed.

Imagine for a moment that we want to design a machine to make large quantities of peanut butter (and this author prefers creamy to crunchy!). We can imagine a mixer that takes all of the ingredients (i.e., roasted peanuts, sugar, and "less than 2%" of molasses and partially hydrogenated vegetable oil) and mixes them into a smooth, creamy spread. Moving a knife through a jar of peanut butter requires a noticeably larger force than stirring a glass of water. Similarly, the forces in a vat-like mixer would be considerable, as would the power needed to run that mixer in an automated food assembly line, as illustrated in Figure 2.1. Thus, the electro-mechanical design of an industrial-strength peanut butter mixer depends on estimates of the forces required to mix the peanut butter. How can we get some idea of what those forces are?

It turns out, as you might expect, that the forces depend in large part on properties of the peanut butter, but on which properties, and how? We can

answer those questions by performing a series of experiments in which we push a blade through a tub of peanut butter and measure the amount of force required to move the blade at different speeds. We will call the force needed to move the blade through the peanut butter the drag force, F_D, because it is equal to the force exerted by the moving (relatively speaking) peanut butter to retard the movement of the knife. We postulate that the force depends on the speed V with which the blade moves, on a characteristic dimension of the blade, say the width d, and on two characteristics of the peanut butter. One of these characteristics is the mass density, ρ, and the second is a parameter called the *viscosity*, μ, which is a measure of its "stickiness." If we think about our experiences with various fluids, including water, honey, motor oil, and peanut butter, these two characteristics seem intuitively reasonable because we do associate a difficulty in stirring (and cleaning up) with fluids that feel heavier and stickier.

Thus, the five quantities that we will take as derived quantities for this initial investigation into the mixing properties of peanut butter are the drag force, F_D, the speed with which the blade moves, V, the knife blade width, d, the peanut butter's mass density, ρ, and its viscosity, μ. The fundamental physical quantities we would apply here are mass, length, and time, which we denote as M, L, and T, respectively. The derived variables are expressed in terms of the fundamental quantities in Table 2.1.

Table 2.1 The five derived quantities chosen to model the peanut butter stirring experiments.

Derived quantities	Dimensions
Speed (V)	$\mathsf{L/T}$
Blade width (d)	L
Density (ρ)	$\mathsf{M}/(\mathsf{L})^3$
Viscosity (μ)	$\mathsf{M}/(\mathsf{L}\times\mathsf{T})$
Drag force (F_D)	$(\mathsf{M}\times\mathsf{L})/(\mathsf{T})^2$

How did we get the fundamental dimensions of the viscosity? By a straightforward application of the principle of dimensional homogeneity to the assumptions used in modeling the mechanics of fluids: The drag force (or force required to pull the blade through the butter) is directly proportional both to the speed with which it moves and the area of the blade, and inversely proportional to a length that characterizes the spatial rate of change of the speed. Thus,

$$F_D \propto \frac{VA}{L}, \tag{2.5a}$$

or

$$F_D = \mu \frac{VA}{L}. \tag{2.5b}$$

If we apply the principle of dimensional homogeneity to eq. (2.5b), it follows that

$$[\mu] = \left[\frac{F_D}{A} \times \frac{L}{V} \right]. \tag{2.6}$$

It is easy to show that eq. (2.6) leads to the corresponding entry in Table 2.1.

Now we consider the fact that we want to know how F_D and V are related, and yet they are also functions of the other variables, d, ρ, and μ, that is,

$$F_D = F_D(V; d, \rho, \mu). \tag{2.7}$$

Equation (2.7) suggests that we would have to do a lot of experiments and plot a lot of curves to find out how drag force and speed relate to each other while we are also varying the blade width and the butter density and viscosity. If we wanted to look at only three different values of each of d, ρ, and μ, we would have nine (9) different graphs, each containing three (3) curves. This is a significant accumulation of data (and work!) for a relatively simple problem, and it provides a very graphic illustration of the need for dimensional analysis. We will soon show that this problem can be "reduced" to considering two dimensionless groups that are related by a single curve! Dimensional analysis is thus very useful for both designing and conducting experiments.

Problem 2.1. Justify the assertion made just above that "nine (9) different graphs, each containing three (3) curves" are needed to relate force and speed.

Problem 2.2. Find and compare the mass density and viscosity of peanut butter, honey, and water.

2.4 How Do We Do Dimensional Analysis?

Dimensional analysis is the process by which we ensure dimensional consistency. It ensures that we are using the proper dimensions to describe the problem being modeled, whether expressed in terms of the correct number of properly dimensioned variables and parameters or whether written in terms of appropriate dimensional groups. Remember, too, that we need consistent dimensions for logical consistency, and we need consistent units for arithmetic consistency.

How do we ensure dimensional consistency? First, we check the dimensions of all derived quantities to see that they are properly represented in terms of the chosen primary quantities and their dimensions. Then we identify the proper *dimensionless groups* of variables, that is, ratios and products of problem variables and parameters that are themselves dimensionless. We will explain two different techniques for identifying such dimensionless groups, the *basic method* and the *Buckingham Pi theorem.*

2.4.1 The Basic Method of Dimensional Analysis

The basic method of dimensional analysis is a rather informal, unstructured approach for determining dimensional groups. It depends on being able to construct a functional equation that contains all of the relevant variables, for which we know the dimensions. The proper dimensionless groups are then identified by the thoughtful elimination of dimensions.

For example, consider one of the classic problems of elementary mechanics, the free fall of a body in a vacuum. We recall that the speed, V, of such a falling body is related to the gravitational acceleration, g, and the height, h, from which the body was released. Thus, the functional expression of this knowledge is:

$$V = V(g, h). \tag{2.8}$$

Note that the precise form of this functional equation is, at this point, entirely unknown—and we don't need to know that form for what we're doing now. The physical dimensions of the three variables are:

$$\begin{aligned} [V] &= \mathsf{L}/\mathsf{T}, \\ [g] &= \mathsf{L}/\mathsf{T}^2, \\ [h] &= \mathsf{L}. \end{aligned} \tag{2.9}$$

The time dimension, T, appears only in the speed and gravitational acceleration, so that dividing the speed by the square root of g eliminates time and yields a quantity whose remaining dimension can be expressed entirely in terms of length, that is:

$$\left[\frac{V}{\sqrt{g}}\right] = \sqrt{\mathsf{L}}. \tag{2.10}$$

If we repeat this thought process with regard to eliminating the length dimension, we would divide eq. (2.10) by $\sqrt{h}$, which means that

$$\left[\frac{V}{\sqrt{gh}}\right] = 1. \tag{2.11}$$

Since we have but a single dimensionless group here, it follows that:

$$V = \text{constant} \times \left(\sqrt{gh}\right) \tag{2.12}$$

Thus, the speed of a falling body is proportional to $\sqrt{gh}$, a result we should recall from physics—yet we have found it with dimensional analysis alone, without invoking Newton's law or any other principle of mechanics. This elementary application of dimensional consistency tells us something about the power of dimensional analysis. On the other hand, we do need some physics, either theory or experiment, to define the constant in eq. (2.12).

Someone seeing the result (2.12) might well wonder why the speed of a falling object is independent of mass (unless that person knew of Galileo Galilei's famous experiment). In fact, we can use the basic method to build on eq. (2.12) and show why this is so. Simply put, we start with a functional equation that included mass, that is,

$$V = V(g, h, m). \tag{2.13}$$

A straightforward inspection of the dimensions of the four variables in eq. (2.13), such as the list in eq. (2.9), would suggest that mass is not a variable in this problem because it only occurs once as a dimension, so it cannot be used to make eq. (2.13) dimensionless.

As a further illustration of the basic method, consider the mutual revolution of two bodies in space that is caused by their mutual gravitational attraction. We would like to find a dimensionless function that relates the period of oscillation, T_R, to the two masses and the distance r between them:

$$T_R = T_R(m_1, m_2, r). \tag{2.14}$$

If we list the dimensions for the four variables in eq. (2.14) we find:

$$\begin{aligned} [m_1], [m_2] &= \mathsf{M}, \\ [T_R] &= \mathsf{T}, \\ [r] &= \mathsf{L}. \end{aligned} \tag{2.15}$$

We now have the converse of the problem we had with the falling body. Here none of the dimensions are repeated, save for the two masses. So, while we can expect that the masses will appear in a dimensionless ratio, how do we keep the period and distance in the problem? The answer is that we need to add a variable containing the dimensions heretofore missing to

the functional equation (2.14). Newton's gravitational constant, G, is such a variable, so we restate our functional equation (2.14) as

$$T_R = T_R(m_1, m_2, r, G), \tag{2.16}$$

where the dimensions of G are

$$[G] = \mathsf{L}^3/\mathsf{MT}^2. \tag{2.17}$$

The complete list of variables for this problem, consisting of eqs. (2.15) and (2.17), includes enough variables to account for all of the dimensions.

Regarding eq. (2.16) as the correct functional equation for the two revolving bodies, we apply the basic method first to eliminate the dimension of time, which appears directly in the period T_R and as a reciprocal squared in the gravitational constant G. It follows dimensionally that

$$\left[T_R\sqrt{G}\right] = \sqrt{\frac{\mathsf{L}^3}{\mathsf{M}}}, \tag{2.18a}$$

where the right-hand side of eq. (2.18a) is independent of time. Thus, the corresponding revised functional equation for the period would be:

$$T_R\sqrt{G} = T_{R1}(m_1, m_2, r). \tag{2.18b}$$

We can eliminate the length dimension simply by noting that

$$\left[\frac{T_R\sqrt{G}}{\sqrt{r^3}}\right] = \sqrt{\frac{1}{\mathsf{M}}}, \tag{2.19a}$$

which leads to a further revised functional equation,

$$\frac{T_R\sqrt{G}}{\sqrt{r^3}} = T_{R2}(m_1, m_2). \tag{2.19b}$$

We see from eq. (2.19a) that we can eliminate the mass dimension from eq. (2.19b) by multiplying eq. (2.19b) by the square root of one of the two masses. We choose the square root of the second mass (do Problem 2.6 to find out what happens if the first mass is chosen), $\sqrt{m_2}$, and we find from eq. (2.19a) that

$$\left[\frac{T\sqrt{Gm_2}}{\sqrt{r^3}}\right] = 1. \tag{2.20a}$$

This means that eq. (2.19b) becomes

$$\frac{T_R\sqrt{Gm_2}}{\sqrt{r^3}} = \sqrt{m_2}\,T_{R2}(m_1, m_2) \equiv T_{R3}\left(\frac{m_1}{m_2}\right), \tag{2.20b}$$

where a dimensionless mass ratio has been introduced in eq. (2.20b) to recognize that this is the only way that the function T_{R3} can be both dimensionless *and* a function of the two masses. Thus, we can conclude from eq. (2.20b) that

$$T_R = \sqrt{\frac{r^3}{Gm_2}}\, T_{R3}\left(\frac{m_1}{m_2}\right). \tag{2.21}$$

This example shows that difficulties arise if we start a problem with an incomplete set of variables. Recall that we did not include the gravitational constant G until it became clear that we were headed down a wrong path. We then included G to rectify an incomplete analysis. With the benefit of hindsight, we might have argued that the attractive gravitational force must somehow be accounted for, and including G could have been a way to do that. This argument, however, demands insight and judgment whose origins may have little to do with the particular problem at hand.

While our applications of the basic method of dimensional analysis show that it does not have a formal algorithmic structure, it can be described as a series of steps to take:

a. List all of the variables and parameters of the problem and their dimensions.
b. Anticipate how each variable qualitatively affects quantities of interest, that is, does an increase in a variable cause an increase or a decrease?
c. Identify one variable as depending on the remaining variables and parameters.
d. Express that dependence in a functional equation (i.e., analogs of eqs. (2.8) and (2.14)).
e. Choose and then eliminate one of the primary dimensions to obtain a revised functional equation.
f. Repeat steps (e) until a revised, *dimensionless* functional equation is found.
g. Review the final *dimensionless* functional equation to see whether the apparent behavior accords with the behavior anticipated in step "b".

Problem 2.3. What is the constant in eq. (2.12)? How do you know that?

Problem 2.4. Apply the basic method to eq. (2.2) for the period of the pendulum.

Problem 2.5. Carry out the basic method for eq. (2.13) and show that the mass of a falling body does not affect its speed of descent.

Problem 2.6. Carry out the last step of the basic method for eqs. (2.20) using the first mass and show it produces a form that is equivalent to eq. (2.21).

2.4.2 The Buckingham Pi Theorem for Dimensional Analysis

Buckingham's Pi theorem, fundamental to dimensional analysis, can be stated as follows:

A dimensionally homogeneous equation involving n *variables in* m *primary or fundamental dimensions can be reduced to a single relationship among* n − m *independent dimensionless products.*

A dimensionally homogeneous (or rational) equation is one in which every independent, additive term in the equation has the same dimensions. This means that we can solve for any one term as a function of all the others. If we introduce Buckingham's Π notation to represent a dimensionless term, his famous Pi theorem can be written as:

$$\Pi_1 = \Phi(\Pi_2, \Pi_3 \ldots \Pi_{n-m}). \tag{2.22a}$$

or, equivalently,

$$\Phi(\Pi_1, \Pi_2, \Pi_3 \ldots \Pi_{n-m}) = 0. \tag{2.22b}$$

Equations (2.22) state that a problem with n derived variables and m primary dimensions or variables requires $n - m$ dimensionless groups to correlate all of its variables.

We apply the Pi theorem by first identifying the n derived variables in a problem: $A_1, A_2, \ldots A_n$. We choose m of these derived variables such that they contain all of the m primary dimensions, say, A_1, A_2, A_3 for $m = 3$. Dimensionless groups are then formed by permuting each of the remaining $n - m$ variables ($A_4, A_5, \ldots A_n$ for $m = 3$) in turn with those m's already chosen:

$$\begin{aligned} \Pi_1 &= A_1^{a_1} A_2^{b_1} A_3^{c_1} A_4, \\ \Pi_2 &= A_1^{a_2} A_2^{b_2} A_3^{c_2} A_5, \\ &\vdots \\ \Pi_{n-m} &= A_1^{a_{n-m}} A_2^{b_{n-m}} A_3^{c_{n-m}} A_n. \end{aligned} \tag{2.23}$$

The a_i, b_i, and c_i are chosen to make each of the permuted groups Π_i dimensionless.

For example, for the peanut butter mixer, there should be two dimensionless groups correlating the five variables of the problem (listed in Table 2.1). To apply the Pi theorem to this mixer we choose the blade speed V, its width d, and the butter density ρ as the fundamental variables ($m = 3$), which we then permute with the two remaining variables—the viscosity μ and the drag force F_D—to get two dimensionless groups:

$$\begin{aligned} \Pi_1 &= V^{a_1} d^{b_1} \rho^{c_1} \mu, \\ \Pi_2 &= V^{a_2} d^{b_2} \rho^{c_2} F_D. \end{aligned} \tag{2.24}$$

Expressed in terms of primary dimensions, these groups are:

$$\begin{aligned} \Pi_1 &= \left(\frac{\mathsf{L}}{\mathsf{T}}\right)^{a_1} \mathsf{L}^{b_1} \left(\frac{\mathsf{M}}{\mathsf{L}^3}\right)^{c_1} \left(\frac{\mathsf{M}}{\mathsf{LT}}\right), \\ \Pi_2 &= \left(\frac{\mathsf{L}}{\mathsf{T}}\right)^{a_2} \mathsf{L}^{b_2} \left(\frac{\mathsf{M}}{\mathsf{L}^3}\right)^{c_2} \left(\frac{\mathsf{ML}}{\mathsf{T}^2}\right). \end{aligned} \tag{2.25}$$

Now, in order for Π_1 and Π_2 to be dimensionless, the net exponents for each of the three primary dimensions must vanish. Thus, for Π_1,

$$\begin{aligned} \mathsf{L}: &\quad a_1 + b_1 - 3c_1 - 1 = 0, \\ \mathsf{T}: &\quad -a_1 - 1 = 0, \\ \mathsf{M}: &\quad c_1 + 1 = 0, \end{aligned} \tag{2.26a}$$

and for Π_2,

$$\begin{aligned} \mathsf{L}: &\quad a_2 + b_2 - 3c_2 + 1 = 0, \\ \mathsf{T}: &\quad -a_2 - 2 = 0, \\ \mathsf{M}: &\quad c_2 + 1 = 0. \end{aligned} \tag{2.26b}$$

Solving eqs. (2.26) for the two pairs of subscripts yields:

$$\begin{aligned} a_1 &= b_1 = c_1 = -1, \\ a_2 &= b_2 = -2, \quad c_2 = -1. \end{aligned} \tag{2.27}$$

Then the two dimensionless groups for the peanut butter mixer are:

$$\begin{aligned} \Pi_1 &= \left(\frac{\mu}{\rho V d}\right), \\ \Pi_2 &= \left(\frac{F_D}{\rho V^2 d^2}\right). \end{aligned} \tag{2.28}$$

Thus, there are two dimensionless groups that should guide experiments with prototype peanut butter mixers. One clearly involves the viscosity of the peanut butter, while the other relates the drag force on the blade to

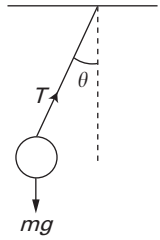


Figure 2.2 The classical pendulum oscillating through angles θ due to gravitational acceleration g.

the blade's dimensions and speed, as well as to the density of the peanut butter.

In Chapter 7 we will explore one of the "golden oldies" of physics, modeling the small angle, free vibration of an ideal pendulum (viz. Figure 2.2). There are six variables to consider in this problem, and they are listed along with their fundamental dimensions in Table 2.2. In this case we have $m = 6$ and $n = 3$, so that we can expect three dimensionless groups. We will choose l, g, and m as the variables around which we will permute the remaining three variables (T_0, θ, T) to obtain the three groups. Thus,

$$\begin{aligned} \Pi_1 &= l^{a_1} g^{b_1} m^{c_1} T_0, \\ \Pi_2 &= l^{a_2} g^{b_2} m^{c_2} \theta, \\ \Pi_3 &= l^{a_3} g^{b_3} m^{c_3} T. \end{aligned} \tag{2.29}$$

Table 2.2 The six derived quantities chosen to model the oscillating pendulum.

Derived quantities	Dimensions
Length (l)	L
Gravitational acceleration (g)	L/T^2
Mass (m)	M
Period (T_0)	T
Angle (θ)	1
String tension (T)	$(\mathsf{M} \times \mathsf{L})/\mathsf{T}^2$

The Pi theorem applied here then yields three dimensionless groups (see Problem 2.9):

$$\begin{aligned}\Pi_1 &= \frac{T_0}{\sqrt{l/g}},\\ \Pi_2 &= \theta,\\ \Pi_3 &= \frac{T}{mg}.\end{aligned} \tag{2.30}$$

These groups show how the period depends on the pendulum length l and the gravitational constant g (recall eq. (2.2)), and the string tension T on the mass m and g. The second group also shows that the (dimensionless) angle of rotation stands alone, that is, it is apparently not related to any of the other variables. This follows from the assumption of small angles, which makes the problem linear, and makes the magnitude of the angle of free vibration a quantity that cannot be determined.

One of the "rules" of applying the Pi theorem is that the m chosen variables include all n of the fundamental dimensions, but no other restrictions are given. So, it is natural to ask how this analysis would change if we start with three different variables. For example, suppose we choose T_0, g, and m as the variables around which to permute the remaining three variables (l, θ, T) to obtain the three groups. In this case we would write:

$$\begin{aligned}\Pi_1' &= T_0^{a_1} g^{b_1} m^{c_1} l,\\ \Pi_2' &= T_0^{a_2} g^{b_2} m^{c_2} \theta,\\ \Pi_3' &= T_0^{a_3} g^{b_3} m^{c_3} T.\end{aligned} \tag{2.31}$$

Applying the Pi theorem to eq. (2.31) yields the following three "new" dimensionless groups (see Problem 2.10):

$$\begin{aligned}\Pi_1' &= \frac{l/g}{T_0^2} = \frac{1}{\Pi_1^2},\\ \Pi_2' &= \theta = \Pi_2,\\ \Pi_3' &= \frac{T}{mg} = \Pi_3.\end{aligned} \tag{2.32}$$

We see that eq. (2.32) produce the same information as eq. (2.30), albeit in a slightly different form. In particular, it is clear that Π_1 and Π_1' contain the same dimensionless group, which suggests that the number of dimensionless groups is unique, but that the precise forms that these groups

may take are not. This last calculation demonstrates that the dimensionless groups determined in any one calculation are unique in one sense, but they may take on different, yet related forms when done in a slightly different calculation.

Note that our applications of the basic method and the Buckingham Pi theorem of dimensional analysis can be cast in similar, step-like structures. However, experience and insight are key to applying both methods, even for elementary problems. Perhaps this is a context where, as was said by noted British economist John Maynard Keynes in his famous book, *The General Theory of Employment, Interest, and Money*, "Nothing is required and nothing will avail, except a little, a very little, clear thinking."

Problem 2.7. Write out the Buckingham Pi theorem as a series of steps, analogous to the steps described in the basic method on p. 23.

Problem 2.8. Confirm eq. (2.28) by applying the basic method to the mixer problem.

Problem 2.9. Confirm eq. (2.30) by applying the rest of the Buckingham Pi theorem to the pendulum problem.

Problem 2.10. Confirm eq. (2.32) by applying the rest of the Buckingham Pi theorem to the revised formulation of the pendulum problem.

Problem 2.11. Apply the Buckingham Pi theorem to the revolution of two bodies in space, beginning with the functional equation (2.16).

Problem 2.12. What happens when the Pi theorem is applied to the two-body problem, but beginning now with the functional equation (2.14)?

2.5 Systems of Units

We have already noted that units are numerical measures derived from standards. Thus, units are fractions or multiples of those standards. The *British* system has long been the most commonly used system of units in the United States. In the British system, length is typically referenced in *feet*, force in *pounds*, time in *seconds*, and mass in *slugs*. The unit of *pound force* is defined as that force that imparts an acceleration of 32.1740 ft/sec^2 to a mass of 1/2.2046 of that piece of platinum known as the standard kilogram. While keeping in mind the distinction between dimensions and

Table 2.3 The British and SI systems of units, including abbreviations and (approximate) conversion factors.

Reference Units	British System	SI System
length	foot (ft)	meter (m)
time	second (sec)	second (s)
mass	slug (slug), pound mass (lbm)	kilogram (kg)
force	pound force (lb)	newton (N)
Multiply number of	**by**	**to get**
feet (ft)	0.3048	meters (m)
inches (in)	2.540	centimeters (cm)
miles (mi)	1.609	kilometers (km)
miles per hour (mph)	0.447	meters per second (m/s)
pounds force (lb)	4.448	newtons (N)
slugs (slug)	14.59	kilograms (kg)
pounds mass (lbm)	0.454	kilograms (kg)

units, it is worth noting that the fundamental reference quantities in the British system of units (foot, pound, second) are based on the primary dimensions of length (L), force (F), and time (T).

In a belated acknowledgment that the rest of the world (including Britain!) uses *metric* units, a newer system of units is increasingly being adopted in the United States. The *Système International*, commonly identified by its initials, SI, is based on the *mks* system used in physics and it references length in *meters*, mass in *kilograms*, and time in *seconds*. The primary dimensions of the SI system are length (L), mass (M), and time (T). Force is a derived variable in the SI system and it is measured in *newtons*. In Table 2.3 we summarize some of the salient features of the SI and British systems, including the abbreviations used for each unit and the (approximate) conversion factors needed to navigate between the two systems.

Finally, we show in Table 2.4 some of the standard prefixes used to denote the various multiplying factors that are commonly used to denote fractions or multiples of ten (10). To cite a familiar example, in the metric system distances are measured in millimeters (mm), centimeters (cm), meters (m), and kilometers (km).

It is worth noting that some caution is necessary in using the prefixes listed in Table 2.4 because this usage is not universally uniform. For example, the measures used for computer memory are kilobytes (KB), megabytes (MB) and increasingly often these days, gigabytes (GB) and terabytes (TB). The B's stand for bytes. However, the prefixes kilo, mega,

Table 2.4 Some standard numerical factors that are commonly used in the SI system.

Numerical factors (SI)	Prefix (symbol)
10^{-9}	nano (n)
10^{-6}	micro (μ)
10^{-3}	milli (m)
10^{-2}	centi (c)
10^{3}	kilo (k)
10^{6}	mega (M)
10^{9}	giga (G)
10^{12}	tera (T)

giga, and tera, respectively stand for 2^{10}, 2^{20}, 2^{30}, and 2^{40}, which is rather different than what we are using!

To finish off our discussion of numbers we add one final set of approximate equalities, for their interest and possible usefulness:

$$2^{10} \cong e^{7} \cong 10^{3}. \tag{2.33}$$

2.6 Summary

In this chapter we have described an important aspect of problem formulation and modeling, namely, dimensional analysis. Reasoning about the dimensions of a problem requires that we (1) identify all of the physical variables and parameters needed to fully describe a problem, (2) select a set of primary dimensions and variables, and (3) develop the appropriate set of dimensionless groups for that problem. The last step is achieved by applying either the basic method or the more structured Buckingham Pi theorem. The dimensionless groups thus found, along with their numerical values as determined from experiments or further analysis, help us to assess the importance of various effects, to buttress our physical insight and understanding, and to organize our numerical computation, our data gathering, and our design of experiments.

We close by noting that while our use of the formal methods of dimensional analysis will be limited, we will use the concepts of dimensional consistency and dimensional groups extensively. We will see these concepts when we discuss scaling, when we formulate models, and when we solve particular problems. In so doing, we will also keep in mind the distinction

between dimensions and units, and we will also ensure the consistency of units.

2.7 References

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

J. M. Keynes, *The General Theory of Employment, Interest, and Money*, Harcourt, Brace & World, New York, 1964.

H. L. Langhaar, *Dimensional Analysis and Theory of Models*, John Wiley, New York, 1951.

H. Schenck, Jr., *Theories of Engineering Experimentation*, McGraw-Hill, New York, 1968.

R. P. Singh and D. R. Heldman, *Introduction to Food Engineering*, 3rd Edition, Academic Press, San Diego, CA, 2001.

E. S. Taylor, *Dimensional Analysis for Engineers*, Oxford (Clarendon Press), London and New York, 1974.

D. F. Young, B. R. Munson, and T. H. Okiishi, *A Brief Introduction to Fluid Mechanics*, 2nd Edition, John Wiley, New York, 2001.

2.8 Problems

2.13. Consider a string of length l that connects a rock of mass m to a fixed point while the rock whirls in a circle at speed v. Use the basic method of dimensional analysis to show that the tension T in the string is determined by the dimensionless group

$$\frac{TL}{mv^2} = \text{constant}.$$

2.14. Apply the Buckingham Pi theorem to confirm the analysis of Problem 2.13.

2.15. The speed of sound in a gas, c, depends on the gas pressure p and on the gas mass density ρ. Use dimensional analysis to determine how c, p, and ρ are related.

2.16. A dimensionless grouping called the Weber number, W_e, is used in fluid mechanics to relate a flowing fluid's surface tension, σ, which has dimensions of force/length, to the fluid's speed, v, density, ρ, and a characteristic length, l. Use dimensional analysis to find that number.

2.17. A pendulum swings in a viscous fluid. How many groups are needed to relate the usual pendulum variables to the fluid viscosity, μ, the fluid mass density, ρ, and the diameter d of the pendulum? Find those groups.

2.18. The volume flow rate Q of fluid through a pipe is thought to depend on the pressure drop per unit length, $\Delta p/l$, the pipe diameter, d, and the fluid viscosity, μ. Use the basic method of dimensional analysis to determine the relation:

$$Q = (\text{constant}) \left(\frac{d^4}{\mu} \right) \left(\frac{\Delta p}{l} \right).$$

2.19. Apply the Buckingham Pi theorem to confirm the analysis of Problem 2.18.

2.20. When flow in a pipe with a rough inner wall (perhaps due to a build-up of mineral deposits) is considered, several variables must be considered, including the fluid speed v, its density ρ and viscosity μ, and the pipe length l and diameter d. The average variation e of the pipe radius can be taken as a measure of the roughness of the pipe's inner surface. Determine the dimensionless groups needed to determine how the pressure drop Δp depends on these variables.

2.21. Use dimensional analysis to determine how the speed of sound in steel depends on the modulus of elasticity, E, and the mass density, ρ. (The modulus of elasticity of steel is, approximately and in British units, 30×10^6 psi.)

2.22. The flexibility (the deflection per unit load) or compliance C of a beam having a square cross-section $d \times d$ depends on the beam's length l, its height and width, and its material's modulus of elasticity E. Use the basic method of dimensional analysis to show that:

$$CEd = F_{CEd} \left(\frac{l}{d} \right).$$

2.23. Experiments were conducted to determine the specific form of the function $F_{CEd}(l/d)$ found in Problem 2.22. In these experiments it was found that a plot of $\log_{10}(CEd)$ against $\log_{10}(l/d)$ has a slope of 3 and an intercept on the $\log_{10}(CEd)$ scale of -0.60. Show that the deflection under a load P can be given in terms of the second moment of area I of the cross section ($I = d^4/12$) as:

$$\text{deflection} = \text{load} \times \text{compliance} = P \times C = \frac{Pl^3}{48EI}.$$

3
Scale

In this chapter we will continue dealing with dimensions, but focusing now on issues of *scale*, that is, issues of *relative size*. Size, whether absolute or relative, is very important because it affects both the form and the function of those objects or systems being modeled. Scaling inßuencesÑ indeed, often controlsÑ the way objects interact with their environments, whether we are talking about objects in nature, the design of experiments, or the representation of data by smooth, nice-looking curves. We even Þnd references to scaling in literature, such as in the depiction by satirist Jonathan Swift of the treatment accorded the traveler Gulliver when he arrived in the land of Lilliput:

His Majesty's Ministers, finding that Gulliver's stature exceeded theirs in the proportion of twelve to one, concluded from the similarity of their bodies that his must contain at least 1728 of theirs, and must needs be rationed accordingly.

This chapter is devoted to explaining where the factor of 1728 came from, as well as discussing abstraction and scale, size and shape, size and function, scaling and conditions that are imposed at an objectÕs boundaries, and some of the consequences of choosing scales in both theory and experimental measurements.

3.1 Abstraction and Scale

We start with some thoughts about the process of deciding on the appropriate level of detail for whatever problem is at hand, which also means

deciding on the appropriate level of detail for the corresponding model. We call this process *abstraction*. It typically requires an organized, thoughtful approach to identifying those phenomena to which we really want to pay attention. In addition, thinking about *scaling* often requires that we think in terms of the magnitude or size of quantities measured with respect to a standard that has the same physical dimensions.

For example, a linear elastic spring can be used to model more than just the relation between force and relative extension of a simple coiled spring, as in an old-fashioned butcherÕs scale or an automobile spring. We could, for example, use $F = kx$ to describe the static load-deßection behavior of a diving board, but the spring constant k should reßect the stiffness of the diving board taken as a whole, which in turn reßects more detailed properties of the board, including the material of which it is made and its own dimensions. The validity of using a linear spring to model the board can be ascertained by measuring and plotting the deßection of the boardÕs tip as it changes with standing divers of different weight.

We had noted in Section 1.3.1 that the classic spring equation is also used to model the static and dynamic behavior of tall buildings as they respond to wind loading and to earthquakes. These examples suggest that we can use a simple, highly abstracted model of a building by aggregating various details within the parameters of that model. That is, the stiffness k for a building would incorporate or lump together a great deal of information about how the building is framed, its geometry, its materials, and so on. For both a diving board and a tall building, we would need detailed expressions of how their respective stiffnesses depended on their respective properties. We could not do a detailed design of either the board or of the building without such expressions. Similarly, using springs to model atomic bonds means that their spring constants must be related to atomic interaction forces, atomic distances, sub-atomic particle dimensions, and so on.

Another facet of the abstraction process is that in each case we are saying that, for some well-deÞned purposes, a Òreal,Ó three-dimensional object behaves like a simple spring. We are thus introducing the concept of a *lumped element* model wherein the actual physical properties of some real object or device are aggregated or *lumped* into a less detailed, more abstract expression. An airplane, for example, can be modeled in very different ways, depending on our modeling goals. To lay out a ßight plan or trajectory, the airplane can simply be considered as a point mass moving with respect to a spherical coordinate system. The mass of the point can simply be taken as the total mass of the plane, and the effect of the surrounding atmosphere can also be modeled by expressing the retarding drag force as acting on the mass point itself with a magnitude related to the relative speed at which the mass is moving. If we want to model and analyze the more immediate, more local effects of the movement of air over the planeÕs wings, we would build

a model that accounts for the wing's surface area and is complex enough to incorporate the aerodynamics that occur in different flight regimes. If we want to model and design the flaps used to control the plane's ascent and descent, we would develop a model that includes a system for controlling the flaps and also accounts for the dynamics of the wing's strength and vibration response.

Clearly, as we talk about finding the right level of abstraction or the right level of detail, we are simultaneously talking about finding the right *scale* for the model we are developing. *Scaling* or imposing a scale includes assessing the effects of geometry on scale, the relationship of function to scale, and the role of size in determining limits. We must think about all of these ideas when we are determining how to scale a model in relation to the "reality" we want to capture.

Lastly, we often look at the scale of things with respect to a magnitude set within a standard. Thus, when talking about freezing phenomena, we expect to reference temperatures near the freezing point of materials included in our model. Similarly, we know that the models of Newtonian mechanics work extraordinarily well for virtually all of our earth- and space-bound applications. Why is that so? Simply because the speeds involved in all of these calculations are far, far smaller than c, the speed of light in a vacuum. Thus, even a rocket fired at escape speeds of 45,000 km/hr seems to stand still when its speed is compared to $c \approx 300{,}000\,\text{km/s} = 1.080 \times 10^9\,\text{km/hr}$! These scaling ideas also represent something of an extension of the ideas behind dimensionless variables that we discussed in Chapter 2. For example, in Einstein's general theory of relativity, the mass of a particle moving at speed, v, is given as a (dimensionless) fraction of the rest mass, m_0, by

$$\frac{m}{m_0} = \frac{1}{\sqrt{1 - (v/c)^2}}. \tag{3.1}$$

The scaling issue involved here, as we will discuss in Section 3.4, is ensuring that the square of the dimensionless speed ratio is always much less than 1, so that $m \cong m_0$.

3.2 Size and Shape: Geometric Scaling

In Figure 3.1 we show two cubes, one of which has sides of unit length in any system of units we care to choose, that it, the cube's volume could be 1 in^3 or 1 m^3 or 1 km^3. The other cube has sides of length L in the same system of units, so its volume is either L^3 in^3 or L^3 m^3 or L^3 km^3. Thus, for comparison's sake, we can ignore the units in which the two cubes' sides

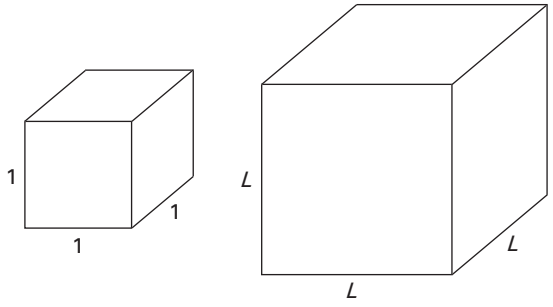


Figure 3.1 Two geometrically similar cubes, one with sides of *unit length* (that is, having lengths equal to 1 measured in any system of units), and the second with sides of length L as measured in the same units as the "unit cube."

are actually measured. The total area and volume of the Þrst cube are, respectively, 6 and 1, while the corresponding values for the second cube are $6L^2$ and L^3. We see immediately an instance of *geometric scaling*, that is, the area of the second cube changes as does L^2 and its volume scales as L^3. Thus, doubling the side of a cube increases its surface area by a factor of four and its volume by a factor of eight.

3.2.1 Geometric Scaling and Flight Muscle Fractions in Birds

Geometric scaling has been used quite successfully in many spheres of biology, for example, for comparing the effects of size and age in animals of the same species, and for comparing qualities and attributes in different species of animals. As an instance of the latter, consider Figure 3.2 wherein are plotted the total weight of the ßight muscles, W_{fm}, of quite a few birds against their respective body weights, W_b. How many birds are Òquite a fewÓ? The Þgure caption states that the underlying study actually included 29 birds, but the Þgure shows data only within the range $10 \leq$ bird number ≤ 23. For the 14 birds shown in Figure 3.2 there seems to be a fairly nice straight line Þt for the data presented. While Þtted by eye, that straight line can be determined to be:

$$W_{fm} \cong 0.18\, W_b. \tag{3.2}$$

Equation (3.2) suggests that ßight muscle makes up about 18% of a birdÕs body weight and that ßight muscle weight *scales linearly* withÑ or is

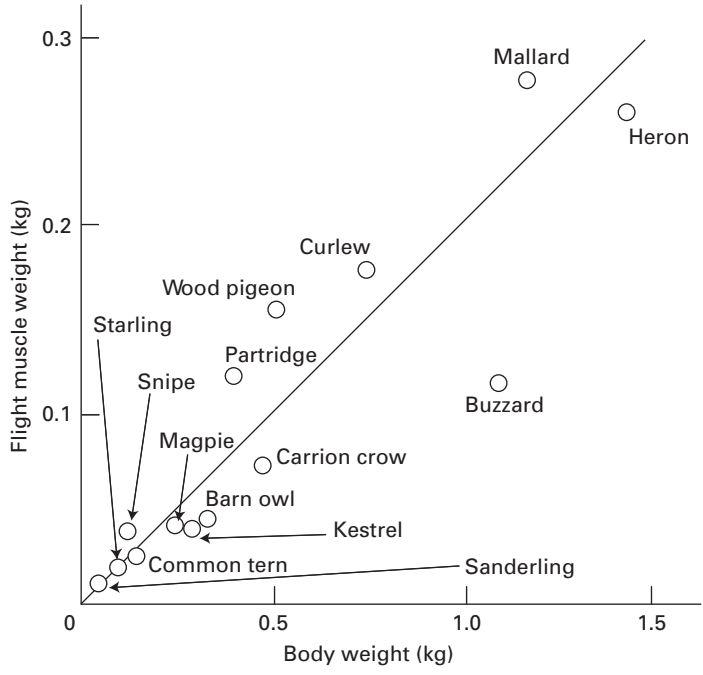


Figure 3.2 A simple linear fit on a plot of the total weight of the flight muscles against body weight for 14 of the 29 birds studied, including starlings, barn owls, kestrels, common terns, mallards, and herons (after Figure 1–2 of Alexander, 1971).

proportional toÑ body weight, a result that seems reasonable enough from our everyday observations of the birds around us.

3.2.2 Linearity and Geometric Scaling

These straightforward geometric scaling arguments can also be used to demonstrate some ideas about linearity in the context of *geometrically similar* objects, that is, objects whose basic geometry is essentially the same. In Figure 3.3 we show two pairs of drinking glasses: One pair are right circular cylinders of radius r, the second pair are right circular inverted cones having a common semi-vertex angle α. If the Þrst pair are Þlled to heights h_1 and h_2 respectively, the total ßuid volume in the two glasses is

$$V_{cy} = \pi r^2 h_1 + \pi r^2 h_2 = \pi r^2 (h_1 + h_2). \tag{3.3}$$

Equation (3.3) demonstrates that the volume is *linearly proportional* to the height of the ßuid in the two cylindrical glasses. Further, since the total

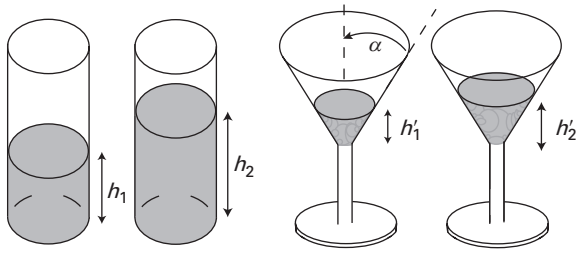


Figure 3.3 Two pairs of drinking glasses: One pair are cylinders of radius r, the second pair are inverted cones (sometimes referred to as martini glasses) having a common semi-vertex angle α.

volume can be obtained by adding or *superposing* the two heights, the volume V_{cy} is a *linear function* of the height h. (Recall the discussion in Section 1.3.4.) Note, however, that the volume is *not* a linear function of the radius, r.

For the two conical glasses, we see that their radii vary with height. In fact, the volume, V_{co}, of a cone with semi-vertex angle, α, Þlled to height, h, is

$$V_{co} = \frac{\pi}{3}\frac{h^3}{\tan^2\alpha}. \tag{3.4}$$

Hence, the total volume of ßuid in the two conical glasses of Figure 3.3 is

$$V_{co} = \frac{\pi}{3}\frac{h_1'^3}{\tan^2\alpha} + \frac{\pi}{3}\frac{h_2'^3}{\tan^2\alpha} \neq \frac{\pi}{3}\frac{(h_1' + h_2')^3}{\tan^2\alpha}. \tag{3.5}$$

The relationship between volume and height is nonlinear for the conical glasses, so we cannot calculate the total volume just by superposing the two ßuid heights, h_1' and h_2'.

3.2.3 "Log-log" Plots of Geometric Scaling Data

We now choose to ask a question: What happened to the other 15 birds in the small scaling study of Section 3.2.1? (Among those discriminated against in Figure 3.2 are hummingbirds, wrens, robins, skylarks, vultures, and albatrosses.) These birds were not included because the bird weights studied spanned a fairly large range, which made it hard to include the heavier birds (e.g., vultures and albatrosses) in the plot of Figure 3.2 without completely squashing the data for the very small birds (e.g., hummingbirds and goldcrests). This suggests a problem in organizing and presenting data, in itself an interesting aspect of scaling.

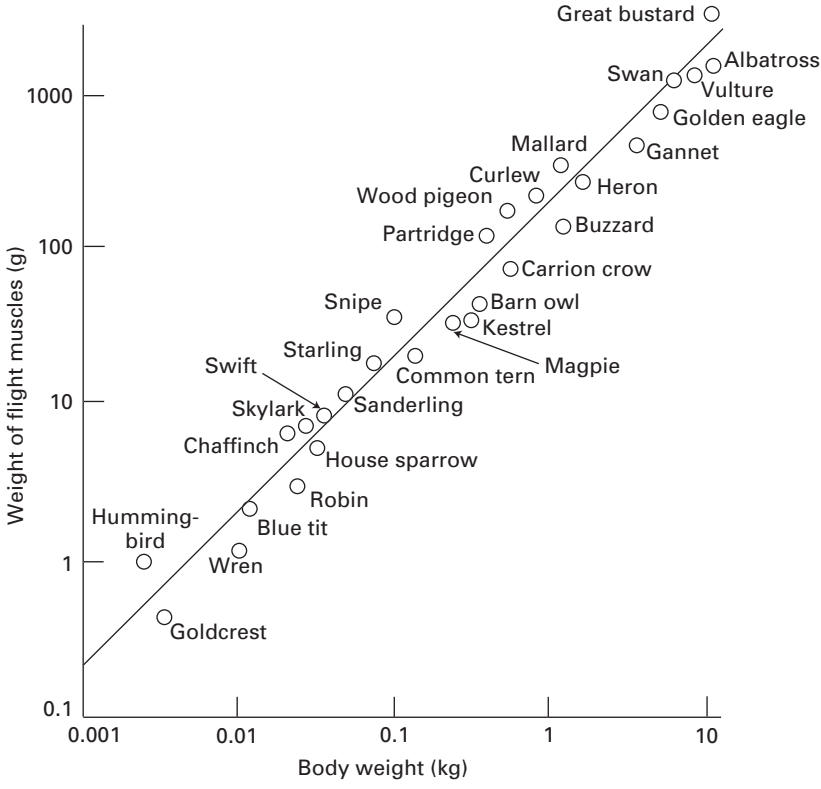


Figure 3.4 A "log-log" plot of the total weight of the flight muscles against body weight for 29 birds, including hummingbirds, wrens, terns, mallards, eagles, and albatrosses. Compare this with the linear plot of the data of Figure 3.2 (after Figure 1–4 of Alexander, 1971).

There is a straightforward way to include the heretofore left-out data: Construct *log-log* plots in which the *logarithms of the data* (normally to base 10) are graphed, as shown in Figure 3.4. In fact, the complete data set was plotted, essentially doubling the number of included data points, and a statistical regression analysis was applied to determine that the straight line shown in Figure 3.4 is given by:

$$W_{fm} = 0.18 W_b^{0.96}. \tag{3.6}$$

We could observe that eq. (3.6) is not exactly linear because, after all, $0.96 \neq 1$. However, it is clear that eqs. (3.2) and (3.6) are sufÞciently close that it is still quite reasonable to conclude that ßight muscle weight scales linearly with body weight.

However, this second look at the ßying muscle weight of birds raises two interesting scaling issues of scaling: First, how do we handle nonlinearities?

Second, how do we handle large ranges of data? In fact, we have just seen that these two questions are not unrelated because we found the *almost linear*, small nonlinearity in eq. (3.6), as a result of looking at an extended range of data.

We also have already provided an answer to the second question, namely, introducing log-log plots to extend our graphical range. Of course, with modern computational capabilities, we could skip the Òold fashionedÓ method of laboriously plotting data and simply enter tables of data points and let the computer spit out an equation or a curve. But something is gained by thinking through these issues without a computer.

Consider the data that emerged from a study of medieval churches and cathedrals in England. Large churches and cathedrals of that area (see Figures 3.5) were generally laid out in a cruciform pattern (viz., Figure 3.6) so that the *nave* was the principal longitudinal area, extending from a front entrance to a *chancel* or altar area at the back, and the *transept* was set out as a section perpendicular to the nave, quite close to the chancel. Was the cruciform shape ecclesiastically motivated, that is, was it inspired by religious feeling? In fact, research suggests that scaling dictated the

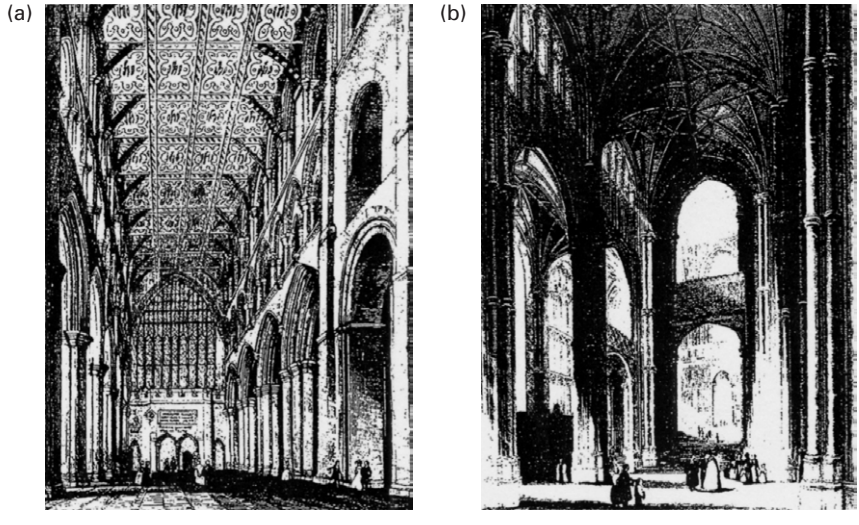


Figure 3.5 Interior views of two church *naves*: (a) The oldest Romanesque cathedral in England, St. Albans, has a nave with a relatively low height-to-length ratio; (b) The late Gothic style, also called the *perpendicular style*, is exemplified by the Canterbury Cathedral, whose nave has a relatively high height-to-length ratio (used by permission of the late Professor S. J. Gould of Harvard University).

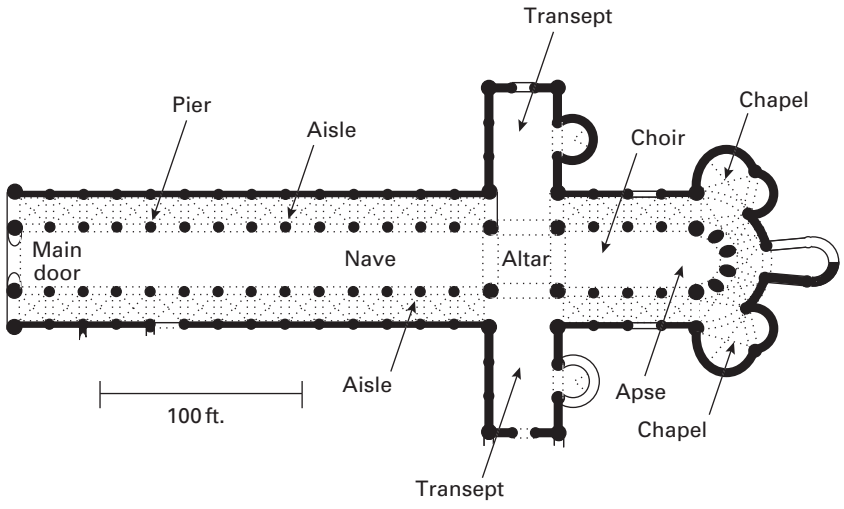


Figure 3.6 The plan of the Norwich Cathedral showing its cruciform shape, including the longitudinal nave, extending from the front door (left) to the rear apse (right), and the perpendicular transept (after Gould, 1975).

cruciform shape, and that the scaling was inspired by the need for both good lighting and sound structures.

We start by taking the length of a church as the Þrst-order indicator of its size. Thus, the longer its length, the larger the church. Then we examine the data displayed in Figure 3.7, which is a log-log plot of nave height against church length for a variety of medieval cathedrals and churches in England and on the European continent. We see from that data that as church length (and size) increase, the heights of their naves increases in absolute terms but falls off in *relative* terms. That is, as churches get longer (and larger), their naves get relatively smaller. Further, although we do not give the data to buttress this assertion, the bigger churches tend to have *narrower* naves. *Why don't the nave height and width increase with church size?* The answer has to do with the scaling of surface areas and enclosed volumes, that is, with geometric scaling.

The relevant scaling refers to the change in the area enclosed in a church as it is made longer (and larger). A longer church has a longer perimeter. In buildings of constant shape, the surface area of the enclosing wall increases linearly with the perimeter length, while the enclosed volume increases as (perimeter length)^2. Thus, problems emerge as it becomes more difÞcult for light and fresh air to penetrate into the churchÕs interior as its perimeter increases. (Remember that these marvelous structures were built long before the invention of the light bulb and air conditioning!) However,

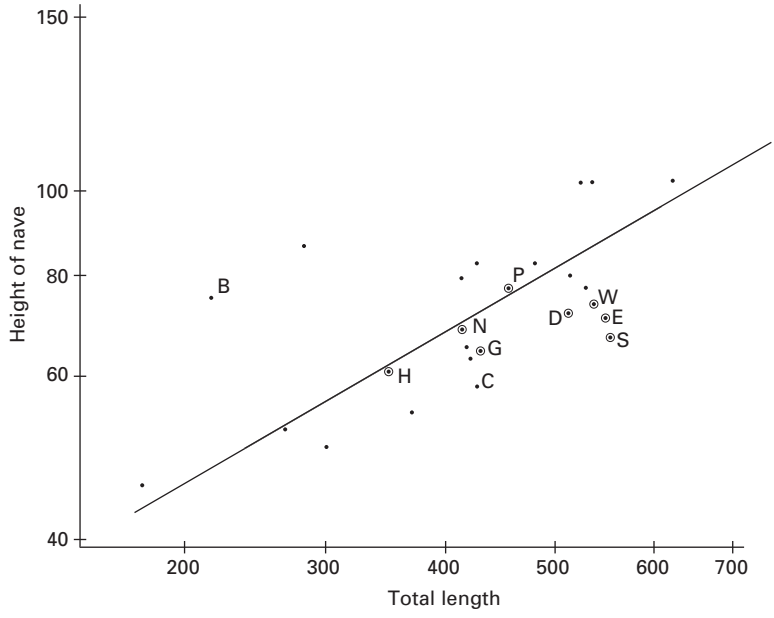


Figure 3.7 A plot of log (church nave height) against log (church length), with circled dots indicating Romanesque churches and letters standing for English churches as follows: B, Earls Barton; C, Chichester; D, Durham; E, Ely; G, Gloucester; H, Hereford; N, Norwich; P, Peterborough; S, St. Albans; and W, Winchester (used by permission of the late Professor S. J. Gould of Harvard University).

the severity of the lighting and ventilation problems can be reduced by introducing the transept because it enables a relatively constant nave width, thus taking away the Òconstant shapeÓconstraint. If the width is kept constant, then the enclosed area increases linearly with perimeter length, as does the churchÕslength (and size). And, of course, such a church will then appear to be relatively narrow!

Increasing a naveÕswidth along with its length is another way to increase the enclosed area, but this approach also exacerbates interior lighting and ventilation problems. And it creates still another problem, namely that of building a roof with a larger surface area to cover the enlarged, enclosed area. Since roofs of cathedrals and churches were built to sit atop stone *vaults* and arches, roof spans became critical because it was very hard to build wide stone arches and vaults. The difÞculty of building wide arches also interacts with the height of the nave for it is the nave walls that support the outward thrust developed in the roof vaults, even when the nave walls are supported by ßying buttresses (see Figure 3.8). Thus, higher nave walls

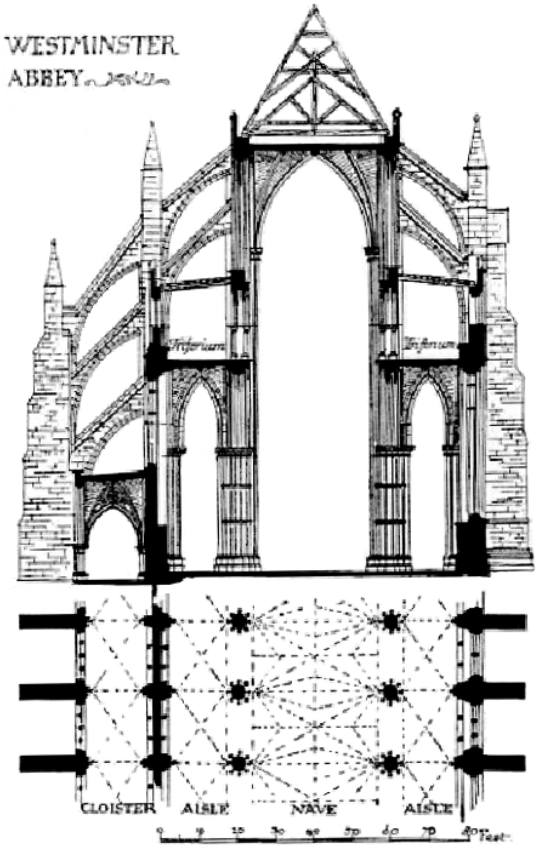


Figure 3.8 A cross-section of London's Westminster Abbey looking down the axis of the *nave*, showing the roof *vault* and the *flying buttresses* and their piers that support both roof vaults and nave walls. On the left (south) side there are *two* sets of flying buttresses, and the main buttressing piers are located beyond the *cloister* that abuts the church along that side (after Heyman, 1995).

had to be thicker to support both their own weight and the weight of the roofs supported by the vaults, which were in turn supported at the more-ßexible tops of the walls. Therefore, in sum, the width and height of cathedral naves had to be scaled back as overall church length (or size) was increased lest problems of lighting, ventilation, *and* structural safety become insoluble.

Problem 3.1. On what basis did the Lilliputians conclude that Gulliver needed 1728 times as much food as they did?

Problem 3.2. How would the LilliputiansÕ conclusion change if they had thought about the exchange of energy between a person (of any size) and the surrounding environment?

Problem 3.3. How do the surface area and volume of a sphere scale? Why? (*Hint*: Analyze spheres of radii 1 and R.)

Problem 3.4. Explain what would happen to an angle between two lines inscribed on a balloon as it was inßated to a radius R from a radius of 1.

Problem 3.5. ConÞrm that eq. (3.2) does adequately portray the straight line drawn in Figure 3.2.

Problem 3.6. Show how the equation $y = mx^b$ becomes a linear equation in a log-log plot.

Problem 3.7. Write eq. (3.6) in a form suitable for a log-log plot.

3.3 Size and Function–I: Birds and Flight

We now examine another set of empirical data, taken from a study of the aerodynamics of birds in ßight and displayed in Figure 3.9. It appears from this plot that a straight line can be penciled in to Þt the data, and it also seems that there is no data for bird weights greater than 35Ð40 pounds. We are thus prompted to ask two questions: Can the general form of this data be explained by dimensional analysis, along the lines of our discussions of Chapter 2? And, is there an upper limit to the weight of a ßying[1] bird? The answers to both questions are afÞrmative.

The answer to the Þrst question can be found by looking at the Þt of the straight line in the log-log plot of Figure 3.9. A close examination of the Þtted line shows that its slope is 1:3, which suggests that

$$\text{Weight} \propto (\text{Wing loading})^3. \tag{3.7}$$

But does eq. (3.7) make dimensional sense?

For birds that soar (e.g., gulls and buzzards), we argue that the lift forces needed to sustain them in the air should be proportional to the wing areas, or in dimensional terms, proportional to (length)^2. The *wing loading* is

[1] Remember that several creatures categorized as birds have never taken wing, including penguins and ostriches, so we really do need the adjective ÒßyingÓ

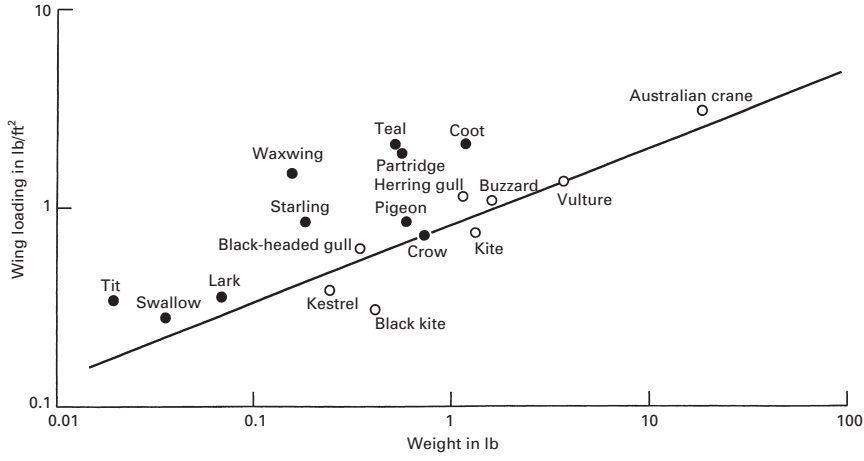


Figure 3.9 A set of empirical data, taken from a study of the aerodynamics of birds in flight (von Karman, 1954). It appears from this plot that a straight line can be penciled in to fit the data, and it also seems that there is no data for bird weights greater than 35–40 lb.

the load a bird has to carry, which is just its weight, which is proportional to its volume. Thus, in dimensional terms, the wing loading is proportional to $(\text{length})^3$. Then the wing loading per unit of wing area would be proportional to $(\text{length})^3/(\text{length})^2$, or to (length). Since the weight is proportional to $(\text{length})^3$ and the wing loading to (length), eq. (3.7) is dimensionally consistent.

The second question, about the existence of an upper bound to ßying weight, is harder to answer. We will answer it, but in the somewhat restricted domain of *hovering flight* because the aerodynamic arguments are simpler. We will formulate the problem by examining the dimensions of both the power *needed* to sustain hovering and the power *available* to sustain hovering.

3.3.1 The Power *Needed* to Hovering

A bird ßaps its wings when it is hovering. In so doing, the bird generates the needed hovering power by moving a mass of airÑ and so transferring momentumÑ downward. NewtonÕs second law says that the time rate of change of the momentum of that jet of air must be equal to the total lift force on the wings, which is, in turn, equal to the birdÕs weight. The mass of air moving through the jet can be estimated in terms of the air density, ρ,

the wing area, A, and the jet speed, v, as

$$\text{mass/time} = \rho A v. \tag{3.8}$$

The time rate of change of momentum is then just the product $v \times$ (mass/time), which is, again, equal to the bird weight, W:

$$W = v \times \text{mass/time} = \rho A v^2. \tag{3.9}$$

In view of the dimensional dependencies of the bird's weight and wing area, it follows from eq. (3.9) that the velocity of the air mass for hovering must be such that

$$v \propto L^{1/2}. \tag{3.10}$$

The power needed to sustain the hovering jet is equal to the time rate of change of the kinetic energy of the mass of air in the jet. Thus,

$$\text{power needed} \propto \tfrac{1}{2}\rho A v \times v^2. \tag{3.11}$$

In view of eqs. (3.10) and (3.11) taken together, the scaling of the power needed for a bird to hover scales with length according to:

$$\text{power needed} \propto L^{7/2}. \tag{3.12}$$

Equation (3.11) is valid for forward flight as well as hovering. It can be confirmed by more complete, more complex aerodynamic arguments.

3.3.2 The Power *Available* for Hovering

There are three ways we can estimate the power available to a bird to enable it to hover. We can estimate its heat loss during hovering, the rate at which its heart supplies oxygen, and the maximum stresses in its bones and muscles.

The *heat loss* estimate is simple, if not altogether compelling. Muscles turn chemical energy into mechanical energy at a 25% efficiency rate. The excess energy is dissipated as heat loss through the bird's surface area. The heat transfer thus decreases at a rate proportional to $(\text{length})^2$. Hence, in order to prevent the bird from overheating, the available power must also be proportional to L^2.

The *oxygen supply* estimate reduces to the consideration of the time rate of change of the volume of blood delivered by the heart. This volumetric rate is proportional to the cross-sectional area of the bird's blood vessels. Thus, we again find that the available power is proportional to L^2 because it is proportional to the oxygen flow, which is in turn proportional to the rate of blood delivery.

The *maximum stress* estimate begins with the assessment of the work done by a contracting muscle. By the principle of conservation of energy that work must equal the resulting change in the kinetic energy of the limb moved by the muscle's contraction. Thus,

$$\text{muscle force} \times \text{contraction} \propto \text{limb mass} \times v^2, \tag{3.13}$$

where v is now the speed of the moving limb. Since the force in the muscle is limited by the maximum tensile strengths of the bird's muscles and tendons, it must be proportional to L^2 as representative of the cross-sectional area of those muscles and tendons.

Now the muscle contraction is proportional to L, and the limb mass to L^3, so that eq. (3.13) tells us that the speed of the hovering bird is independent of L or size. If this is true, the time it takes for a muscle to contract would be found from the ratio L/v, or simply the length L. Then the power exerted by the muscle is

$$\text{power} \propto \frac{\text{muscle force} \times \text{contraction}}{\text{time}} \propto \frac{L^2 \times L}{L}, \tag{3.14}$$

so once again we find that the available power is proportional to L^2.

3.3.3 So There Is a Hovering Limit

We have seen in Section 3.3.1 that the power needed for flight is proportional to $L^{7/2}$, while in Section 3.3.2 we showed that the power available to the bird to sustain flight is proportional to L^2. Since the power needed to hover increases so much faster with the bird size, it is clear that a limit to hovering size must indeed exist.

Problem 3.8. Confirm the dimensional relationship of eq. (3.10).

Problem 3.9. Use dimensional analysis to confirm that power is equal to the time rate of change of kinetic energy.

Problem 3.10. Confirm the dimensional relationship of eq. (3.12).

Problem 3.11. Confirm that eq. (3.13) does show that v is independent of L.

3.4 Size and Function–II: Hearing and Speech

Human hearing and speech are areas of human physiology where scaling has interesting and important effects. Size, shape, and function are clearly

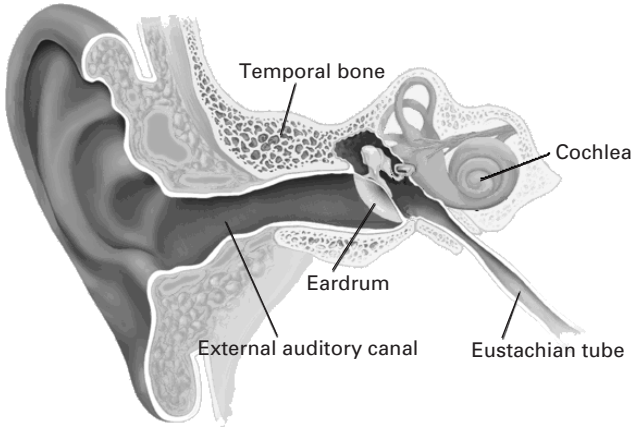


Figure 3.10 A cross-section of the human ear, including the eardrum, which is the mechanism with which we hear (from Encyclopedia Britannica Online, www.brittanica.com, 1997).

intertwined in the ear and eardrum (Figure 3.10), and in the vocal cords and *larynx*, that is, the Òvoice boxÓthat contains the vocal cords (Figure 3.11). We are inclined to wonder about scale effects in hearing because we know that humans hear sounds in the range of 20 to 20,000 Hz (or hertz or cycles per second), dogs hear sounds that have frequency components up to 50,000 Hz, and bats hear sounds as highs as 100,000 Hz. The unit hertz is named after the acoustician Gustav Ludwig Hertz (1887Ð1975). Since larger animals seem to have more limited frequency ranges, it is worth exploring whether size could play a role in these differences.

3.4.1 Hearing Depends on Size

The eardrum is just one part of a complex hearing apparatus (see Figure 3.10) that starts at the outer ear and goes through the cochlea to the auditory nerve that transmits signals to the brain for interpretation. When a sound is generated by a source, the result is that air (or another mediumÕs) particles immediately adjacent to the source are set into motion, creating the acoustic signals that are transmitted through the intervening air (or medium, or media) to the receiverÕs ear. The eardrum itself converts the mechanical vibration of the Òincoming Óair particles that form the acoustic signal into a mechanical vibration of three bonesÑ called the *hammer*, *anvil*, and *stirrup*Ñ that in turn carry the vibratory signal into the inner ear. Eventually, the inner ear converts these mechanical signals into electrical signals that are transmitted through the nervous system to the

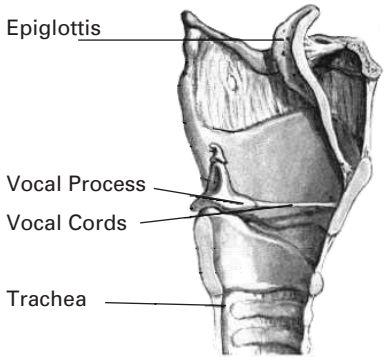


Figure 3.11 A cross-section of the human *larynx* or "voice box" showing the vocal cords that are the mechanism with which we speak (www.sfu.ca/~saunders/L33098/L5/L5Fset.html, 2002, by courtesy of R. Saunders, Simon Fraser University, Burnaby, British Columbia, Canada).

brain by way of the organs of Corti. As the Þrst pickup of the incoming mechanical signal, it is important that the eardrum remain quite stiff in order to pick up and accurately reproduce the higher frequencies of that signal.

In mechanical terms, the eardrum is a stretched membrane, much like a trampoline. Like every other mechanical device, the eardrum has *natural frequencies* at which it can vibrate freely (and indeÞnitely, if only there were no damping!). As we will see in Chapter 8, an elastic system responds just like a linear spring when it is forced or excited at frequencies below the lowest natural frequency, sometimes called the *fundamental frequency*. Thus, if we want the eardrum to be stiff, we want its fundamental frequency to be very high. It turns out that the fundamental frequency of a stretched circular membrane of radius, r, and thickness, h, is given by

$$f_{membrane} = \frac{2.40}{2\pi r}\sqrt{\frac{F}{\rho h}}, \tag{3.15}$$

where F is the tensile (stretching) force per unit length of the membrane circumference and ρ is the mass density of the material of which the membrane is made. It is easily veriÞed that the dimensions of the membrane's fundamental frequency are 1/(time) or 1/T, which is quite appropriate for

frequency. However, it is also interesting to note in eq. (3.15) that the fundamental frequency varies inversely with the radius, r, of the membrane (or eardrum). So, for similar values of the tensile force, F, and the mass density, ρ, we would expect the range of hearing to extend into higher frequencies for smaller animals, and this is just what we have seen in the hearing ranges of humans, dogs, and bats.

3.4.2 Speech Depends on Size

A similar situation exists with regard to human vocal cords and voice boxes. We know from everyday experience that men generally have deeper, lower-pitched voices than do women, and we are also accustomed to the facts that birds chirp and bears growl. So we are inclined to imagine that the sound of speech would scale with size.

The mechanism that creates speech is the forced vibration of the vocal cords as air is expelled from the lungs and pushed past (and through) the cords in the *larynx* or voice box (viz., Figure 3.11). In order to develop and produce volume at low frequencies, the vocal cords must be able to vibrate at low frequencies, and the voice box must be able to amplify the low-frequency signals produced by the vocal cords.

The vibration characteristics of vocal cords can be modeled just as we would model the vibration of violin or piano strings, whose fundamental frequency is given by

$$f_{string} = \frac{1}{2l}\sqrt{\frac{F}{\rho A}}, \tag{3.16}$$

where l is the string length, A its cross-sectional area, F is the tensile force applied to the string, and ρ its mass density. Note the very strong resemblance between eqs. (3.15) and (3.16). Further, we see that this frequency (3.16) scales inversely with both the string length and its mass density. Thus, a larger animal with longer and more dense vocal cords will make sounds that have components at lower frequencies.

We can also look at the fundamental frequency of an *acoustic cavity* as a model for the larynx. Such a cavity, which we examine in more detail in Chapter 8, is also called an *acoustic resonator* and it has a fundamental frequency given by [see eq. (8.47)]

$$f_{cavity} = \frac{c_0}{2\pi}\sqrt{\frac{A}{lV_0}}, \tag{3.17}$$

where A and l are, respectively, the area and the length of the "neck" leading into an acoustical cavity of volume V_0 that is filled with a gas in which

sound waves travel at a speed c_0 (see Figure 8.7). Clearly, the fundamental frequency of the cavity scales inversely with the cavity's volume. So once again we find that larger humans and animals have deeper voices than do their smaller counterparts.

Problem 3.12. What is the hearing range of an elephant? A whale? How do these ranges compare with those of humans?

Problem 3.13. Confirm that the dimensions of eq. (3.15) are $1/T$.

Problem 3.14. Confirm that the dimensions of eq. (3.16) are $1/T$.

Problem 3.15. Confirm that the dimensions of eq. (3.17) are $1/T$.

3.5 Size and Limits: Scale in Equations

In Section 3.3, while discussing size and function, we found that there is an upper limit to the weights of hovering birds. This limit is due to the fact that birds could not supply enough power to sustain hovering flight as they grew bigger and heavier. Thus, the birds' ability to hover was limited by the power available to them. Limits occur quite often in mathematical modeling, and they may control the size and shape of an object, the number and kind of variables in an equation, the range of validity of an equation, or even the application of particular physical models—or "laws," as they are often called.

Modern electronic components and computers provide ample evidence of how limits in different domains have changed the appearance, performance, and utility of a wide variety of devices. The bulky radios that were made during the 1940s, or the earliest television sets, were very large because their electronics were all done in old-fashioned circuits using vacuum tubes. These tubes were large and threw off an enormous amount of heat energy. The wiring in these circuits looked very much like that in standard electrical wiring of a house or office building. Now, of course, we carry television sets, personal digital assistants (PDAs), and wireless telephones on our wrists. These new technologies have emerged because we have learned to dramatically change the limits on fabricated electrical circuits and devices, and on the design and manufacturing of small mechanical objects. And this is true beyond electronics. The scale at which surgery is done on people has changed because of our ability to "see" inside the human body with greater resolution—with increasingly sophisticated scans and imagers, as well as with fiber-optic television cameras—and to design visual, electronic, and mechanical devices that can operate inside a human

eye, and in arteries and veins. In the emerging Þeld of *nanotechnology* we are learning to engineer things at the molecular level. Thus, our mathematical models will change, as will the resulting devices and Ònmachines.Ó

3.5.1 When a Model Is No Longer Applicable

As we hinted in Section 3.1, one interesting example of the interaction of scale and limits is Newtonian mechanics. We are accustomed to taking the masses or weights of objects as constants in our everyday lives and in our normal engineering applications of mechanics. We do not expect a box of candy to weigh any more whether we are standing still, riding in a car at 120 km/hr (75mph), or ßying across the country at 965 km/hr (600 mph). Yet, as we noted in Section 3.1, according to the general theory of relativity, the mass of a particle moving at speed, v, is given as a (dimensionless) fraction of the rest mass, m_0, by

$$\frac{m}{m_0} = \frac{1}{\sqrt{1-(v/c)^2}}, \tag{3.18}$$

where c is the speed of light (3×10^8 m/s = 186,000 mi/sec). For the box of candy ßying across the country at 965 km/hr = 268 m/s, the factor in the denominator of the relativistic mass formula (3.18) becomes

$$\sqrt{1-\left(\frac{v}{c}\right)^2} = \sqrt{1-7.98\times10^{-13}} \cong 1 - 3.99\times10^{-13} \cong 1. \tag{3.19}$$

Clearly, for our practical day-to-day existence, we can neglect such relativistic effects. However, it remains the case that Newtonian mechanics is a good model only on a scale where all speeds are very much smaller than the speed of light. If the ratio v/c becomes sufÞciently large, the mass can no longer be taken as the constant rest mass, m_0, and Newtonian mechanics must be replaced by relativistic mechanics.

3.5.2 Scaling in Equations

In certain situations, scaling may shift limits or perhaps points on an objectÕs boundary where *boundary conditions* are applied. For example, suppose we want to approximate the hyperbolic sine function,

$$\sinh x = \frac{1}{2}(e^x - e^{-x}). \tag{3.20}$$

We know that for large values of x, the term e^x will be much larger than the term e^{-x}. The approximation problem is one of deÞning an appropriate

criterion for discarding the smaller term, e^{-x}. For dimensionless values of x greater than 3, the second term on the right-hand side of eq. (3.20), e^{-x}, does become very small (less than 4.98×10^{-2}) compared to e^{x} for $x = 3$, which is 20.09. Hence, we could generally take $\sinh x \cong (\frac{1}{2})e^{x}$. All we have to do is decide on a value of x for which we are willing to accept the approximation $e^{2x} - 1 \cong e^{2x}$.

We can also approach this problem by introducing a *scale factor*, λ, after which we can look for values of x for which we can make the approximation

$$\sinh(x/\lambda) \cong \frac{1}{2}e^{x/\lambda}. \tag{3.21}$$

Putting a scale factor, λ, in the approximation of eq. (3.21) obviously means that it will affect the value of x for which that approximation is acceptable. Now the comparison is one in which we want

$$e^{2x/\lambda} - 1 \cong e^{2x/\lambda}. \tag{3.22}$$

For $\lambda = 1$, the approximation is good for $x \geq 3$, while for $\lambda = 5$ the approximation works for $x \geq 15$. Thus, by introducing the scale factor λ we can make the approximation valid for different values of x because we are now saying that $e^{-x/\lambda}$ is sufficiently small for $x/\lambda \geq 3$. Changing λ has in effect changed a boundary condition because it has changed the expression of the boundary beyond which the approximation is acceptable to $x \geq 3\lambda$.

Recall that functions such as the exponentials of eqs. (3.21) and (3.22), as well as sinuosoids and logarithms, are *transcendental functions*. Transcendental functions can always be represented as power series, as we will detail in Section 4.1.2. For example, the power series for the exponential function is:

$$e^{x/\lambda} = 1 + \frac{x}{\lambda} + \frac{1}{2!}\left(\frac{x}{\lambda}\right)^2 + \frac{1}{3!}\left(\frac{x}{\lambda}\right)^3 + \cdots + \frac{1}{n!}\left(\frac{x}{\lambda}\right)^n + \cdots \tag{3.23}$$

It is clear that the argument of the exponential must be dimensionless because without this property eq. (3.23) would not be a rational equation. Further, we could not calculate numerical values for the exponential function, or any other transcendental function, if its argument was not dimensionless. The presence of a scale factor in eq. (3.22) makes the exponential's argument dimensionless, and so numerical calculations can be performed.

In addition, the scale factor, λ, often represents a characteristic aspect of the problem being modeled, so that a ratio such as x/λ becomes a useful measure of whether something is truly large or small. For example, the hyperbolic sinusoid in eq. (3.20) might describe the deflection or downward displacement of a catenary cable as a function of its length.

The variable x could be a coordinate measured along the projected cable length and λ could represent its total projected length, which could be regarded as the cable's *characteristic length*.

3.5.3 Characteristic Times

We often see rate effects in first-order differential equations (a brief review of which can be found in Section 5.2.2). For example, it will be shown that a charged capacitor draining through a resistor produces a voltage drop $V(t)$ at a rate proportional to the actual value of the voltage at any given instant. The mathematical model would be:

$$\frac{dV(t)}{dt} = -\lambda V(t). \tag{3.24}$$

We can rewrite this equation in the equivalent form

$$\frac{dV(t)}{V(t)} = -\lambda dt. \tag{3.25}$$

Now, in order for this rate equation to be a rational equation, the net dimensions of each side of the equation must be the same. For eq. (3.25) that means each side must be dimensionless. The left-hand side is clearly dimensionless because it is the ratio of a voltage change to the voltage itself. The right-hand will be dimensionless only if the scale factor, λ, has physical dimensions such that $[\lambda] = 1/\mathsf{T}$. We will soon see this below and then will reconfirm it when we solve the differential equation (3.24) in Chapter 5.

We can use the dimensionless product λt to derive a measure of the time that it takes to discharge the capacitor being modeled. For example, we could define a *decay time*, often called a *characteristic time*, as the time it takes for the voltage to decrease to a specified fraction of its initial value. Suppose we choose that specified value to be 1/10. The characteristic or decay time of the charged capacitor would then be

$$V(t_{decay}) \equiv \frac{V_0}{10}. \tag{3.26}$$

How would we calculate t_{decay}? As we will show in Chapter 6, it is easily confirmed that the solution to the differential equations (3.24) and (3.25) is

$$V(t) = V_0 e^{-\lambda t}, \tag{3.27}$$

which in view of eq. (3.26) means that

$$\lambda \cong \frac{2.303}{t_{decay}}. \tag{3.28}$$

Equation (3.28) simply says that the scale factor λ for the discharging capacitor is inversely proportional to the characteristic (decay) time, and that the voltage in the capacitor can then be written as

$$V(t) \cong V_0 e^{-2.303(t/t_{decay})}. \tag{3.29}$$

Problem 3.16. Under what conditions is eq. (3.24) dimensionally consistent?

Problem 3.17. ConPrm that the voltage of eq. (3.27) does satisfy eq. (3.24).

Problem 3.18. ConPrm that eq. (3.28) is correct.

3.6 Consequences of Choosing a Scale

Since all actions have consequences, it should come as no surprise that the acquisition of experimental data, its interpretation, and its perceived meaning(s) generally can be very much affected by the choice of scales for presenting and organizing data.

3.6.1 Scaling and Data Acquisition

Scales affect the ways in which data is taken during experiments. Carefully chosen scales can reduce errors, save time and money, and they can highlight important details.

Consider, for example, the simple apparatus shown in Figure 3.12, which can be used to determine the *rotational inertia*, I_{rot}, (the second moment of inertia of the mass around an axis through its centroid or center) of the wheel shown as it turns or spins around an axis through its center. This experiment uses a falling weight connected to the wheel by a string to produce a torque that, in turn, causes the wheel to rotate. That torque, τ, is related to the rotational inertia I_{rot} by

$$I_{rot} = \frac{\tau}{\alpha}, \tag{3.30}$$

where α is the angular acceleration of the wheel, measured in units of radians per second squared (rad/sec^2). As we describe the inßuence of scale on experimental observation, we will focus on the angular acceleration as the important parameter through which we can determine I_{rot}. We conduct

the experiment itself by letting the falling weight cause the wheel to spin, during which we measure or read the speed, v, of any point on the wheel both as the experiment begins at some time $t = t_0$, and at a later time ($t = t_f$) that denotes the end of the experiment. The angular acceleration is then calculated in terms of the wheel's radius, R, and the measured speeds and measurement times as:

$$\alpha_{exp} = \frac{(v_f - v_0)}{R(t_f - t_0)}, \tag{3.31}$$

where the speeds v_0 and v_f are measured at the times t_0 and t_f, respectively.

Clearly, the time scale for this experiment is the time interval $t_f - t_0$. It will control the amount of error between the experimentally determined value of I_{rot} and its actual (or theoretically calculated) value.

We know that the wheel is set into motion by releasing or dropping the falling weight, because that action pulls the string taut and causes the wheel to start spinning. As the weight falls, the wheel rotates at an increasingly faster rate. Since the wheel is at rest when we initiate each experimental run, we can safely take $t_0 = v_0 = 0$. Then the values of α_{exp} determined experimentally are found from eq. (3.31) as

$$\alpha_{exp} = \frac{v_f}{Rt_f}. \tag{3.32}$$

Now, while we have argued above that the apparatus shown in Figure 3.12 produces a constant acceleration, that is not exactly true. Since we are starting from the state $t_0 = v_0 = 0$, static friction must be overcome as the wheel starts from rest at the beginning of each run of the experiment. After a short while, the wheel motion does, in fact, settle into spinning with a fairly constant acceleration. But what exactly is a "short while"? How do we know the correct value of t_f at which we can terminate each experimental run? Is 2 seconds enough time? Or do we need 4 seconds, or a still longer time?

In Table 3.1 we show some data obtained in one run of this experiment. Note that the number of revolutions or spins of the wheel goes up rapidly as time elapses, as does the speed of rotation. Further, and most importantly, if we calculate the angular acceleration as it varies with time (or with the estimated number of revolutions, a number that we can also count), we see that α_{exp} appears to approach a constant value (which means the torque also approaches a constant value). Why is this so? It is so because when we allow the experiment to run for a longer time (or through more turns of the wheel), we are changing the time scale over which the drag due to static friction has an influence. In a very short experiment, the time taken to overcome static friction takes up a much larger percentage of the time scale

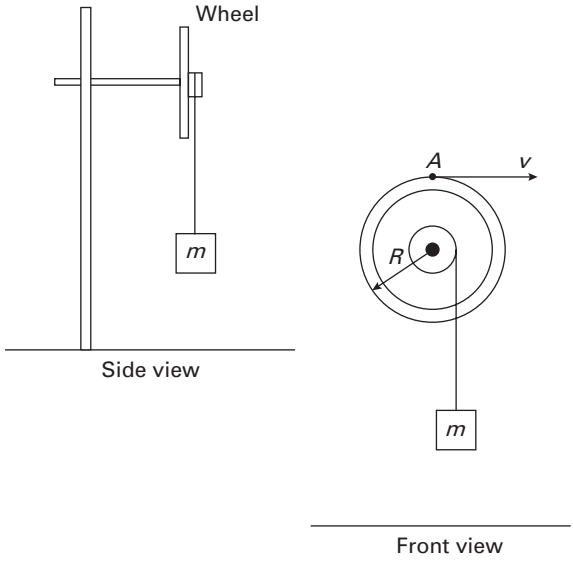


Figure 3.12 A simple piece of apparatus that can be used to measure the rotational inertia of a wheel of radius, R, as it spins around an axis through its center.

of the experiment, and so it has a disproportionate inßuence. In a longer experiment, conditions approach a steady state in which the predominant effect is the torque applied by the falling weight, so the static friction occupies an increasingly smallÑ and negligibleÑ part of the experimentõs run time.

Table 3.1 The data taken in the experimental determination of the rotational inertia of the wheel (as shown in the apparatus of Figure 3.12), along with an estimate of the actual number of revolutions that had occurred when v_f was measured.

t_f (s)	Estimated number of revolutions	Measured v_f (m/s)	Calculated $\alpha_{exp} = v_f/Rt_f$ (rad/s^2)
2	1/5	0.55	0.55
6	2.4	2.47	0.82
10	7	4.48	0.90
20	30	9.50	0.95
100	790	49.68	0.99

As another illustration of how scaling affects data acquisition, consider the diagnosis of a malfunctioning electronic device such as an audio amplifier. Such amplifiers are designed to reproduce their electrical input signals without any distortion. The outputs are distorted when the input signal has frequency components beyond the amplifier's range, or when the amplifier's power resources are exceeded. Distortion also occurs when an amplifier component fails, in which case we must diagnose the failure to identify the particular failed component(s).

A common approach to doing such diagnoses is to display (on an oscilloscope) the device's output to a known input signal. If the device is working properly, we expect to see a clear, smooth replication of the input. One standard test input is the square wave shown in Figure 3.13 (a). A nice

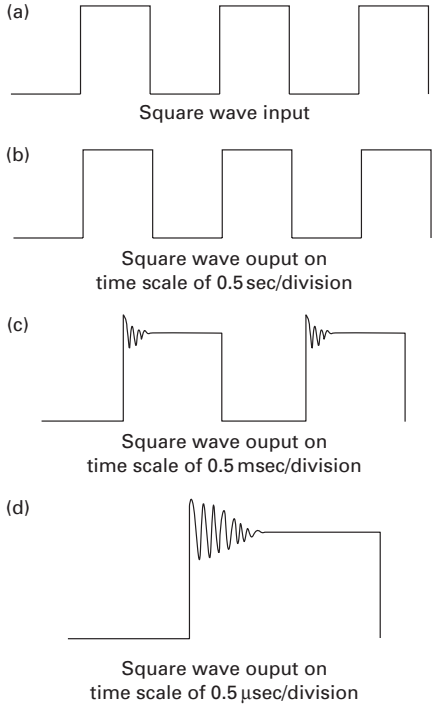


Figure 3.13 A square wave (a) is the input signal to a (hypothetical) malfunctioning electronic device. Traces of the output signals are shown at three different time scales (i.e., long, short, shorter): (b) 0.5 second/division; (c) 0.5 millisecond/division; and (d) of 0.5 microsecond/division.

replication of that square wave is shown in Figure 3.13(b), and it seems just Þne until we notice that the horizontal time scale is set at a fairly high value, that is, 0.5 second/division. To ensure that we are not overlooking something that might not show up on this scale, we spread out the same signal on shorter time scales of 0.5 millisecond/division (Figure 3.13(c)) and 0.5 microsecond/division (Figure 3.13(d)), neither of which is the nice sinusoid we originally thought. This suggests that the device is malfunctioning. Had we not set the oscilloscope to shorter, more appropriate time scales, we might have come to an erroneous conclusion.

3.6.2 Scaling and the Design of Experiments

Scale also affects the ways in which experiments are designed, especially when the context is that of ensuring that models replicate the prototypes or ÒrealÓ artifacts that they are intended to stand for or model. This aspect of scaling is, as we will now show, intricately intertwined with the notions of dimensional analysis discussed in Chapter 2.

Scale models or reproductions of physical phenomena or devices are usedÑ as they have been for quite some timeÑ to do experiments and study behavior for which a comprehensive analytical model is not available. Often such studies are done because a laboratory experiment is more easily developed than is a full-scale experiment. For example, it is easier to study the vibration characteristics of a model of a proposed bridge design than it is to build the designed bridge and hope for the best, just as it is easier to test models of rockets in simulated spaceßight or models of buildings in simulated earthquakes or Þres. But such experimental models won't be of much use unless some preliminary analysis is done and clear physical hypotheses are developed in advance. We will illustrate how is that done with one simple example.

Consider a simple beam, such as the one shown in Figure 3.14.

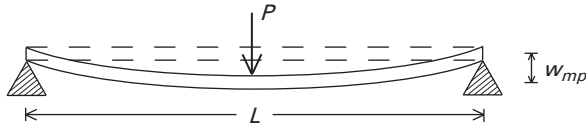


Figure 3.14 Prototype and model of a simple elastic beam of length, L, elastic modulus, E, and second moment of cross-section, I, as it deflects an amount, w_{mp}, at its center due to the load of magnitude, P, applied there.

We assume that it is known that the deßection, $w(L/2)$, of the midpoint of the beam when concentrated load P is applied at the same point is a function of the load and three other parameters, that is,

$$w(L/2) \equiv w_{mp} = f(P, EI, L), \tag{3.33}$$

where L is the beam's length, E is the elastic modulus of the beam material, and I is the second moment of its cross-sectional area. The product EI is commonly called the beam's *bending stiffness*. This example has two dimensionless groups (see Problem 3.19):

$$\Pi_1 = \frac{w_{mp}}{L}, \quad \Pi_2 = \frac{PL^2}{EI}. \tag{3.34}$$

Thus, it follows that

$$\frac{w_{mp}}{L} = f\left(\frac{PL^2}{EI}\right). \tag{3.35}$$

Suppose we want to determine the functional form of eq. (3.35) for a beam, which we will call the *prototype* beam, but that the beam is too big and heavy, and the load P too large, for us to do an experiment on the beam itself. We propose instead to test a *model* beam. But then we immediately face a question: How should the properties of the model beam relate to those of the prototype? The answer lies in the results obtained by applying the principles of dimensional analysis: The model properties and prototype properties must be such that the two dimensional groups have the same numerical values for both model and prototype. Stated in mathematical terms, with subscripts m for model and p for prototype,

$$(\Pi_1)_m = (\Pi_1)_p, \quad (\Pi_2)_m = (\Pi_2)_p. \tag{3.36}$$

Thus, to a certain extent we can scale the geometry, the material, or the load for our own convenience, but we cannot scale all of the independent variables independently. In order to preserve the property of *complete similarity* between model and prototype, we must preserve the equality between model and prototype of each dimensionless group needed to deÞne a particular problem.

Applying the general similarity rule of eq. (3.36) to the speciÞc case of the beam whose dimensionless groups are given in eq. (3.34), we Þnd that we can preserve complete similarity by requiring that

$$\left(\frac{w_{mp}}{L}\right)_m = \left(\frac{w_{mp}}{L}\right)_p, \quad \left(\frac{PL^2}{EI}\right)_m = \left(\frac{PL^2}{EI}\right)_p. \tag{3.37}$$

Having established in eq. (3.37) the overall conditions needed for complete similarity, we can now go into further detail to see both what we *must* do

and what we *may* do in terms of *scaling factors* deÞned for each of the problemÕs variables, that is, for the factors

$$n_w = \frac{(w_{mp})_p}{(w_{mp})_m},\ n_P = \frac{P_p}{P_m},\ n_E = \frac{E_p}{E_m},\ n_I = \frac{I_p}{I_m},\ n_L = \frac{L_p}{L_m}. \qquad (3.38)$$

Thus, we see that the scaling factors in eq. (3.38)Ñ which should not be confused with the graphical scale factors, λ, introduced in Section 3.5.2Ñ are simply ratios of the values of each of the variables in the prototypes to the values of the same variable in the model. Equation (3.38) shows that we have Þve such scaling factors for this problem, while eq. (3.37) shows that there are two overall similarity conditions that must be satisÞed. We can, in fact, write the similarity conditions (3.37) in terms of the scaling factors (3.38) by straightforward substitution:

$$\frac{n_w}{n_L} = 1, \quad \frac{n_P n_L^2}{n_E n_I} = 1. \qquad (3.39)$$

So, if we choose a length scale (n_L) for this problem, we have also chosen a deßection scale (n_w) by the Þrst of eq. (3.39). However, this means that we may still freely choose two of the three remaining scaling factors (n_P, n_E, and n_I). If we chose the scaling factors of the elastic modulus (n_E) and of the moment of inertia (n_I) because we had appropriate materials or small beams lying around our laboratory, then the single remaining scaling factor n_P would be determined by the second of eq. (3.39):

$$n_P = \frac{n_E n_I}{n_L^2}. \qquad (3.40)$$

Suppose we wanted to model the deßection of a steel beam by doing experiments on a small model made of balsa wood. Assume a typical laboratory scenario in which the length scale is twenty-to-one, that is, $n_L = 20$, the scaling factor of the moments of inertia is about $n_I = 1000$, and that the scaling factor of the moduli of elasticity is approximately $n_E = 50$. For a similar experiment, we would then expect that the resulting deßection would be one-twentieth of the actual deßection when we apply a load to the model that is equal to the anticipated actual load divided by 125.

Note that this introduction to the consequences of scaling in modeling is just that, a very short and very limited introduction. Clearly, not all experiments are so easily analyzed or scaled, and so there are many more issues to be explored in a comprehensive look at scaling in the design of experiments.

3.6.3 Scaling and Perceptions of Presented Data

The scales used to present modeling ÒresultsÓ also signiÞcantly inßuence how such data is perceived, no matter whether those models are analytical or experimental in nature. Indeed, individuals and institutions have been known to choose scales and portrayals to disguise or even deny the realities they purport to present. Thus, whether by accident or by intent, scales can be chosen to persuade. While this is more of a problem in politics and the media than it is in the normal practice of engineering and science, it seems useful to touch on it brießy here since the underlying issue is a consequence of scale.

We start by reconsidering some calculations we have already performed (in Section 3.5.2) to show how we can use a scale factor to effectively move

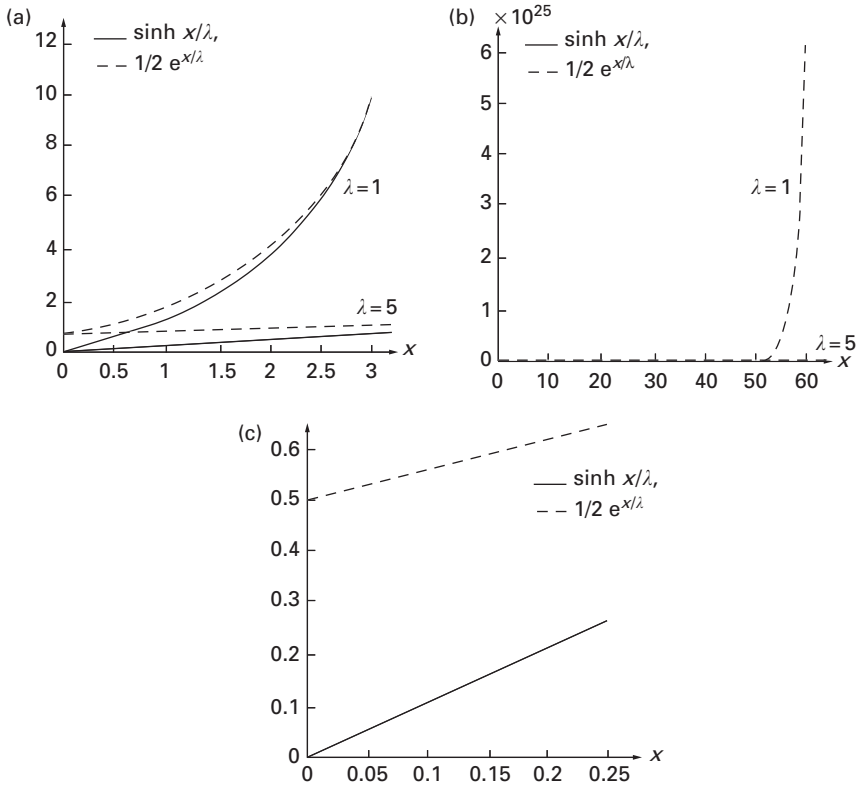


Figure 3.15 Plots of sinh (x/λ) (solid line) and its one-term exponential approximation (1/2) exp (x/λ) (dashed line): (a) for $\lambda = 1, 5$, with length scale $0 \leq x \leq 3.0$; (b) for $\lambda = 1, 5$, with length scale $0 \leq x \leq 60$; and (c) for $\lambda = 1$, with length scale $0 \leq x \leq 0.25$.

a boundary. In Figure 3.15(a) we display plots of $\sinh(x/\lambda)$ (solid line) and its one-term exponential approximation (dashed line). We now see what we have previously described, namely, for $\lambda = 1$ the approximation is good for $x \geq 3$, while for $\lambda = 5$ the approximation works for $x \geq 15$. The scale factor λ makes the approximation valid for different values of x because of the argument that $e^{-x/\lambda}$ can be neglected when compared to 1 for $x/\lambda \geq 3$.

The same two functions have been redrawn in Figure 3.15(b) where the horizontal scale has been very much contracted, as a result of which we don't see any difference between the hyberbolic sinusoid and its elementary approximation. That is, it looks like $\sinh(x/\lambda)$ and $1/2e^{x/\lambda}$ are the same for all values of x, when we know that is not the case. In other words, we have lost (or hidden) some information about the behavior at small values of x. To emphasize this, we show in Figure 3.15(c) a plot for the case $\lambda = 1$ with a much-elongated horizontal scale where, as a result, the difference between the two functions is very much exaggerated.

Lastly on graphical display, scaling, and perception, we show in Figures 3.16 and 3.17 two illustrations of the consequences of scale in contexts somewhat beyond the normal professional concerns of engineers and scientists. We show both examples because they use the same technique of carefully choosing a scale in a figure in order to present data out of context. In Figure 3.16(a) we show a rather dated picture of traffic deaths in the state of Connecticut during the time interval 1956–1957, and we see that a sharp drop in traffic deaths occurred then. But, was that drop real? And, in comparison to what? It turns out that if more data are added, as in Figure 3.16(b), we see that the drop followed a rather precipitous increase in the number of traffic fatalities. Further, if we added data from adjacent states and normalized the number of deaths against a common base, as shown in Figure 3.16(c), we then find that the numbers of Connecticut's traffic fatalities was similar to those of its neighbors, although the impact of the stricter enforcement is still visible after 1955.

Similarly, one of the most often shown graphics in the financial pages of newspapers, or in their televised equivalents, are graphics such as that shown in Figure 3.17 (see p. 65). Here, the immediate sense conveyed is that the bottom has dropped out of the market because the scale used on the ordinate (or y- or vertical axis) has been so foreshortened that it includes only one week's trading activities, Thus, a decline of a few percent in a stock market barometer such as the Dow Jones Industrial Average (DJIA) is made to look like a much steeper decline— especially if the curve itself is drawn in red ink!

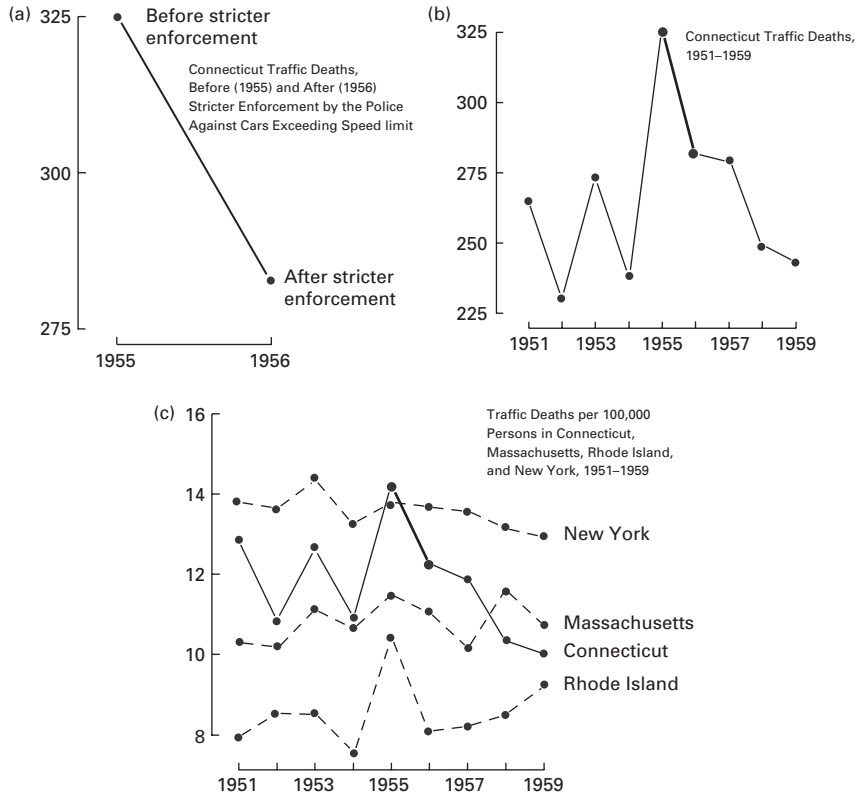


Figure 3.16 Plots of traffic fatalities in the state of Connecticut, showing the dangers of truncating scales and deleting comparative data: (a) Connecticut data for 1955–56; (b) Connecticut data for 1951–59; and (c) normalized data for Connecticut and three neighboring states for 1951–59 (from Tufte, 1983).

Problem 3.19. Show that the deßection w_{mp} of a beam with bending stiffness EI, length L, and under a concentrated load P is governed by the two dimensionless groups in eq. (3.34).

Problem 3.20. Why is the torque, τ, in the apparatus of Figure 3.12 a constant?

Problem 3.21. Expressed in terms of the wheelÕs geometric and gravitational properties, what is the magnitude of the torque in Problem 3.20?

Problem 3.22. ConÞrm that eq. (3.30) is dimensionally correct.

Problem 3.23. ConÞrm that eq. (3.31) is dimensionally correct.

Problem 3.24. Calculate and conÞrm the estimated number of revolutions in the last column of Table 3.1.

Problem 3.25. ConÞrm that eq. (3.39) is the correct representation of eq. (3.37) in terms of the Þve scaling factors of a simple beam.

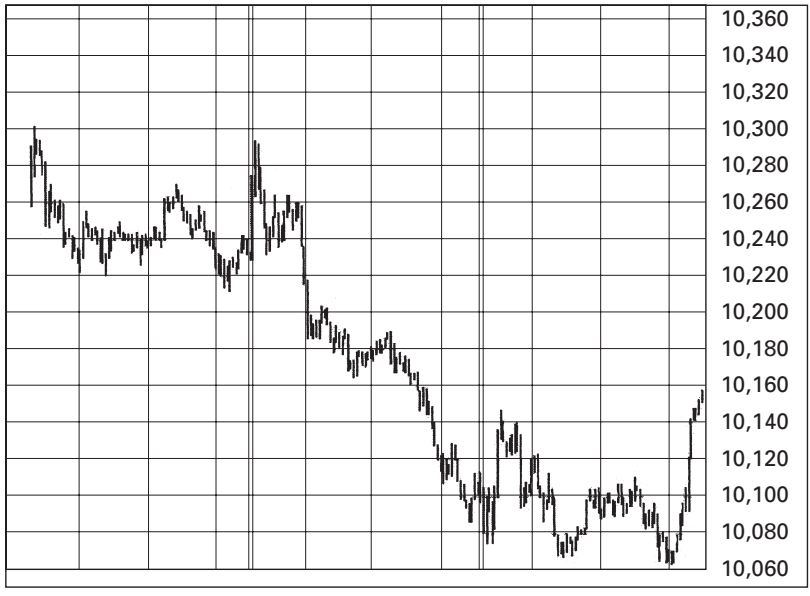


Figure 3.17 A plot of the performance of the New York Stock Exchange during 13–15 May 2002, as exemplified by that universally-cited barometer, the Dow Jones Industrial Average (DJIA) (www.bigcharts.com, 2002).

3.7 Summary

Continuing the discussion of issues involving dimensions that began in Chapter 2, here we have focused on very important effects of scale. We have shown how scaling effects inßuenced the growth of cathedrals and large churches, and we have demonstrated how size affects function in the ability of birds to hover and in peopleÕs ability to hear and to speak. In fact, we have shown that the nature of hearing and speech in animals is determined in large part by the relative size of the relevant parts of their anatomy.

We have also discussed the fact that scaling has a signiÞcant effect on experiments, both in terms of how data is acquired and how it is interpreted. The choice of scale(s) for experiments is a crucial part of the design of experiments. More generally, we have seen that the way that data is scaled for presentation can signiÞcantly inßuence how people perceive the meaning of that data. This is also a very important part of modeling because it speaks directly to the perceived credibility of the results of any modeling endeavor.

3.8 References

R. M. Alexander, *Size and Shape*, Edward Arnold, London, 1971.

W. Burns, *Noise and Man*, Lippincott, Philadelphia, 1973.

P. D. Cha, J. J. Rosenberg, and C. L. Dym, *Fundamentals of Modeling and Analyzing Engineering Systems*, Cambridge University Press, New York, 2000.

F. W. David and H. Nolle, *Experimental Modelling in Engineering*, Butterworths, London, 1982.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

A. A. Ezra, ÒScaling Laws and Similitude Requirements for Valid Scale Model Work,Ó in W. E. Baker (Ed.), *Use of Models and Scaling Shock and Vibration*, American Society of Mechanical Engineers, New York, 1963.

S. J. Gould, ÒSize and Shape,Ó *Harvard Magazine*, 78(2), 43Ð50, October 1975.

J. Heyman, *The Stone Skeleton*, Cambridge University Press, Cambridge, 1995.

L. E. Kinsler and A. R. Frey, *Fundamentals of Acoustics*, John Wiley, New York, 1962.

S. J. Kline, *Similitude and Approximation Theory*, McGraw-Hill, New York, 1965.

G. Murphy, *Similitude in Engineering*, Ronald Press, New York, 1950.

D. J. Schuring, *Scale Models in Engineering: Fundamentals and Applications*, Pergamon Press, Oxford, UK, 1977.

J. M. Smith, *Mathematical Ideas in Biology*, Cambridge University Press, London and New York, 1968.

D'A. W. Thompson, *On Growth and Form*, Cambridge University Press, London and New York, 1969. (Abridged edition, J. T. Bonner (Ed.))

E. R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT, 1983.
E. R. Tufte, *Envisioning Information*, Graphics Press, Cheshire, CT, 1990.
T. von Karman, *Aerodynamics*, McGraw-Hill, New York, 1954.

3.9 Problems

3.26. Formulate a hypothesis to explain why a wood pigeon and a buzzard seem to have such different ratios of W_{fm}/W_b in Figure 3.2.

3.27. Show that the equation that describes the log-log plot of Figure 3.7 can be found to be $h \cong 1.23\, l^{0.68}$, where h and l are, respectively, the nave height and church length rendered dimensionless by dividing each by 1 ft.

3.28. Using reasoning similar to that which brought us to eq. (3.13), show that the maximum speed at which animals can run is independent of size.

3.29. The velocity of blood in the aorta is related to the difference in pressure between the heart and the arteries. Find the relationship between the velocity of the blood and the pressure difference. (*Hint*: Use the work-energy theorem as we did for bird hovering in Section 3.3.2.)

3.30. The stilt, a little long-legged bird, was described in *Gulliver's Travels* as weighing 4.5 ounces and having legs that are 8 in long. A ßamingo has a similar shape and weighs 4 lb. Apply scaling arguments to show that ßamingoÕs legs should be about 20 in long (as they actually are!).

3.31. Given that a robin weighs about 2 ounces, could we scale the length of its legs from the stilt data given in Problem 3.30? Explain your answer.

3.32. A certain cucumber was found to have cells that divided when they had grown to 1.5 times the volume of Ò restingÓ cells. Cells normally divide so that the ratio between their surface and their mass remains constant. Is the cucumber described a Ò normalÓ cucumber?

3.33. Find the range of values of the variable x for which the approximation

$$\cosh(x/\lambda) \cong \frac{1}{2} e^{x/\lambda}$$

is acceptable, for scaling factors $\lambda = 1$ and $\lambda = 6$. Plot both functions for each of the two scaling factors.

3.34. An experiment to determine the natural or fundamental period of oscillation of a simple spring-mass system (see Figure P3.34) is set up as follows. A spring of stiffness k is Þxed at one end and connected to a mass m at its other, with the mass being able to move along an ideal, frictionless air track. The mass is displaced a distance x_0 from its initial resting position, after which it oscillates along the air track around that initial position. The time needed for a complete oscillationÑ that is, the periodÑ is measured several times for several periods in succession, with the results being compared to the theoretical formula for the period, T:

$$T = 2\pi\sqrt{\frac{m}{k}}.$$

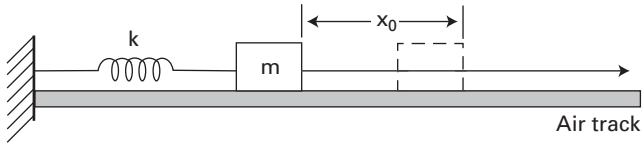


Figure P3.34 An experimental device for determining the period, T, of a spring-mass system, wherein the mass, m, moves on an ideal, frictionless air track.

Assuming that k and m are known, and that the timer used to measure the period is accurate to within $\pm 1\%$. What are the possible pitfalls that could prevent the successful experimental determination of T?

3.35. When the structural elements called *beams* vibrate freely, their natural frequencies, ω, depend on a beamÕs mass density, ρ, its modulus of elasticity, E, and its length, l, depth, h, and cross-sectional area, A. If a model and prototype are to be built of the same material and tested, and their lengths are scaled in the ratio 1:5, how will their natural frequencies relate? (*Hint*: Use dimensional analysis to determine the various dimensionless parameters that relate ω to the various beam properties.)

3.36. A steel beam of length of 20 cm is to be used to model a prototype timber beam whose span is 3.6 m.

(a) Verify that the dimensionless group containing the load, the modulus, and the length is P/EL^2.

(b) If the timber beam is to carry a load of 9000 N at a point 1.5 m from the left end, what load must be applied to the model to determine whether the prototype can carry its intended load? (Assume that the load-carrying capacity is the only behavior of interest here.)

3.37. The data given in the table immediately below were recorded as the growth of a colony of bacteria was observed. (a) Plot this data as a function of time. (b) Write an equation that expresses the bacterial population as a function of time.

Time (min)	Population (p) $\times 10^6$
0	10
5	15
10	22
20	50
30	110
40	245
50	546
60	1,215
70	2,704
80	6,018
90	13,394
100	29,810

4
Approximating and Validating Models

We devote this last chapter on fundamentals to discussions of elementary mathematical approximation techniques and of model testing and validation. Approximations are used to simplify both models (as we will see in Chapter 7 where the nonlinear model of the pendulum is simplified to obtain a linear estimate of the pendulum's behavior) and the numerical calculations made with the models. Such approximations and their numerical implementations introduce *error*, but the magnitudes of these errors can be estimated and limited. We will also discuss means of model validation: checking dimensions and units, testing qualitative behavior and limits, and applying basic statistics.

4.1 Taylor's Formula

Engineering and scientific calculations abound with mathematical approximations, in some measure because linear problems are easier to solve, but in larger measure because many of our linear models are validated and justified by experiment and by experience. Distinctions such as those between a linearized model and its full nonlinear counterpart also involve mathematical approximations such as those described in this section. How do we approximate a function to properly estimate the behavior it describes?

Many analytical approximations are derived from Taylor's formulas. Advanced numerical techniques such as the finite element method also use Taylor's formulas to approximate functions as polynomials with unknown coefficients that are determined numerically. Thus, we now review some basic results about Taylor's formula and series, including Taylor formulas of trigonometric functions and binomial expansions.

4.1.1 Taylor's Formula and Series

Any function that is continuous and has derivatives can, in general, be expanded into and approximated by a Taylor's formula. For values of the independent variable, x, in a region near $x = a$, a function $f(x)$ can be approximated by the polynomial

$$f(x) \cong f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{4.1}$$

where $f'(a)$ represents the first derivative of $f(x)$, $f''(a)$ the second derivative, and $f^{(n)}(a)$ the nth derivative of $f(x)$ evaluated at the point $x = a$. The series given in eq. (4.1) is called the *Taylor formula of f(x) in the neighborhood of the point* $x = a$. The point $x = a$ must be such that all derivatives of $f(x)$ exist there and are finite. In addition, and most important for this discussion, if the difference $(x - a)$ is very small, then we need only a few terms of the series (4.1) to render a good approximation of $f(x)$ in the neighborhood of $x = a$. The corresponding *Taylor's series* that renders the approximate equality in eq. (4.1) an exact equality is the limit of eq. (4.1) as $n \to \infty$:

$$f(x) = \lim_{n\to\infty} \left[f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \right]. \tag{4.2}$$

If we want to approximate the function $f(x)$ at another point, say $x = b$, we evaluate eq. (4.1) at that point to find Taylor's formula for $f(b)$:

$$f(b) \cong f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n. \tag{4.3}$$

If we use only the first term of eq. (4.3), we are approximating $f(b)$ as being equal to $f(a)$, as shown in Figure 4.1(a). If we use the first two terms of eq. (4.3), our approximation is improved by incorporating the effect of the slope change $f'(a)$, as shown in Figure 4.1(b). This value is closer to the true value than our simple one-term approximation. Our approximation is still further improved when three terms of the expansion (4.3) are used to approximate $f(b)$, as shown in Figure 4.1(c).

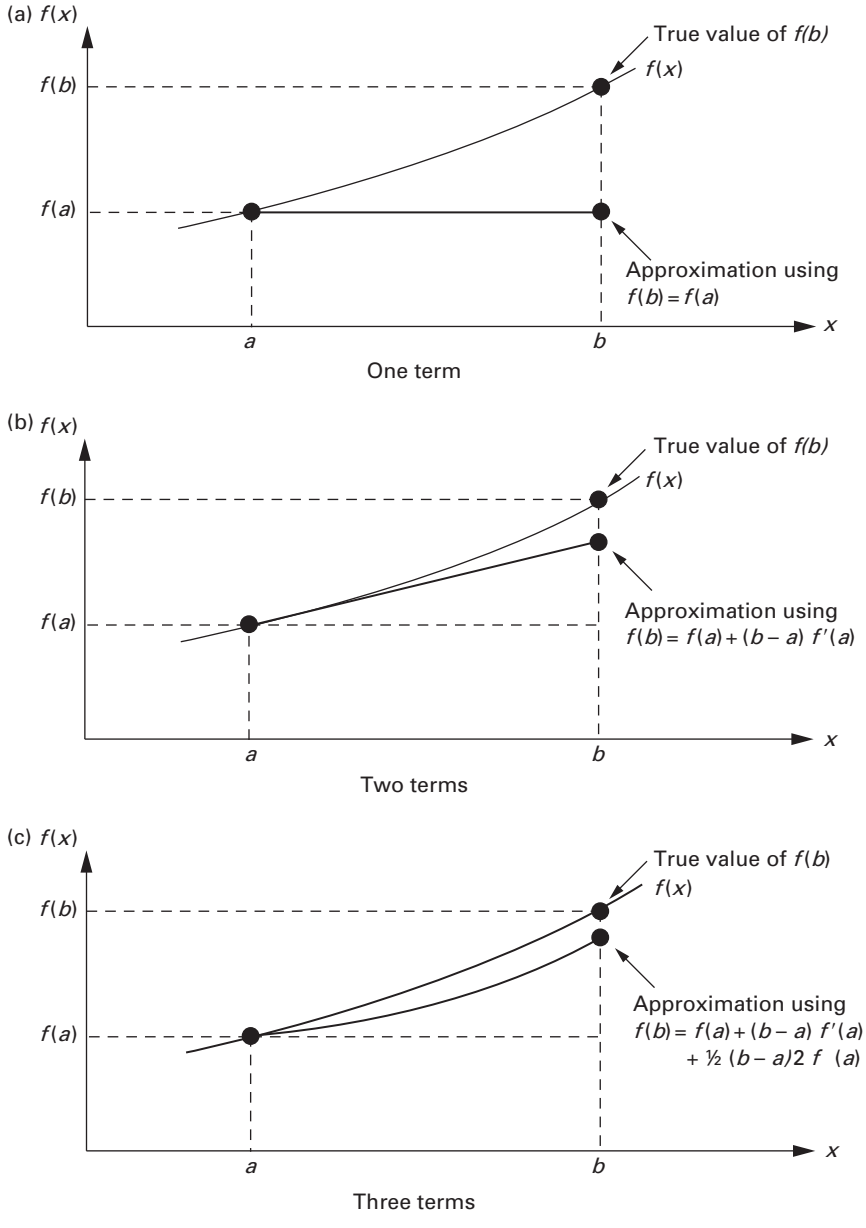


Figure 4.1 Improving the approximations obtained with a Taylor expansion by retaining more terms: (a) a one-term series approximation; (b) a two-term estimate; and (c) a three-term approximation. Note that the higher-order approximations depend on derivatives of $f(x)$ at the reference point of the Taylor series, $x = a$.

The accuracy of an approximation for any function $f(x)$ improves with the number of terms used in the expansion. Similarly, the approximation in eq. (4.1) can be turned into an exact formula like eq. (4.2) by adding a *remainder term* R_{n+1} to eq. (4.1):

$$f(x) = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\cdots+\frac{f^{(n)}(a)}{n!}(x-a)^n+R_{n+1}, \tag{4.4}$$

where the remainder term (which can be cast in several forms) is here shown as:

$$R_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x-a)^{n+1}\,. \tag{4.5}$$

The derivative in eq. (4.5) is calculated at a "suitably chosen" point ξ somewhere in the interval between a and x. Even though the precise location of ξ is not known, the remainder formula can be used to estimate the error made if a Taylor formula to order n is applied (see Problem 4.31). How many terms do we have to keep in a Taylor formula to ensure that the error is negligible, or at least acceptable? As we will see below, it depends on what we're trying to do, on the specifics of the model we're trying to build.

4.1.2 Taylor Series of Trigonometric and Hyperbolic Functions

The Taylor series expansions of the trigonometric functions for $a = 0$ are:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\,, \tag{4.6a}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\,. \tag{4.6b}$$

where x is expressed in (dimensionless) radians to ensure dimensional homogeneity. The corresponding Taylor expansions for the hyperbolic functions are:

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots\,, \tag{4.7a}$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots\,. \tag{4.7b}$$

We will now use a Taylor formula for the hyperbolic cosine (eq. (4.7b)) to estimate the sag of a tightly stretched string or cable that is weighted down only by its own weight. Such a cable is called a *catenary* after the

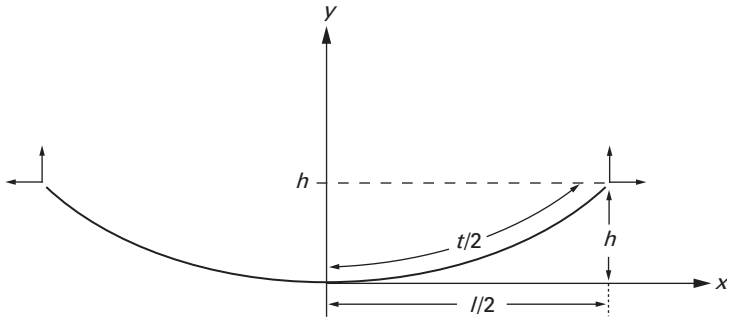


Figure 4.2 A long measurement tape stretched between two fixed points, A and B, for which the sag, h, is exaggerated. The mathematical model of the stretched tape is a hyperbolic cosine that can be approximated to varying degrees, depending on the relative magnitude of the ratio, h/l. This dependence signifies the fact that actual tape readings, t, must be corrected to properly measure the distance, l, on the ground.

Latin word for chain. Estimating the sag of a catenary may not sound all that interesting, but it does have a practical side that had been, until recently, a real engineering application. Until theodolites were introduced to measure large distances in construction projects, surveyors and engineers relied on tape measures. A surveyor's tape acts as a catenary because its only vertical load while measuring is its self-weight. We show such a tape in Figure 4.2, stretched between two supports at the same elevation that are separated by the length, l, with the cable's sag, h, exaggerated. Since $\cosh(0) = 1$, the equation of the catenary is

$$y(x) = c\left(\cosh\frac{x}{c} - 1\right), \tag{4.8}$$

where c is the *catenary parameter* and the coordinates of the *vertex* or low point of the cable are $(x = 0, y = c)$. The catenary parameter is a function of T_0, the (constant) horizontal component of the tension in the stretched cable, and of γ, the string's weight per unit length (see Problem 4.4). We see from Figure 4.2 that the sag is given by

$$h = y(l/2) = c\left(\cosh\frac{l}{2c} - 1\right). \tag{4.9}$$

Now we substitute the Taylor series (4.7b) of the hyperbolic cosine to find the sag:

$$h = c\cosh\frac{l}{2c} - c = c\left(1 + \frac{1}{2!}\frac{l^2}{4c^2} + \frac{1}{4!}\frac{l^4}{16c^4} + \frac{1}{6!}\frac{l^6}{64c^6} + \cdots - 1\right)$$
$$= c\left(\frac{1}{2!}\frac{l^2}{4c^2} + \frac{1}{4!}\frac{l^4}{16c^4} + \frac{1}{6!}\frac{l^6}{64c^6} + \cdots\right). \tag{4.10}$$

Note that this Taylor series for the sag has the correct physical dimensions since both c and h are measures of length and the ratio l/c is dimensionless, as it should be as the argument of the hyperbolic function. Further, for a tightly stretched string, the sag, h, is very small compared to the length, l, that is, $h/l \ll 1$. This suggests that the ratio $l/2c$ is also quite small compared to 1 because a one-term approximation of eq. (4.9) is found by retaining only the first term in the last of eq. (4.10):

$$h \cong c\left(\frac{1}{2!}\frac{l^2}{4c^2}\right) = \frac{l^2}{8c}. \tag{4.11}$$

Equation (4.11) confirms the suggestion that large values of the dimensionless catenary parameter, $2c/l$, correspond to small values of the dimensionless sag, h/l, because this result can be arranged as:

$$\frac{2c}{l} = \frac{l}{4h} \gg 1. \tag{4.12}$$

Further, had we approximated the hyperbolic cosine for small values of $l/2c$ independently of eqs. (4.9) and (4.10), we would have calculated that

$$c\cosh\frac{l}{2c} \cong c\left(1 + \frac{1}{2!}\frac{l^2}{4c^2} + \frac{1}{4!}\frac{l^4}{16c^4} + \frac{1}{6!}\frac{l^6}{64c^6}\right) \cong c, \tag{4.13}$$

and we would then have found, quite mistakenly, that the sag was identically zero because we had used an inadequate approximation!

How do these results affect the measurements of long distances with a tape? The answer is found by calculating the length of tape, t, needed to measure the horizontal distance, l, as shown in Figure 4.2. An element of arc length along the tape, ds, is given by

$$ds = \sqrt{(dx)^2 + (dy)^2} = dx\sqrt{1 + (y'(x))^2}. \tag{4.14}$$

If we substitute the catenary shape (4.8) into eq. (4.14) and apply a standard identity, we find that

$$ds = \cosh\frac{x}{c}\,dx. \tag{4.15}$$

Equation (4.15) can be straightforwardly integrated, that is,

$$\int_0^{t/2} ds = \int_0^{l/2} \cosh \frac{x}{c}\, dx,$$

to yield (see Section 4.8):

$$t = 2c \sinh \frac{l}{2c}. \tag{4.16}$$

We can expand eq. (4.16) in a Taylor formula, again based on the assumption that $l/2c$ is quite small, but for reasons that will soon become evident, we will retain the first two terms in the series, that is:

$$t \cong 2c \left(\frac{l}{2c} + \frac{1}{3!} \frac{l^3}{8c^3} \right). \tag{4.17}$$

With the aid of either eq. (4.11) or eq. (4.12), eq. (4.17) can be written as a quadratic equation in the distance l:

$$l^2 - lt + \frac{8}{3} h^2 = 0. \tag{4.18}$$

The quadratic equation (4.18) can be solved for its roots:

$$2l = t \left(1 \pm \sqrt{1 - \frac{32}{3} \left(\frac{h}{t} \right)^2} \right). \tag{4.19}$$

Only the positive root is physically viable here. In the next section, we will see that the radicand in eq. (4.19) is an ideal candidate to be written as a *binomial expansion*, which is a special form of Taylor's formula. For small values of h/l and to two term accuracy,

$$2l \cong t \left(1 + \left(1 - \frac{32}{6} \left(\frac{h}{t} \right)^2 \right) \right) = t \left(2 - \frac{32}{6} \left(\frac{h}{t} \right)^2 \right). \tag{4.20}$$

Thus, the actual length, l, that is measured by a tape reading of t is given by

$$l \cong t \left(1 - \frac{8}{3} \left(\frac{h}{t} \right)^2 \right). \tag{4.21}$$

Obviously, the larger the sag, h, the larger the correction that must be applied to the tape reading, t, to ensure an accurate measurement of the distance, l.

Lastly on the expansion (4.10), we point out that it is an approximation in the spirit of the *small angle* approximation that appears frequently in

engineering and scientific models. For example, from eq. (4.6b) we know that the second-order Taylor formula for the elementary cosine can be written as

$$\cos x \cong 1 - \frac{x^2}{2!}, \tag{4.22}$$

where x is measured in radians. To approximate the cosine function for very small angles in the neighborhood of $x = 0$, we can safely ignore the second-order term in eq. (4.22) and take $\cos x \cong 1$. However, as we will see in the formal development of the pendulum model in Chapter 7, we often have reason to approximate a slightly different function, $(1 - \cos x)$. If we neglected or ignored the second-order term here, the resulting approximation would be $(1 - \cos x) \cong 0$, which is a bad approximation that results from throwing out the dependence on x. Thus, as in so many other aspects of modeling, it is important to know where we're going when truncating Taylor formulas or series.

There is another approach to approximating trigonometric functions that is worth mentioning. Suppose we wanted to replace $\sin x$ by x in a model or a calculation. We could look at the numerical values of both functions to see where the substitution would be acceptable. For example, if we are willing to accept an error of 5%, we could replace $\sin x$ by x for $x \leq \pi/6$. For an error of only 2%, the substitution would be acceptable for $x \leq \pi/12$. (And while it is important that all angles in these arguments be either rendered as dimensionless ratios of variables or expressed as angles measured in radians, it is worth noting that the two examples just given correspond to small angles of, respectively, 30° and 15°.) Thus, by exploring the numerical ranges of interest and the associated errors, we can often justify replacing a trigonometric function by an algebraic approximation.

4.1.3 Binomial Expansions

Another Taylor series that is used often in engineering and science is the *binomial expansion*:

$$(a + x)^n = a^n + na^{n-1}x + \frac{n(n-1)}{2!}a^{n-2}x^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}x^3 + \cdots . \tag{4.23}$$

Equation (4.23) is valid for all values of n, and it converges for $x^2 < a^2$. Further, when n is a positive integer, the series (4.23) has only a finite number of terms.

Equation (4.23) is very useful in applications when x is rendered dimensionless with respect to a. (Recall that the principle of dimensional

homogeneity requires that x and a have the same physical dimensions.) If we divide eq. (4.23) by a^n, we find that

$$\left(1+\frac{x}{a}\right)^n = 1+n\left(\frac{x}{a}\right)+\frac{n(n-1)}{2!}\left(\frac{x}{a}\right)^2+\frac{n(n-1)(n-2)}{3!}\left(\frac{x}{a}\right)^3+\cdots . \tag{4.24}$$

This is an ideal form for extracting expansions valid for values of $(x/a) \ll 1$.

We will illustrate the use of binomial expansions by looking at a familiar mechanics problem, the estimation of the weight of a mass, m, that is held at some height, h, above the surface of the earth. The weight, W, is the gravitational force, F_g, as expressed by *Newton's law of gravitational attraction*, which can be expressed in scalar form as:

$$F_g = -\frac{Gm_e m}{R^2} = -W, \tag{4.25}$$

where G is the universal gravitational constant, m_e the mass of the earth, and R is the distance between the centers of m and m_e. The minus sign in front of W follows because of the sign convention implied in eq. (4.25) wherein the gravitational force, F_g, would be positive directed away from the earth, while we would customarily draw W as a positive quantity (an arrow) directed toward the earth. Now, if we measure the distance to the mass, m, from the earth's surface as z, it follows that

$$R = R_e + z, \tag{4.26}$$

where R_e is the average radius of the earth. If we substitute eq. (4.26) into eq. (4.25), we find that the weight can now be written as:

$$W = \frac{Gm_e m}{(R_e+z)^2} = \frac{Gm_e m}{R_e^2}\left(1+\frac{z}{R_e}\right)^{-2}. \tag{4.27}$$

The collection of terms involving the earth's properties and the universal gravitational constant are normally expressed in the gravitational constant, g:

$$g \equiv \frac{Gm_e}{R_e^2}, \tag{4.28}$$

so that the weight at height z above the earth's surface is expressed in the form

$$W = mg\left(1+\frac{z}{R_e}\right)^{-2}. \tag{4.29}$$

Equation (4.29) looks strange at first glance. We are accustomed to $W = mg$, so the presence of the dependence on z is unfamiliar. On the other hand, the function of z looks very much like the binomial expansion (4.24). We can assume that $z \ll R_e$, but what does that mean? If we ignore

the dependence on z altogether, then we obtain a very familiar result, that is, $W \cong mg$. If we expand eq. (4.29) in the manner of eq. (4.24) and keep only the first two terms in that expansion, we find that

$$W \cong mg\left(1 - \frac{2z}{R_e}\right). \tag{4.30}$$

This clearly indicates a dependence of weight on height that we do not ordinarily experience. On the other hand, it at least raises the questions, "When does the dependence on height become a significant factor on weight?" and "When does a mass become truly weightless?". The first question can be answered by some straightforward calculations (see Problems 4.9 and 4.10), while the second deserves a bit of discussion. For a body to be weightless, the truncated binomial expansion (4.30) suggests that it would have to be weighed at an altitude $z = R_e/2$. This altitude is sufficiently large that it violates the assumption made in this binomial expansion, that is, $z \ll R_e$. If we look at the exact result (4.29), we see that the body only becomes truly weightless when $z \to \infty$, which is a very different result!

In fact, when the altitude or distance becomes so large that $z \gg R_e$, we would rewrite eq. (4.29) in the form

$$W = mg\left(\frac{R_e}{z}\right)^2\left(1 + \frac{R_e}{z}\right)^{-2}. \tag{4.31}$$

Equation (4.31) can be expanded and truncated as:

$$W \cong mg\left(\frac{R_e}{z}\right)^2\left(1 - \frac{2R_e}{z}\right) \cong mg\left(\frac{R_e}{z}\right)^2. \tag{4.32}$$

The expansion (4.32) clearly indicates that, within a strictly Newtonian world, a body becomes truly weightless only at heights or distances from the earth's surface that are infinitely larger than the radius of the earth. No doubt there are distances for which the weight is significantly less and for which there are practical applications. But, for our purposes, the main point is that the same function can be expanded into different binomial expansions, depending on what information we are seeking. Also, in either instance, we are defining large and small as always, with respect to another dimension or distance. That is, we never say, "z is small" or "z is large." Instead we say that $z \ll R_e$ or that $z \gg R_e$, or, in words, "z is small compared to R_e" or "z is large compared to R_e."

Problem 4.1. Show that eq. (4.7) can be obtained by substituting ix for x in eq. (4.6).

Problem 4.2. Determine the first four terms of the Taylor expansions of $\tan x$ and $\cot x$ about $x = 0$.

Problem 4.3. Determine the first four terms of the Taylor expansions of $\tanh x$ and $\coth x$ about $x = 0$.

Problem 4.4. Use dimensional analysis to determine how the catenary parameter, c, is related to the constant horizontal component of the cable tension, T_0, and its weight per unit length (or unit weight), γ.

Problem 4.5. How much tape sag is permissible to measure a 50 m distance accurately to within 5% ? Within 2% ?

Problem 4.6. What does a body that weighs 10 N at the earth's surface weigh at a height of 10 m? At the peak of Mt. Everest? (*Hint*: You might have to look up some facts about our planet!)

Problem 4.7. According to eq. (4.30), at what altitude would the weight of 10 N at the earth's surface drop to 9 N? To 5 N?

Problem 4.8. Compare the results obtained in Problem 4.7 with more exact results obtained by using eq. (4.29).

Problem 4.9. What does a body that weighs 10 N at the earth's surface weigh on the surface of the moon? On the surface of the planet Pluto? On the surface of the planet Mars? (*Hint*: You might have to look up some facts about our planet's environment!)

Problem 4.10. If the gravitational potential corresponding to Newton's law of gravitation (eq. (4.25)) is given by

$$V_g = -\frac{Gm_e m}{R},$$

find the exact expression that defines this potential as a function of altitude, z, from the earth's surface.

Problem 4.11. Write a binomial expansion of the results of Problem 4.10 to determine the potential energy above the earth's surface to the first order in z.

Problem 4.12. Fill in the missing elements of the following table to two-term order.

Function	Approximation
$\sin x$	
$\cos x$	
$1 - \sin x$	
$1 - \cos x$	

Problem 4.13. Fill in the missing elements of the following table to two-term order.

Function	Approximation
$\sinh x$	
$\cosh x$	
$1 - \sinh x$	
$1 - \cosh x$	

4.2 Algebraic Approximations

As we have seen in Section 4.1, we often drop terms that are of higher order in Taylor series expansions because they will not affect the final answer very much, that is, neglecting those terms does not introduce unacceptable error. We will now look very briefly at some elementary equations of thermal expansion so we can illustrate how we might more generally drop analytical terms to simplify calculations.

When we heat a solid body, the average distance between that solid's atoms increases. Consequently, the linear dimensions of that body—that is, its length, width, or its height—also increase. Thus, assuming that any of the solid's three dimensions is originally of length, L_0, upon heating that produces a temperature difference, ΔT, that dimension increases to the length $L_0 + \Delta L$, where the change in length, ΔL, is given by:

$$\Delta L = \alpha L_0 \Delta T. \tag{4.33}$$

Equation (4.33) tells us that the change in length of a linear dimension is directly proportional to the temperature increase, and that the constant of proportionality is the coefficient of thermal expansion, α. We can rewrite eq. (4.33) as an expression for the heated length of the dimension, L:

$$L = L_0(1 + \alpha \Delta T). \tag{4.34}$$

Suppose the solid we are considering is a sheet of material originally of length L_0 and width W_0. After heating, these two dimensions would each expand according to eq. (4.34) and the plate's original area $A_0 = L_0 W_0$ would expand to the area A:

$$\begin{aligned} A &= L_0(1 + \alpha \Delta T) W_0(1 + \alpha \Delta T) = A_0(1 + \alpha \Delta T)^2 \\ &= A_0 \left[1 + 2\alpha \Delta T + (\alpha \Delta T)^2\right]. \end{aligned} \tag{4.35}$$

Table 4.1 Coefficients of thermal expansion, α, for several common materials.

Material	$\alpha[(°C)^{-1}$, per °C]
Aluminum	24×10^{-6}
Brass	20×10^{-6}
Copper	14×10^{-6}
Glass	$4 - 9 \times 10^{-6}$
Steel	12×10^{-6}
Zinc	26×10^{-6}

The question then arises: Do we need to keep (and use) all three terms in eq. (4.35) to calculate the area change due to heating or cooling?

The answer to the foregoing question depends in part on the coefficient of thermal expansion, α, which is typically a very small number, as can be seen in Table 4.1. Thus, it is tempting to say that because α is small we can neglect the quadratic term in eq. (4.35). And while this may, in fact, be practically alright, in principle it would be wrong, for two reasons. First, if the temperature difference ΔT is large enough, the product $\alpha \Delta T$ might not be negligible. Second, we have cautioned that comparisons should always be made to some reference, so we normally say that it is some dimensional ratio that is small, as we did for $l/2c$ for the catenary. This means that we are making a straightforward numerical estimate. For the present case, the comparable—and proper—statement is that the product $\alpha \Delta T$ is small, so that we can approximate eq. (4.35) as:

$$A \cong A_0(1 + 2\alpha \Delta T). \tag{4.36}$$

From this truncation we can define a surface coefficient of expansion,

$$\gamma \cong 2\alpha, \tag{4.37}$$

where γ is thus derived from our approximating $(1 + \alpha \Delta T)^2$ by $(1 + 2\alpha \Delta T)$.

Problem 4.14. Develop a *volume* coefficient of expansion, β, for a solid of length L_0, width W_0, and height H_0, that parallels the surface coefficient, γ, of eq. (4.37).

Problem 4.15. To what temperature difference would an aluminum solid have to be subjected for the surface coefficient of

expansion to produce errors of 1% in the area change compared to the exact area change?

Problem 4.16. To what temperature difference would an aluminum solid have to be subjected for the volume coefficient of expansion to produce errors of 1% in the area change compared to the exact area change?

4.3 Numerical Approximations: Significant Figures

We now shift our attention to approximations that we make both in measurements and in calculations, that is, we turn to the subject of significant figures. All measurements and virtually all calculations involve approximations or truncations and, therefore, they involve error. When measuring things we try to minimize these errors by being very careful about what we read and record. Although analog displays have been almost completely displaced by digital displays, it is worth revisiting the "good old days" to emphasize an important point about what we regard as significant.

In Figure 4.3 we show an old-fashioned analog display with a graduated scale that goes from 0 to 50 V. The needle points to a number between 12 and 14, and since there are no lines or gradations between 12 and 14, we have to estimate where the needle points within that 2 V interval. Since the needle appears to be about 20% of the distance between these numbers, we estimate that the added voltage measured is $0.20 \times (14 - 12) \cong 0.40$ V, so that the correct reading is 12.4 V. We would characterize this reading as "good to *three significant figures*" because two digits are read directly from the graduated scale, and the third digit is estimated.

It is important to recognize that the number of significant figures is *not* determined by the placement of the decimal point. Had the voltage scale been from 0 to 5 V on the meter in Figure 4.3, we would have recorded a voltage of 1.24 V good to the same three significant figures because we would have directly read 1.2 V plus 20% of the distance between 1.2 and 1.4 V.

We show some examples of how numbers are written in Table 4.2, together with assessments of the number of significant figures of each. The confusion arises because of the presence of terminal zeros. In general, we don't know whether those zeroes are intended to signify something, or whether they are placeholders to fill out some arbitrary number of digits.

It is equally important to recognize that a very similar situation is confronted when doing calculations. Much of the data that engineers and

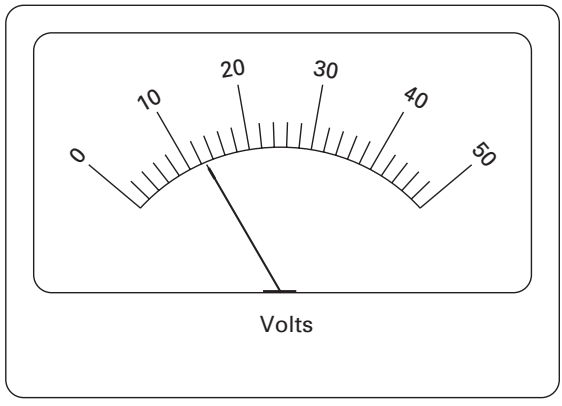


Figure 4.3 An old fashioned analog meter, the standard face before the advent of the digital display. We still see them in many automobile instrument panels, and there are some people who still wear analog watches—but these are uses in which accuracy beyond the gradations is seldom critical. When measuring in the lab and interpreting the results, however, it becomes quite important to know just how many significant figures should be recorded.

Table 4.2 Examples of the ways numbers are typically written and assessments of the number of significant figures that can be assumed or inferred. Confusion arises because of the unstated meaning of the terminal zeroes.

Measurement	Assessment	Significant Figures
9415	Clear	Four
9400	Possibly Confusing	Two (94×10^2) or three (940×10^1) or four (9400)
52.0	Clear	Three
63.2	Clear	Three
6.32	Clear	Three
0.00632	Clear	Three
6.32×10^5	Clear	Three
0.041	Clear	Two
0.0410	Possibly Confusing	Two (0.41) or three (0.0410)
0.00008	Clear	One

scientists use is given only to a limited number of significant figures, sometimes as few as one. For example, there is a much-used material parameter called the *modulus of elasticity*. Denoted by E, this modulus is 30×10^6 lbf/in^2 in British units, implying that at most only two figures (i.e., 30) are significant. It is possible to infer that there is only one significant figure here, but in that case we should write $E = 3 \times 10^7$ lbf/in^2.

Much of the confusion could be mitigated or even eliminated if all technical calculations and experimental data were written in *scientific notation*, wherein numbers are written as products of another number and a power of 10, and where the "new" number is normally between 1 and 10. Thus, numbers both large and small can be written in one of two equivalent, yet unambiguous forms:

$$256{,}000{,}000 = 2.56 \times 10^8 = 0.256 \times 10^9,$$
$$0.000075 = 7.5 \times 10^{-5} = 0.75 \times 10^{-4}.$$

In scientific notation, the number of *significant figures* is equal to the number of digits counted starting from the first nonzero digit on the *left* to either (a) the last nonzero digit on the right if there is no decimal point, or (b) the last digit (zero or nonzero) on the right when there is a decimal point. This notation or convention assumes that terminal zeroes without decimal points to the right signify only the magnitude or power of ten.

We should always remember that *we cannot generate more significant digits or numbers than the smallest number of significant digits in any of our starting data.* In other words, the results of any calculation or measurement are only as accurate as the least accurate starting value. We illustrate this with three examples of multiplication and division showing that the number of significant figures in the result is equal to the smallest number of significant figures in any of the calculation's components:

$$21.982 \times 3.72 = 81.77304 \rightarrow 81.8,$$
$$101.572 \times 0.0031 = 0.3147337 \rightarrow 0.31,$$
$$789.30 \div 0.05 = 15{,}786 \rightarrow 2 \times 10^4.$$

It is far too easy to become captivated by all of the digits that pop up in the displays of our electronic calculators or in computer printouts, but it is really important to remember that *any calculation is only as accurate as the least accurate value we started with.*

In addition and subtraction, the same principle applies. Thus, here we compare the positions of the last significant figure of each number relative to its decimal point because the one that is furthest to the left defines the position of the last allowable significant figure of the sum or difference

being calculated. For example,

$$\begin{array}{r} 53.24 \\ +3.333 \\ +2.4 \\ \hline 58.9 \end{array} \qquad \begin{array}{r} 489.3213 \\ -5.487 \\ \hline 483.834 \end{array}$$

Another important issue when dealing with numbers is that of *round off*, that is, when should we round off numbers, or should we round them off at all? We generally round off numbers at the end of a calculation because dropping insignificant numbers earlier increases uncertainty. The standard convention for rounding off uses the number 5 as its benchmark: Numbers less than 5 following the last retained significant digits are dropped, while numbers greater than 5 cause us to add 1 to the last significant digit retained. If the digit to be rounded or dropped is itself a five, we make the preceding digit even (i.e., even digits are left so, while odd digits are "rounded up" to the next even digit). Thus, for example,

$$5.017 \rightarrow 5.02,$$
$$5.015 \rightarrow 5.02,$$
$$5.014 \rightarrow 5.01,$$
$$5.025 \rightarrow 5.02.$$

These results also indicate the degree of uncertainty in the true value of a number that has been rounded off. From the data just given and the rules behind it, we see that the number 5.02 could mean a number that is actually between 5.015 and 5.025.

Finally, it is worth noting that there are numbers that have unlimited significant figures. Some are whole numbers representing an exact count, and thus contain an unlimited number of significant figures. They are usually written without any digits after the decimal point, or they may not have a decimal point at all. To indicate such a number, we might write "35." or, as in the formula for the circumference of a circle, $C = 2\pi r$, wherein the number "2" represents an exact count and is written without a decimal point. The number π is itself a number that has an infinite number of significant figures, as does e, the base of Naperian logarithms. However, we write "35.0" or "2.0" when we want to indicate that we are measuring something to the first decimal place.

Whether reporting measurement data or doing calculations, we should always keep in mind the significance of our initial data so that we can assess the validity of our results.

Problem 4.17. Round off each of the following numbers to two (2) significant figures:
(a) 5.237 (b) 0.82549 (c) 81.356 (d) π
(e) 6.2305 (f) 0.0428 (g) 10.45 (h) 4.035

Problem 4.18. Round off each of the following numbers to three (3) significant figures:
(a) 5.237 (b) 0.82549 (c) 81.356 (d) π
(e) 6.2305 (f) 0.0428 (g) 10.45 (h) 4.035

Problem 4.19. Complete the following multiplications and express the results to the correct number of significant figures:
(a) $(6.28 \times 10^3) \times 2.712$ (b) 43.32×0.3
(c) 928×4.23

Problem 4.20. Do 99.9 and 100.1 have the same number of significant figures? Explain your answer.

Problem 4.21. Estimate the ranges within which each of the following numbers lie:
(a) 7.7 (b) 7.70 (c) 1200 (d) 1.200×10^{-3}

4.4 Validating the Model–I: How Do We Know the Model Is OK?

There are two issues that arise when we speak of the validity or correctness of a model. The more obvious one is whether or not the model can predict the measured or observed behavior of whatever object or device is being modeled. Thus, if we are modeling the period of the oscillations of a pendulum, as we started to do in Chapter 2, we could reasonably expect that changes in the pendulum length would produce oscillations at correspondingly different periods or frequencies. As we see from eq. (2.2), if we double the length l of a pendulum, we would expect its period to increase by about 41% . Similarly, were we doing pendulum experiments on the moon, we would expect to see an increase in the period of about 145%. These predictions of the pendulum's behavior are confirmed by the available experimental data, and so the model is validated. Alternatively, given empirical data without an underlying theory, we could construct a model to explain the empirical data—although it is also quite likely that the (new) model or theory would be further tested by making predictions about experiments as yet undone or measurements as yet untaken.

(We note parenthetically that the measurement [and containment] of experimental error is a complex subject that is closely linked to the field or discipline in which the experiment is intellectually housed. However, there

are some fundamental ideas about error and about statistics that apply generally, and we will introduce them in Sections 4.5–4.8.)

The less obvious question about model validity is concerned with the inherent consistency and validity of the model. If we hark back to the modeling meta-principles outlined in Section 1.2, we see issues and questions that pertain directly to model validation. For example, have we identified the right governing principles? Have we used the right equations? And, is the model consistent with its principles and assumptions? The first two of these questions are about ensuring that we apply the proper principles and formulations when we try to find what we are seeking. Again, when modeling the pendulum, our basic principles are Newton's law of motion, and our assumptions will depend on whether we are anticipating small angles of oscillation or large. As we will see in Chapter 7, a linear equation of motion suffices in the former case, while a complete nonlinear formulation is needed for the latter (large oscillations).

4.4.1 Checking Dimensions and Units

There are several checks or tests we can bring into play while we build models and approximate the mathematics. The first is the application of the principle of dimensional homogeneity (cf. Section 2.2), which requires that each term in an equation has the same net dimensions. For example, the *stiffness* or spring constant of a cantilever beam, k, can be written in terms of the beam's length, L, second moment of its cross-sectional area (commonly but erroneously called the "moment of inertia"), I, and modulus, E, as:

$$k = \frac{3EI}{L^3}. \tag{4.38}$$

The physical dimensions of the parameters in eq. (4.38) are $\mathsf{F/L}$ for the spring constant, L for the beam length, L^4 for I, and $\mathsf{F/L}^2$ for the modulus. Thus, we can apply the principle of dimensional homogeneity to ensure that eq. (4.38) has the correct dimensions and is dimensionally consistent:

$$[k] = (\mathsf{F/L}) = \left[\frac{3EI}{L^3}\right] = \frac{1 \times (\mathsf{F/L}^2) \times \mathsf{L}^4}{\mathsf{L}^3} = (\mathsf{F/L}). \tag{4.39}$$

If the dimensions of all the terms in an equation or model are not known, as is sometimes the case, then the principle of dimensional homogeneity can be applied to properly determine the dimensions of the unknown quantity. In the case of the cantilever beam, if we didn't know the dimensions of I, we would solve eq. (4.38) for I and then apply the principle of

dimensional homogeneity again:

$$[I] = \left[\frac{kL^3}{3E}\right] = \frac{(\mathsf{F}/\mathsf{L})\mathsf{L}^3}{\mathsf{F}/\mathsf{L}^2} = \mathsf{L}^4. \tag{4.40}$$

We can also take the principle of dimensional homogeneity one step further and use it as a guiding principle for checking the specific *units* used in a numerical calculation. If we measured the properties of a particular cantilever beam, say a standard (12 in) steel ruler to be used in a classroom project, we would find

$$\begin{aligned} E &= 2.05 \times 10^2\,\text{GPa}, \\ I &= 6.78 \times 10^{-5}\,\text{cm}^4, \\ L &= 2.81 \times 10^{-1}\,\text{m}. \end{aligned} \tag{4.41}$$

If we substitute these values into eq. (4.38), we see immediately that we have a mismatch of units:

$$k = \frac{3(2.05 \times 10^2\,\text{GPa})(6.78 \times 10^{-5}\,\text{cm}^4)}{(2.81 \times 10^{-1}\,\text{m})^3}. \tag{4.42}$$

The units' mismatch is easily rectified if we use proper unit conversions, that is,

$$k = \frac{3\left[2.05 \times 10^2 \times 10^9\ \text{Pa}\left(\frac{\text{N/m}^2}{\text{Pa}}\right)\right]\left[6.78 \times 10^{-5}\ \text{cm}^4\left(\frac{\text{m}}{10^2\text{cm}}\right)^4\right]}{(2.81 \times 10^{-1}\ \text{m})^3}, \tag{4.43}$$

or

$$k = \frac{3\left[2.05 \times 10^{11}\ \text{N/m}^2\right]\left[6.78 \times 10^{-13}\ \text{m}^4\right]}{(2.81 \times 10^{-1}\ \text{m})^3} = 1.88 \times 10^1\ \text{N/m}. \tag{4.44}$$

Two final notes here. First, it is generally a better strategy to write all of the data to be used in the same system of units at the beginning of a calculation as this reduces the chance for error. Thus, here we could have converted the units immediately after the measurements were taken. Second, note that we have used scientific notation in both writing the measurements and performing the arithmetic. Thus, there can be no doubt about the number of significant figures in the answer (4.44).

4.4.2 Checking Qualitative and Limit Behavior

Model validation is integral to the modeling process. Models are validated by having their predictions confirmed experimentally, or statistically, or by some other quantitative means. In both our physical and mathematical reasoning we must make explicit our assumptions and their limits, and we must ensure that our mathematics does indeed reflect the physics we are modeling. In addition to looking at numbers, the mathematical behavior should "feel right" in qualitative terms. We did just such qualitative analysis at the beginning of this section when we described the expected behavior of the pendulum as a function of its length, l. Similarly, as also indicated by eq. (2.2), it feels intuitively right that pendulums will swing faster and have shorter periods in stronger gravitational fields. Thus, when we are constructing mathematical models, and especially when we are making mathematical approximations, we need to take care that we are admitting mathematical behaviors that are qualitatively appropriate.

Still another example of such reasoning is available from our just-completed dimensional check of the stiffness of a beam. Here we rewrite eq. (4.38) in a form that explicitly identifies the physical meaning of each parameter that appears in the equation:

$$(k = \text{beam stiffness}) \propto \frac{(E = \text{material stiffness})(I = \text{cross} - \text{sectional 2nd moment})}{(L = \text{beam length})^3}. \tag{4.45}$$

Equation (4.45) can be viewed through the eyes of a structural engineer talking about the meaning of its mathematical version, eq. (4.38). It supports the engineer's intuitions as follows. It stands to reason that the beam's stiffness is proportional to the material stiffness, that is, it increases or decreases as does E. The beam's stiffness is also proportional to the second moment of the beam's cross-section, I. It also is intuitively pleasing that the stiffness is inversely dependent on the length, so that the beam's stiffness increases as L becomes very small and decreases as L becomes very large. Finally, if we look at the limiting cases of each parameter decreasing to zero or becoming indefinitely large, we would see that each of the trends exhibited by eq. (4.45) is consistent with the reasoning just outlined, as well as with our practical experience of beams in the real world.

Reasoning about the way that variables appear in equations is of second nature in mathematical modeling, and we will have many opportunities to invoke such reasoning in the discussions of applications that follow. One simple example is afforded by the fundamental frequency of free vibration of a cantilever beam, ω, of mass density, ρ, and cross-sectional area, A,

with a mass, m, at its tip. That frequency is, approximately,

$$\omega \cong \sqrt{\frac{3EI/L^3}{\rho AL(1 + m/\rho AL)}}. \tag{4.46}$$

Does eq. (4.46) exhibit the right qualitative and limit behavior? It does. It reduces to a well-known result for a cantilever beam when the tip mass, m, vanishes, and eq. (4.46) correctly describes the frequency of a mass-less beam with a tip at its end when that tip mass gets so large that it dominates the beam mass.

It may seem that much of what has been said in this section is *common sense*. It is, as long as it is commonly applied! To invert a popular saying, "If we expect our model to be a duck, then it should look like a duck, walk like a duck, and quack like a duck."

Problem 4.22. By what percentage would the period of a pendulum change if its length was halved? If it was reduced by one-third? If the length was reduced to one-third of its original length?

Problem 4.23. Explain why the pendulum period increases by 145% on the moon.

Problem 4.24. How would the period of a pendulum change, compared to its value on earth, if the pendulum was on Mars? On Pluto?

Problem 4.25. How would the period of a pendulum change as a function of its height, h, above the surface of the earth? (*Hint*: The variation of the gravitational acceleration g can be represented as a function of h from Newton's law of gravitational attraction.)

4.5 Validating the Model–II: How Large Are the Errors?

Building mathematical models means using numbers derived from experimental or empirical data, or from analytical or computer-based calculations. Errors are thus always present, whether due to data reading or data manipulation. Since error is always present, we turn now to a discussion of error and statistics—the way we deal with error.

4.5.1 Error

Error is defined as the difference between a measured (or calculated) value and its true or exact value. Error is *always* present. How much error is present depends on how skillfully the data is read or manipulated. Therefore, error analysis should be a part of every modeling process.

There are two types of error. *Systematic error* occurs whenever an observed or calculated value deviates from the true value in a consistent way. Systematic error occurs in experiments when instruments are improperly calibrated because their output varies during use. Thus, instruments must be properly calibrated before an experiment is run and before data is measured and recorded. Improper calibration affects both analog and digital data recorders, although analog displays are also subject to other kinds of systematic error, such as a bent needle on a meter face such as that shown in Figure 4.3. Systematic error also affects calculations, although this is more controllable as it is likely due to using incorrect values of "known" variables or to improper control of the number of significant figures retained during the calculation process.

Random errors are, not surprisingly, due to chance. They arise largely in experimental work because unpredictable things happen and because not everything in an experimental set-up is known with complete certainty: Connections can be loose or break altogether, dirt may get into a sensitive moving part, or the amount of friction present in a moving part may not be controllable. The resulting random error varies in both magnitude and sign. The laws of statistics help us to describe and account for the distribution of such random errors. Indeed, it has been said that randomness is a mathematical model for variability that cannot be explained in a deterministic way.

The *absolute error* is defined as the difference between the true or expected value, X_e, and the measured value, X_m, that is, as $X_e - X_m$. The true value, X_e, may be known or it may have an expected value based on a calculation or some other data source. The *relative error* is the absolute error divided by the measured value, that is, $(X_e - X_m)/X_m$.

The statistic found most useful is the *percentage error*, which is the percentage-based relative error:

$$\% \text{ error} = (100)\frac{(X_e - X_m)}{X_m}. \tag{4.47}$$

For example, suppose that an ammeter has a systematic error of +2 A (amperes) because of either a bent needle (analog) or improper calibration

(digital or analog). When the display reads 100 A the percentage error is

$$\% \text{ error} = (100)\frac{(102 - 100)}{100} = 2\%,$$

while if the same ammeter reads 20 A the percentage error is

$$\% \text{ error} = (100)\frac{(22 - 20)}{20} = 10\%.$$

The percentage error is much larger in this instance, providing another example of how scale affects results!

Similarly, errors are introduced when series expansions are truncated (cf. Section 4.1.2). For example, for $\theta = \pi/12(15^\circ)$, the percentage error incurred by replacing $\sin x$ with x is:

$$\% \text{ error} = (100)\frac{(\sin \pi/12 - \pi/12)}{\pi/12} = -1.14\%.$$

Note that *errors* and *mistakes* are not the same thing. Errors are defined as the difference between a true or expected value and a measured (or calculated) value. Further, as we discussed above, some error is unavoidable. On the other hand, *mistakes* are blunders made by the person doing the experiment (or analysis or calculation). Blunders are made by reading or recording erroneous data, using instruments inappropriately (e.g., improperly calibrated instruments, inadequately sensitive meters), using the wrong formulas, using inconsistent or wrong units, and so on. These kinds of mistakes can—and obviously should—be avoided.

4.5.2 Accuracy and Precision

Since we have to contend with systematic and random errors, as well as with the hopefully rare mistake, it is important that we be able to estimate the effects of these errors and mistakes.

Accuracy is defined as a representation of how close a measured or calculated value is to an established or true value. In experimental work, accuracy is usually expressed as a percentage of the maximum scale value. Thus, voltages read on a 100 V scale with an accuracy of 5% are accurate to within ± 5 V.

Precision is defined in terms of the ability to reproduce a set of data with a specified accuracy. The more precise a set of readings or calculations, the closer the individual readings or calculations are *to each other*. Thus, suppose we measured an input voltage that is known to be 50 V with the voltmeter having an accuracy of 5%. Five individual readings are taken and recorded as, respectively, 54, 53, 55, 53, and 55 V. These clearly fall

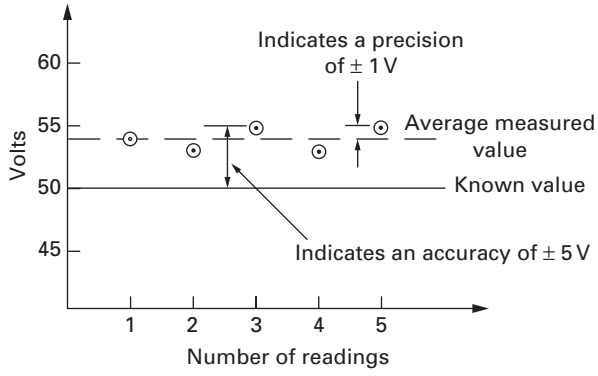


Figure 4.4 Some (made-up) experimental data that illustrates: *accuracy*, the closeness of the measured value to an established value, and *precision*, the ability to reproduce a set of measurements within a specified accuracy. These data reflect measurements that are rather precise, yet relatively inaccurate.

within the meter's accuracy bounds of ± 5 V. Since the average or mean reading of the five readings is 54 V, and since the maximum deviation from this mean of any one of the measurements is 1 V, the precision of the five measurements is determined to be $\pm 1\%$ (remember that the meter has a 100 V scale). As we illustrate in Figure 4.4, our little virtual experiment has produced precise but relatively inaccurate readings.

It is worth noting that the accuracy of a measuring device is controlled by its sensitivity because it is the *sensitivity* that identifies the minimum amount of change that the device can detect and indicate. Suppose we wanted to measure very small voltages, say less than 1 millivolt (mV). Our trusty voltmeter allows us to choose one of three measurement ranges: 0–50 V, 0–2.5 V, or 0–5 mV. With either of the first two ranges we will see no reading at all. However, with the third scale, 0–5 mV, there will be a noticeable measurement that can be recorded. Thus, moving from either of the first two scales to the third produces a more sensitive voltmeter, and so our readings will be more accurate. Hence, we see how scale influences sensitivity and, therefore, accuracy.

Problem 4.26. Draw two circular archery targets and use them to depict the "hit" patterns of (a) an archer who is accurate, but not precise; and (b) an archer who is precise, but not accurate.

4.6 Fitting Curves to Data

Graphical presentations of calculations and experimental results are the most convenient—and often the most informative—presentation of data available. We can spot trends, identify discontinuities, and generally get an intuitive "feel" for what the data "says" when we look at plots or curves. Given this very human proclivity, how do we draw curves for a given collection of points? That is, since plotted data points rarely align themselves perfectly on a known or identifiable curve, how do we fit a curve through them? Still further, how do we generate the "best fit" of a curve through the data?

The short answer to these questions is in a familiar spirit: It depends on what you want. If the accuracy of the curve is not too important, and if we're only looking for a rough, qualitative idea of how one variable depends on another, then we can draw the curve "by eye." That is, we draw a smooth curve that seems to go through the plotted data points with an eye to perhaps "distributing" the data in roughly equal amounts above and below the curve drawn, as we have done in Figure 4.5.

Often, greater accuracy is desirable, as when we want to *interpolate* to obtain values between measured values, or even more so when we want to *extrapolate* to estimate values beyond the range of the measured values. Extrapolation can easily magnify errors in the estimated values, so that greater accuracy is quite important. Further, extrapolation is most accurate when the curve drawn is a straight line.

The *method of least squares* is the most commonly used approach to obtaining a best straight line through a series of points. It assumes that all of the *scatter*, the variation of the data from the drawn curve, derives

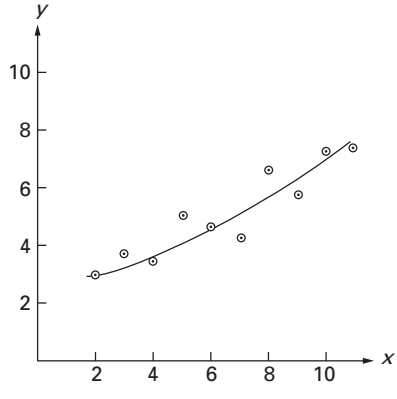


Figure 4.5 A best-fit curve that is drawn by hand using visual estimation (i.e., "drawn by eye").

from error in measuring one of the variables. That variable is chosen as the ordinate for the axes on which the straight line will be plotted. Then the best-fit straight line is the one that has the minimum errors in the ordinate.

We are thus looking for an equation of the usual form

$$y = mx + b, \tag{4.48}$$

where b is the y-intercept with [b] = [y], and m is the slope with [m] = [y/x]. We first define the error in each reading as the difference in the ordinate between the measured value, y_i, and the straight line's ordinate, $(mx_i + b)$, for all values of the abscissa, x_i:

$$E_{y_i} = y_i - (mx_i + b). \tag{4.49}$$

We define a measure S of the total error as the sum of the square of the errors at every point on the abscissa, x_i, where values of the ordinate, y_i, are given, that is, as

$$S = \sum_{i=1}^{n} (E_{y_i})^2 = \sum_{i=1}^{n} [y_i - (mx_i + b)]^2. \tag{4.50}$$

The minimum of the measure of the total error is then found by differentiating S with respect to m and b and so determining the values of m and b needed to plot eq. (4.48):

$$\begin{aligned} \frac{\partial S}{\partial m} &= 2 \sum_{i=1}^{n} [(y_i - mx_i - b)(-x_i)] \\ &= -2 \sum_{i=1}^{n} x_i y_i + 2m \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i = 0, \end{aligned} \tag{4.51}$$

and

$$\begin{aligned} \frac{\partial S}{\partial b} &= 2 \sum_{i=1}^{n} [(y_i - mx_i - b)(-1)] \\ &= -2 \sum_{i=1}^{n} y_i + 2m \sum_{i=1}^{n} x_i + 2nb = 0. \end{aligned} \tag{4.52}$$

Equations (4.51) and (4.52) are a pair of linear algebraic equation that can be solved (see Problem 4.28) to yield the following values of m and b:

$$m = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{4.53}$$

and

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{4.54}$$

Note that eqs. (4.53) and (4.54) have different physical dimensions that depend on the particular physical problem being modeled (see Problem 4.28).

Consider now the data displayed in the first two columns of Table 4.3, which are the result of another, virtual experiment. We will now determine the best straight line that can be drawn through the data. First, we calculate the products shown in the third and fourth columns of Table 4.3. Then we sum all four columns to find the data in the last row of the table, which are then substituted into eqs. (4.53) and (4.54) to find $m = 0.85$ and $b = 1.26$. The best straight-line fit through the data of Table 4.3 is, then,

$$y = 0.85x + 1.26. \tag{4.55}$$

Equation (4.55) is plotted in Figure 4.6, together with the data from Table 4.3, and we see that the straight line seems to fit the data pretty well. Can we characterize the *quality* of that fit, that is, just how well does eq. (4.55) fit the given data? The quality of fit is expressed in terms of R^2, called "R squared," which describes how well a curve *regresses* toward the

Table 4.3 A table of data from a virtual experiment used to calculate the best-fit straight line approximation shown in Figure 4.6.

i	x_i	y_i	$x_i y_i$	x_i^2
1	0	1.0	0	0
2	1.0	2.1	2.1	1.0
3	2.0	2.8	5.6	4.0
4	3.0	3.6	10.8	9.0
5	4.0	5.0	20.0	16.0
6	5.0	5.5	27.5	25.0
7	6.0	8.0	48.0	36.0
8	7.0	6.4	44.8	49.0
9	8.0	7.4	59.2	64.0
	$\sum_{i=1}^{9} x_i = 36.0$	$\sum_{i=1}^{9} y_i = 41.8$	$\sum_{i=1}^{9} x_i y_i = 218.0$	$\sum_{i=1}^{9} x_i^2 = 204.0$

data from which it was derived. R^2 is a number between 0, which indicates no fit at all, and 1, which describes a perfect fit. (There are many mathematical and statistical computational packages that include the formulas needed to calculate R^2.)

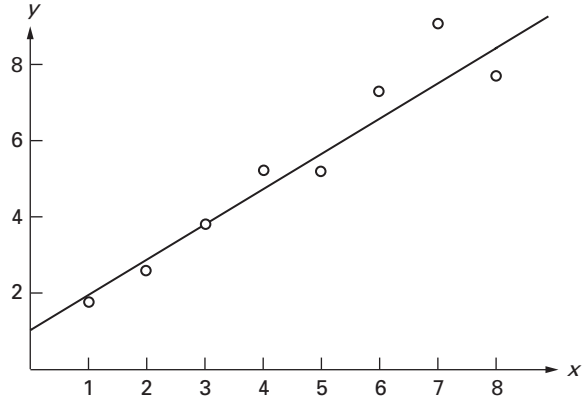


Figure 4.6 A best-fit straight line for the data in Table 4.3 produced by least squares. It is analytically represented as $y = 0.85x + 1.26$.

Problem 4.27. Verify the final forms of eqs. (4.51) and (4.52).

Problem 4.28. Verify the equations for m and b given in eqs. (4.53) and (4.54).

Problem 4.29. Discuss and explain the dimensional differences between eqs. (4.53) and (4.54).

Problem 4.30. Verify the terms in the third and fourth columns of Table 4.3, as well as the sums of all four columns.

Problem 4.31. Verify the calculations of m and b found from the results in Table 4.3.

4.7 Elementary Statistics

What do we do *after* we have recorded a bunch of measurements or calculated several values of something? A more meaningful phrasing of this question would be: How do we organize and present our results so that we are better able to understand and communicate the data? Our answer to this question comes in two parts. In the first, we define the meaning of average, while in the second, we discuss ways of drawing curves through data.

4.7.1 Mean, Median, and Standard Deviation

We often want to average our results when making several measurements or calculating several values of something. There are several ways of defining the meaning of average, but we will limit our discussion to two: the mean and the median.

In Figure 4.4 we showed data from a virtual experiment whose individual measurement readings (and, occasionally, model calculations) vary from one another. We want to deal with a single value, a best estimate of the magnitude of the entire set of readings. We will take the average or mean of a *sample* of n measurements as such a best estimate, where the *arithmetic mean* or *sample mean* $\bar{x}$ is defined as the sum of all of the individual readings x_i divided by the number of readings, n:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{4.56}$$

Note that the calculation of the mean of a set of values given by eq. (4.56) strongly resembles the way that the centroids of areas are calculated, and for good reason!

There is one other measurement that is often cited as a meaningful indicator of an "average" of a number of readings and that is the *median*, which is defined as the measured value that is at the middle of the distribution. The median removes any bias that might be introduced by a few values that differ significantly from the mean. For example, in the virtual voltmeter experiment of Section 4.5.2, the median is 54, which is the same as the mean. On the other hand, had the five readings been 54, 53, 65, 53, and 55 V, then the mean rises to 56 V, while the median stays at 54.

In Table 4.4 we show a collection or sample of 100 noise level measurements of the noise due to traffic as measured in a schoolyard playground. In addition to traffic noise, the microphones also picked up the occasional noise due to children in the playground who, excited by the experiment, made some loud sounds as they passed by. We see that for these measurements the mean is higher than the median, which is likely due to the relatively large number of readings in the 90–91 dB interval.

In addition to identifying the mean as our best estimate, we would like to estimate the spread or dispersion of the set of measurements about the mean. Clearly, if this estimate of the spread is small in some sense, then we can attach a high precision to the mean $\bar{x}$. The accepted statistical measure of this estimate of dispersion is the *sample variance*, s^2, defined in

Table 4.4 A sample of 100 noise level measurements (in decibels (dB)) made in a schoolyard playground.

Decibels		Number of Observations
	90–91	x x x x x x x x
	88–89	x x x x x
	86–87	x x x
	84–85	x x
	82–83	x x x x
	80–81	x x
	78–79	x x x
	76–77	
	74–75	x x x x x x
	72–73	x x x x
Mean	70–71	x x x x x x
	68–69	x x x
Median	66–67	x x x x x x x
	64–65	x x x x x
	62–63	x x x x x x x x x
	60–61	x x x x x x x x x x x x x
	58–59	x x x x x x
	56–57	x x x x x x x x
	54–55	x x x x
	52–53	x
	50–51	

terms of the deviation of each reading from the mean, $(x_i - \bar{x})$, as:

$$\begin{aligned} s^2 &\equiv \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} \\ &= \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2. \end{aligned} \tag{4.57}$$

The *standard deviation*, s, is defined as the square root of the sample variance:

$$s \equiv \left[\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}. \tag{4.58}$$

We often see the symbol σ used for the standard deviation, but that usage is correct only when the calculation is performed for the *total population* or the complete set of all the objects being measured. When we are taking

readings or doing calculations, we are taking a *sample* of all of the values that could, in principle, be obtained. In that case, s is the correct notation for the standard deviation of that sample. (Similarly, when the calculation of a mean is done for an entire population, it is denoted by μ, rather than $\bar{x}$.)

Note that in calculating the standard deviation, the deviation of each value or reading from the mean is squared before being added to the comparable deviations of the rest of the readings. This is done to eliminate the sign differences that occur because the deviation $(x_i - \bar{x})$ can be positive or negative, depending on whether the reading x_i is greater or smaller than the mean $\bar{x}$. Thus, only positive numbers are added when the standard deviation is calculated. Also note that eq. (4.58) clearly suggests that the best way to increase the precision of the answer is to increase the number of readings or calculated values. Indeed, an infinite number of measurements would, in theory, produce perfect precision because the standard deviation vanishes in the limit $n \to \infty$. We also point out that just as the calculation of the mean parallels the calculation of the location of the centroid of an area about one axis, the calculation of the variance (eq. (4.57)) parallels the calculation of the second moment of area about that same axis.

Notwithstanding the physical analogy just given, the interpretation of the standard deviation, s or σ, is difficult because its units are squares of the units of the variable, x. However, we can give meaning to the standard deviation when we relate it to the mean of the data set, $\bar{x}$ or μ. This meaning is embedded in the *Empirical Rule* that tells us, approximately, where the data points lie with respect to the mean. The following heuristics describe the data set that underlies a distribution that is, approximately, a mound shape:

- almost all of the data points lie within 3 standards of deviation of the mean of the data set, that is, within the window $(\bar{x} \pm 3s)$ for samples and within the window $(\mu \pm 3\sigma)$ for complete populations;
- some 95% of the measurements lie within 2 standards of deviation of the mean of the data set, that is, within the window $(\bar{x} \pm 2s)$ for samples and within the window $(\mu \pm 2\sigma)$ for complete populations; and
- some 68% of the measurements lie within 1 standard of deviation of the mean of the data set, that is, within the window $(\bar{x} \pm s)$ for samples and within the window $(\mu \pm \sigma)$ for complete populations.

4.7.2 Histograms

Another way of displaying measured data is the *histogram* or *bar chart* in which a distribution of the frequency of occurrence of the measured quantity is displayed. The histogram's abscissa indicates the values recorded,

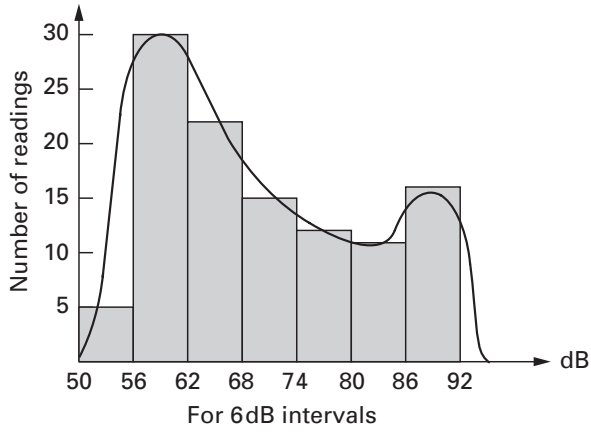


Figure 4.7 A histogram of the noise measurements given in Table 4.4, with a continuous approximation of the same noise level data superposed as a dotted line. Here the data were taken and recorded in 6-dB windows or intervals. For example, there are 30 measurements registered in the 56–62 dB window.

while its ordinate represents the number of times the values occur. The histogram shown in Figure 4.7 displays the same data given in Table 4.4 with the measured sound pressure levels grouped in 6-dB intervals or windows. Thus, the bar between 56 and 62 dB represents the total number of measurements that registered, respectively, 56, 57, 58, 59, 60, or 61 dB. Two questions occur immediately: Why construct histograms? and How big should the intervals be?

The main reason for constructing a histograms is that it offers a graphic depiction of the frequency of events, so that problematic repetitions of particular events are readily identified. Histograms can also be used to generate approximate plots based on the data they express. For example, Figure 4.7 also shows a continuous approximation of its 6-dB histogram. Both the histogram and its continuous counterpart show us that the largest number of readings of outdoor noise in the schoolyard occur in the 56–62 and 86–92 dB windows. This prompts us to inquire about the cause(s) of readings at these two levels. In response, we can identify the peak in the 86–92 dB window as deriving from the children yelling at the microphone, which in turn allows us to note that the playground noise is more generally at levels less than 86 dB, with the remaining peak occurring at the relatively low levels of 56–62 dB.

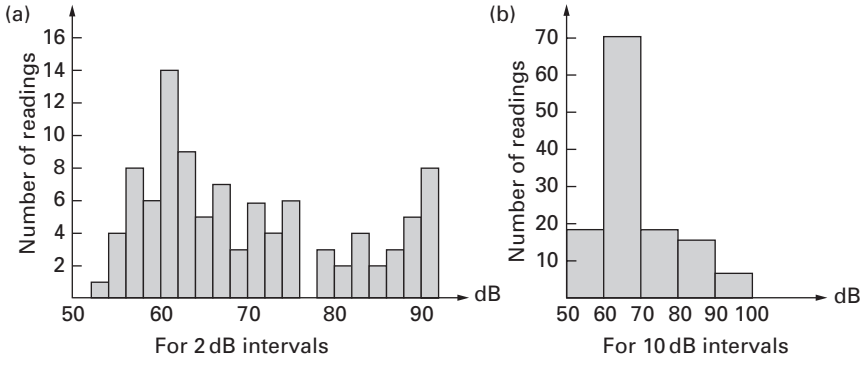


Figure 4.8 Two more histograms of the noise measurements given in Table 4.4. The data were taken and recorded in, respectively, (a) 2-dB and (b) 10-dB windows.

How do we decide on the size of the intervals or windows? We want the interval to be *large enough* to have enough data to minimize the chance of spurious fluctuations, yet *small enough* that we don't throw out data that would indicate serious events within the interval. The data of Table 4.3 (and the 6-dB histogram of Figure 4.7) are displayed in Figure 4.8 in histograms with intervals of (a) 2 dB and (b) 10 dB. We see that with the larger interval we have lost the (identifiable) peak due to the children's screaming, while with the smaller interval we have many more peaks and fluctuations. As a practical matter, experience suggests that the number of bars in a histogram should roughly equal the square root of the number of data entries, $\sqrt{n}$.

How did we draw the curve representing the continuous version of the histogram in Figure 4.7? First, we assumed the validity of the *continuum hypothesis*, which states that such discrete data can be plotted as a continuous curve. Second, we chose the number of intervals to get a relatively smooth and meaningful curve. Just as with the underlying histogram, this meant going back to the original data (i.e., Table 4.4) to choose an interval size large enough to contain a significant number of points, yet not so large that variations within the interval are drowned out. We constructed Table 4.5 to aid in this process of choosing an interval size. Table 4.5 organizes the data in Table 4.4 in terms of the number of points within intervals of length Δ centered around 66 dB: There are 13 readings in the interval of $\Delta = 4$ dB, 27 in the interval of $\Delta = 8$ dB.

A plotted curve of the data of Table 4.5, in Figure 4.9, helps us better visualize and understand the data. If the length of the measuring interval Δ is too small, say < 4 dB, the density fluctuates a lot and is not representative of the complete picture. If the interval Δ is too large, say > 8 dB, the

Table 4.5 An organizing chart of the data in Table 4.4 that allows us to estimate the number of data points in intervals of varying length Δ. This form of the data enables the drawing of the plot shown in Figure 4.9.

Interval length, Δ	1	2	3	4	5	6	7	8	10	20	30	40	50
Interval, $66 \pm \Delta/2$	66.5 65.5	67 65	67.5 64.5	68 64	68.5 63.5	69 63	69.5 62.5	70 62	71 61	76 56	81 51	86 46	91 41
Number of readings in interval	6	9	12	13	17	19	24	27	37	68	78	85	100
Density	6	4.5	4	3.25	3.4	3.17	3.43	3.34	3.7	3.4	2.6	2.38	2.00

density curve is smoothed out to the extent that all of the meaningful variations have disappeared. Thus, an interval such that $4 < \Delta \text{ dB} < 8$ would appropriately approximate the number of readings as a continuous function of the noise level. We have 100 readings here, so $10 = \sqrt{100}$ histogram bars are appropriate for the range 50–90 dB, resulting in the shown width of 4 dB. However, as with other aspects of modeling, the number of histogram bars is to some extent a matter of taste.

We have not offered any criteria to aid in choosing a measuring interval because there are none. The best path is to organize the data as in Table 4.5, use it to plot a curve such as that in Figure 4.9, and then exercise our best judgment as to the size of Δ.

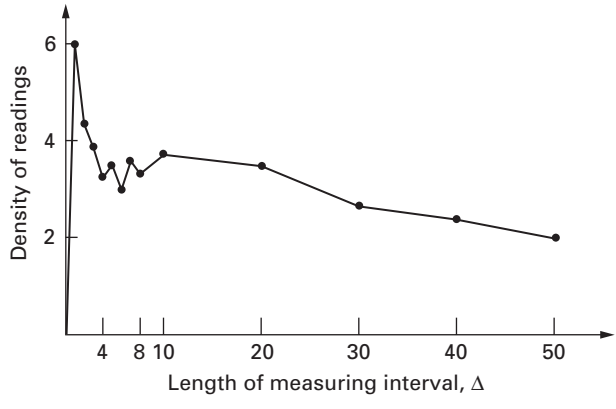


Figure 4.9 An illustration of the *continuum hypothesis,* showing how the density of the readings depends on variation of the measuring interval.

Lastly, all of the calculations outlined in this section can be done on a computer. Care and thought are invaluable because, at best, a computer does only what it's told to do. Given erroneous instructions or bad data, its results will be erroneous and bad!

Problem 4.32. Determine the standard deviation for the data presented in Table 4.4.

Problem 4.33. Draw a histogram for the data in Table 4.4 with 10 intervals of 4 dB width.

Problem 4.34. Show that the square of the sample variance of eq. (4.57) can be cast in the alternative form

$$s^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right].$$

4.8 Summary

We have devoted this chapter to discussions of approximations and their limits, and of model validation, including both qualitative and statistical methods. We have shown the importance of Taylor and algebraic series expansions, including applications to stretched strings (Taylor series of hyperbolic functions), gravitational forces (binomial expansions), and thermal expansion (algebraic approximations). We have emphasized the need to validate models, as well as the roles played by dimensional and qualitative analyses in model validation. We have also stressed the importance of numerical approximations and of significant figures, especially as regards their proper display and interpretation.

Working with mathematical models means that we are constantly using numbers that derive from calculations or experiments. These numbers always incorporate error. We have discussed both random and systematic errors, and how they affect the precision and accuracy of any set of data. We also looked briefly at statistical techniques that could be used to quantify such errors, introducing the concepts of mean, median, and standard deviation. We showed how curve fitting could be used to approximate functions, and we showed illustrative examples using both the least squares method and the continuum hypothesis to develop statistically based numerical approximations.

4.9 Appendix: Elementary Transcendental Functions

The so-called *elementary transcendental functions* are the trigonometric functions ($\sin x, \cos x$), the exponential functions (e^x, a^x), the hyperbolic functions ($\sinh x, \cosh x$), and the logarithmic functions ($\ln x = \log_e x, \log_a x$). We will present some basic results and relationships for these functions, rather than derivations and proofs. Some of the results make use of the notation $i = \sqrt{-1}$, which is central in relating, for example, the trigonometric functions to the exponential. In fact, we will use what famed physicist Richard Feynman called "the most remarkable formula in mathematics":

$$e^{ix} = \cos x + i \sin x. \tag{4A.1}$$

We also note that that the *imaginary* (as it is often called) number i is often denoted instead by $j = \sqrt{-1}$, especially by the electrical engineering community, but we will stick to the traditional i. Thus, this Appendix assumes some comfort with basic notions of the arithmetic of complex numbers.

We begin with the formal definition of the *natural logarithm* (also called the *Naperian* or the *hyperbolic* logarithm), $\ln x$:

$$\ln x \equiv \int_1^x \frac{dt}{t}, \tag{4A.2}$$

where the t in the integrand of eq. (4A.2) is a *dummy variable* of integration, and where three special values of the natural logarithm are noted:

$$\ln 1 = 0, \qquad \ln 0 = -\infty, \qquad \ln e = 1. \tag{4A.3}$$

The number e is defined as:

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.7182818284 \ldots. \tag{4A.4}$$

In view of the properties (4A.3), the Taylor series representation (see eq. (4.1)) of the natural logarithm is defined in terms of an argument that is centered around the value $a = 1$ (for x $\neq -1$ and $|x| \leq 1$):

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots. \tag{4A.5}$$

Further, the natural logarithm is related to the *common* or *Briggs* logarithm, which we colloquially call the *logarithm to base 10*, by

$$\ln x = (\ln 10)(\log_{10} x) \cong 2.303 \log_{10} x. \tag{4A.6}$$

The *exponential function* is defined as the inverse of the natural logarithm, that is, $x = \ln y$ if $y = e^x$. The Taylor series for the exponential function is:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots . \tag{4A.7}$$

The results that now follow are obtained by formal manipulation from this Taylor series. For example, from eq. (4A.7) it can be shown that

$$e^{x+y} = 1 + (x+y) + \frac{(x+y)^2}{2!} + \frac{(x+y)^3}{3!} + \cdots + \frac{(x+y)^n}{n!} + \cdots = e^x e^y, \tag{4A.8}$$

and that for complex numbers a

$$e^{ax} = 1 + ax + \frac{(ax)^2}{2!} + \frac{(ax)^3}{3!} + \cdots + \frac{(ax)^n}{n!} + \cdots = (e^x)^a. \tag{4A.9}$$

Further, building on the result (4A.9), we can confirm the formula (4A.1) that Feynman found remarkable, which is known as the *De Moivre Theorem*:

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\right) = \cos x + i \sin x. \tag{4A.10}$$

In the last step of eq. (4A.10) we are recognizing the standard Taylor series expansions of the trigonometric functions that appear as the middle terms in that equation. Further,

$$e^{-ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right) - i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\right) = \cos x - i \sin x, \tag{4A.11}$$

so that from eqs. (4A.10) and (4A.11) we find we can write the trigonometric functions as:

$$\cos x = \frac{1}{2}\left(e^{ix} + e^{-ix}\right), \quad \sin x = \frac{1}{2i}\left(e^{ix} - e^{-ix}\right) \tag{4A.12}$$

We are now in a position to write down relations for the hyperbolic functions by replacing x by ix, and recalling the definition of i, so that:

$$\cosh x = \frac{1}{2}\left(e^{-x} + e^{x}\right) = \cos(ix), \quad \sinh x = \frac{1}{2}\left(e^{x} - e^{-x}\right) = -i \sin(ix). \tag{4A.13}$$

It also follows from eqs. (4A.13) that

$$\cosh(ix) = \frac{1}{2}\left(e^{ix} + e^{-ix}\right) = \cos x, \ \ \sinh(ix) = \frac{1}{2}\left(e^{ix} - e^{-ix}\right) = i \sin x. \tag{4A.14}$$

While the structure and appearance of the trigonometric and hyperbolic functions appear to be very similar, their behavior is not. The trigonometric functions are periodic, with period 2π, and their values are always bounded by ± 1, that is, $-1 \leq (\sin x, \cos x) \leq 1$. The hyperbolic cosine increases monotonically for both positive and negative values of its argument, while the hyperbolic sinusoid is asymmetric about the origin and so approaches $-\infty$ as $x \to -\infty$. Oh, what a difference an i makes! We show further details of all of the elementary transcendental functions in Table 4A.1.

Table 4A.1 Behavioral features of the elementary transcendental functions.

$f(x)$	Value at $x = 0$	Behavior as $x \to \infty$	Behavior of $f(x)$
$\sin x$	0	$\lvert \sin x \rvert \leq 1$	Oscillates continuously between ± 1
$\cos x$	1	$\lvert \cos x \rvert \leq 1$	Oscillates continuously between ± 1
e^x	1	$\to \infty$	Uniformly increases as $(x > 0) \to \infty$
$\sinh x$	0	$\to \infty$	Uniformly increases as $(x > 0) \to \infty$; Uniformly decreases as $(x < 0) \to -\infty$
$\cosh x$	1	$\to \infty$	Uniformly increases as $x \to \pm\infty$
$\ln x$	$-\infty$	$\to \infty$	Uniformly increases as $(x > 0) \to \infty$
$\log_{10} x$	$-\infty$	$\to \infty$	Uniformly increases as $(x > 0) \to \infty$

Finally, some derivatives and integrals of the elementary transcendental functions are:

$$\frac{d}{dx}\sin x = \cos x, \qquad \frac{d}{dx}\cos x = -\sin x,$$
$$\frac{d^2}{dx^2}\sin x = -\sin x, \qquad \frac{d^2}{dx^2}\cos x = -\cos x. \tag{4A.15}$$

$$\frac{d}{dx}e^x = e^x, \qquad \frac{d^n}{dx^n}e^x = e^x. \tag{4A.16}$$

$$\frac{d}{dx}\sinh x = \cosh x, \qquad \frac{d}{dx}\cosh x = \sinh x,$$
$$\frac{d^2}{dx^2}\sinh x = \sinh x, \qquad \frac{d^2}{dx^2}\cosh x = \cosh x. \tag{4A.17}$$

$$\frac{d}{dx}\ln x = \frac{1}{x}, \qquad \int \ln x\, dx = x\ln x - x. \tag{4A.18}$$

4.10 References

M. Abramowitz and I. A. Stegun (Eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Applied Mathematical Series 55, National Bureau of Standards, Washington, D.C., 1964.

A. H-S. Ang and W. H. Tang, *Probability Concepts in Engineering Planning and Design*, Vol. 1, John Wiley, New York, 1975.

R. A. Becker, *Introduction to Theoretical Mechanics*, McGraw-Hill, New York, 1954.

J. C. Burkill, *A First Course in Mathematical Analysis*, Cambridge University Press, London and New York, 1967.

K. C. Crandall and R. W. Seabloom, *Engineering Fundamentals in Measurements, Probability, Statistics, and Dimensions*, McGraw-Hill, New York, 1970.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

C. E. Ebeling, *An Introduction to Reliability and Maintainability Engineering*, McGraw-Hill, New York, 1997.

R. M. Felder and R. W. Rousseau, *Elementary Principles of Chemical Processes*, John Wiley, New York, 2000.

R. Haberman, *Mathematical Models*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

J. P. Holman, *Experimental Methods for Engineers*, McGraw-Hill, New York, 1971.

R. W. Hornbeck, *Numerical Methods*, Quantum, New York, 1975.

D. Huff, *How to Lie with Statistics*, Norton, New York, 1954.

L. B. W. Jolley, *Summation of Series*, Dover Publications, New York, 1961.

M. A. Lavrent'ev and S. M. Nikol'skiĭ, "Analysis," in A. D. Aleksandrov, A. N. Kolmogorov, and M. A. Lavrent'ev (Eds.), *Mathematics: Its Content, Methods, and Meaning*, Vol. I, MIT Press, Cambridge, MA, 1963.

C. Lipson and N. J. Sheth, *Statistical Design and Analysis of Engineering Experiments*, McGraw-Hill, New York, 1973.

W. Mendenhall and T. L. Sincich, *Statistics for Engineering and the Sciences*, 4th Edition, Prentice-Hall, Englewood Cliffs, NJ, 1995.

B. O. Pierce and R. M. Foster, *A Short Table of Integrals*, 4th Edition, Ginn and Company, Boston, MA, 1956.

H. J. Schenck, Jr., *Theories of Engineering Experimentation*, McGraw-Hill, New York, 1968.

F. W. Sears and M. W. Zemansky, *Modern University Physics*, Addison-Wesley, Reading, MA, 1960.

H. D. Young, *Statistical Treatment of Experimental Data*, McGraw-Hill, New York, 1962.
H. Ziegler, *Mechanics*, Vol. 1, Addison-Wesley, Reading, MA, 1965.

4.11 Problems

4.35. Estimate the error made in approximating $y(x) = \sin x$ with a Taylor's formula to $n = 4$ by evaluating the remainder R_5.

4.36. Do the statements that $\sin x \ll 1$ and $\tan x \ll 1$ produce similar approximations? Confirm and explain your answer.

4.37. The readings of an old-fashioned analog voltmeter—it has dials, not digital readouts!—are subject to some systematic error where all of its readings are too large. The magnitude of the error has been found to vary linearly from 1 V at a dial reading of 5 V to 4 V at a dial reading of 80 V.

(a) What are the correct voltages for dial readings of 80, 100, 50, 1, 35, and 10 V?
(b) What is the percentage error for each of the six (6) readings in part (a)?

4.38. (a) Is it possible to have a set of measurements that are precise but not accurate? Explain.
(b) Is it possible to have a set of measurements that are accurate but not precise? Explain.

4.39. (a) Write the Taylor series expansion for e^x about $x = 0$.
(b) Calculate $e^{0.5}$ to five significant figures using the first four terms of the series found it part (a).

4.40. (a) What percentage error was incurred in the calculation of part (b) of Problem 4.39 if the "true value" of $e^{0.5}$ is 1.6487?
(b) Use the Taylor remainder (eq. (4.5)) to calculate the error in $e^{0.5}$ after only four terms. Is the error calculated in part (a) of this problem acceptable? Explain.

4.41. Evaluate the following function by hand (no calculators or computers, please) for $x = 4$:

$$\left(1 + \frac{2}{x}\right)^{1/4}.$$

4.42. How does an observer know when enough is enough, that enough measurements have been taken?

4.43. Make a list of five new (i.e., not found in the text) examples of systematic errors.

4.44. Make a list of five new (i.e., not found in the text) examples of random errors.

4.45. The resistance of a resistor, R, is made by passing several currents, I, through it and measuring the corresponding voltage drops, V, and currents with imprecise, analog meters. The resulting data are:

$x_i = V(V)$	10	20	30	40	50	60	70	80
$y_i = I(A)$	0.8	1.1	2.5	4.2	4.3	4.7	5.8	6.4

(a) What kinds error will be found in the data?
(b) Assuming that $V = IR$, plot the data (by hand!) and "eyeball" in the best-fit line for that data.

4.46. Use the method of least squares to plot a V versus I curve for the data of Problem 4.45. How does it compare with the "eyeball" result of Problem 4.45?

4.47. The data presented below comprise 100 readings of noise levels taken 6 mi away from an airport, taken late in an evening at 15 s intervals. Find the mean, median, and standard deviation of these data.

Observed Decibel Values (dB), $n = 100$

50	50	53	48	45	51	57*	75*	85*	82*
75*	71*	65*	61*	60*	60*	55*	55*	51	50
49	49	48	51	49	54	48	48	47	49
49	49	49	49	48	47	50	49	48	49
47	48	48	50	50	54	48	47	47	48
48	49	48	47	50	49	48	48	48	48
48	48	52	50	53	49	49	48	49	47
49	55	51	50	49	48	49	45	48	50
50	51	49	50	47	47	47	47	47	47
48	50	49	49	49	49	49	49	56	49

4.48. The starred numbers in the data of Problem 4.47 are readings taken while an aircraft was flying directly overhead. If these data are deleted, what are the mean, median, and standard deviation of the remaining 88 data points?

4.49. Draw (a) a histogram of all of the data of Problem 4.47 and (b) a continuous curve of the number of readings as a function of the measured noise level.

4.50. Determine a *far-field approximation* of the function $f(r)$ given below as a binomial expansion for values of $r \gg a$.

$$f(r) = \sqrt{a^2 + r^2}.$$

4.51. The electric potential, V_e, at a distance, r, along the axis of revolution of a disk of radius a is given by

$$V_e = \frac{q}{2\pi a^2 \varepsilon_o}(\sqrt{a^2 + r^2} - r),$$

where q is the total charge that is distributed uniformly over the surface of the disk and ε_0 is the permittivity constant. Using the results of Problem 4.50, find a far-field approximation for the electric potential for values of $r \gg a$.

4.52. Compare the minimum number of terms kept in the binomial expansions of the solutions to Problems 4.50 and 4.51. Are those numbers the same, or not? Why are those numbers the same, or not?

4.53. Suppose we need to calculate the radial extension or deflection w of a very thin, spherical balloon, meaning that the sphere's radius extends from R to $R + w$ as the balloon is pressurized. It is made of an elastic material. A colleague finds a textbook that shows a formula for the pressure, p, that looks reasonable:

$$\frac{w}{R} = \frac{pR}{Eh},$$

where h is the balloon's wall thickness, and E is the modulus of the material of which the sphere is made. Is this equation dimensionally consistent?

4.54. Analyze the limit behavior of the equation presented in Problem 4.53 as the pressure, modulus, radius, and thickness both go to zero and become infinitely large. Does this limit behavior conform with your intuitive estimate of what should happen?

4.55. Use the equation in Problem 4.53 to derive an estimate of the magnitude of the pressure, p, as a fraction of the modulus, E. Estimate the pressure fraction for a thin-walled sphere, for which $h/R \ll 1$.

5
Exponential Growth and Decay

This chapter is devoted to a discussion of *exponential models* that share a common characteristic: The rate of change of a variable, whether positive (as it grows) or negative (as it decays), is directly proportional to the immediate value of that variable. More often than not, the rate of change is a *time* rate of change that is proportional to the variable's *instantaneous* value. Similar exponential decays also occur spatially, that is, with respect to a spatial coordinate. Here, behaviors decay over some distance so as to have little or no effect at distances sufficiently far from the initiating behavior. We will see that exponential models are ubiquitous and have many applications, including in physics, finance, and population and resource predictions.

5.1 How Do Things Get So Out of Hand?

As we have just indicated, the primary characteristic of exponential growth or decay of a population is the dependence of the rate of growth of the population on its size at any instant. Thus, if a population is large, its growth rate will be proportionately large, and its continuing growth will accelerate with its increasing size. As we will soon see, this kind of growth exhibits itself in nice, smooth curves whose ordinate values increase very rapidly in relatively short periods of time. One application area where **Why?**

this behavior is often modeled is the field of population studies. Indeed, much has been made in recent years of the dangers of overpopulation and of the related resource and environmental issues. In fact, with regard to the principles of modeling outlined in Section 1.2, common sense would indicate in this instance that we have a pretty good idea of what we are looking for, what we know, and what we want to know.

Consider the two population projections shown in Figures 5.1 and 5.2. Even though they are now somewhat dated, both curves project very rapid increases in the world's population in relatively short times. The first curve (Figured 5.1) reflects both historical data for the years prior to 1960 and a projection from a 1960 world population estimate of 3 billion people growing at a rate of 2% per year. The world population was quite small until 1700, but it has been growing rapidly since the end of the 19th century.

Find? However, even though the projections past 1960 are at a modest rate of 2% per year, we should wonder about the validity of the steepness of the projection, especially after the year 2100.

If we were to extend the projection shown in Figure 5.1 for another 700 or 800 years, we would obtain the results shown in Figure 5.2. The assumed annual growth rate is still 2% and the population is still measured in billions. However, the time scale has been expanded by a factor of two and the population projections are now measured in *millions of billions*! While these population projections are almost certainly unrealistic, the

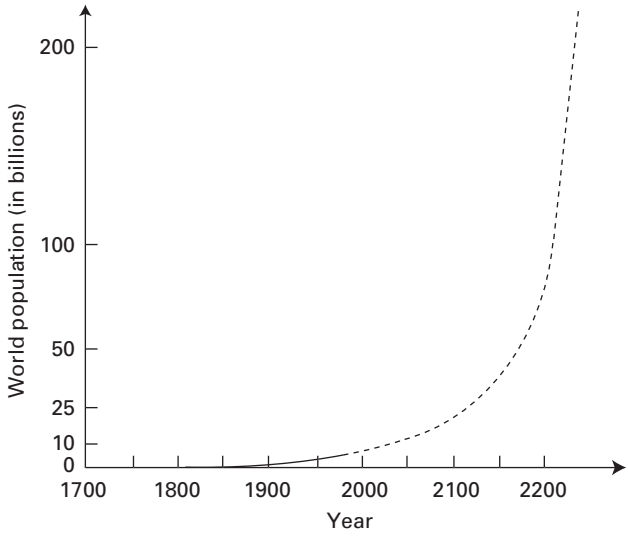


Figure 5.1 A historical view (solid line, for 1700–1960) and a projection (dashed line, for 1960–2165) of the world's population.

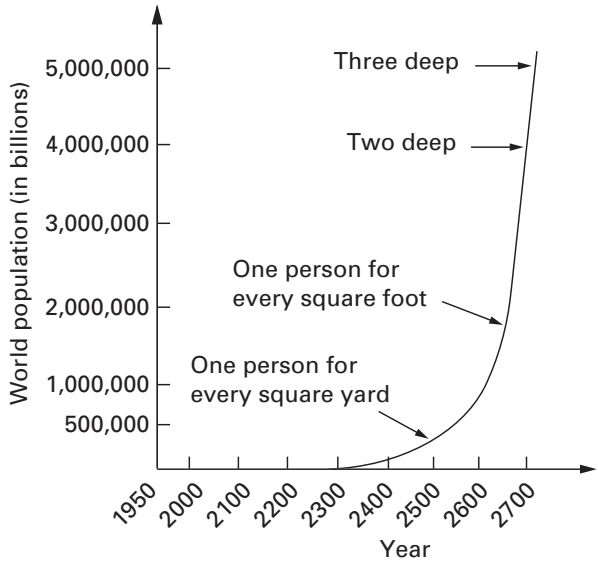


Figure 5.2 A longer projection, for 1950–2700, of the world's population, annotated to show the amount space that each person would be accorded were the projections to become reality.

projection curve clearly illustrates the nature of unrestrained exponential growth: The bigger it is, the faster it grows.

We also emphasize (again, as in Section 3.5.2) the importance of scaling when examining exponential growth. Consider the magnitudes of the numbers involved. For example, at a 2% annual growth rate, the world population grows from 3 billion in 1960 to 5,630,000 billion in 2692 (cf. Figure 5.2). What does it mean to have

$$5{,}630{,}000 \text{ billion people}$$

or

$$5{,}630{,}000{,}000{,}000{,}000 \text{ people}$$

or

$$5.63 \times 10^{15} \text{ people}$$

on earth in the year 2692? Is there room for all of these people? Could we even count this many people in a census? (And if you think that this is not a meaningful question or issue, there were vigorous debates within the U.S. Congress about the role of statistical sampling in the 2000 census—and they were talking about counting "only" some 285 million Americans!)

What? Let's try to answer the "room" issue first, that is, is there space enough on earth for more than five million billion people? The total surface area of the earth is approximately $5.10 \times 10^8\ \text{km}^2$ ($\sim 5.49 \times 10^{15}\ \text{ft}^2$), of which 72%
Given? is water. Assuming that people cannot stand on water, the net "standing area" is approximately $1.43 \times 10^8\ \text{km}^2$ ($\sim 1.54 \times 10^{15}\ \text{ft}^2$). If each person were given just 1 square foot of personal standing space, people would have to be stacked more than three deep in order to accommodate everyone!

How long would it take to physically count all of the people on earth in
Assume? 2692? Suppose we could tally the population at a rate of 1000 people per
Predict? second. Then it would take

$$\frac{5.63 \times 10^{15}\ \text{people}}{1000\ \text{people/s}} = 5.63 \times 10^{12}\ \text{s}.$$

This seems like a lot of counting time. In fact, it easily shown that this simple calculation suggests that it would take almost 200,000 years to count the population growth that occurred in (only!) 800 years at a 2% annual growth rate.

We have presented the above numbers in part because they are patently
Valid? absurd, to show just how things get out of hand. These numbers show how simplistic calculations with exponentials can lead to results that are arithmetically correct yet fail the test of basic credibility. We also note again the effect of scale in displaying such results. The ordinate scales of Figures 5.1 and 5.2 are linear and represent, respectively, 100 billion people per 1.50 in of graph and 2,000,000 billion per in. To express the projected population of 5.63×10^{15} people on the same ordinate scale of Figure 5.1, we would need a piece of paper that is 85,000 in long (you do the math!). It is also readily shown (see, for example, Problems 5.38 and 5.40) that exponential curves do not always portray such dramatic results.

Remember, therefore, that a change in scale does not, by itself, generate or dissipate true exponential behavior. Scale changes add or disguise
Improve? perspective on the underlying mathematics. What is more important is that exponential behavior can express other kinds of response, illustrated in Figure 5.3, both of which occur when the proportionality factor is negative.Figure 5.3(a) shows a classic *decay* or dissipation curve in which an initial value decays to zero, while Figure 5.3(b) shows how some variable grows evermore slowly, *asymptotically*, to a limiting value as time becomes infinite. We will see both of these behaviors in Section 5.4, for example, when we describe the charging and discharging of a capacitor in a very elementary electrical circuit. Thus, after we introduce the necessary mathematics, we should also expect to see mathematical behavior that is

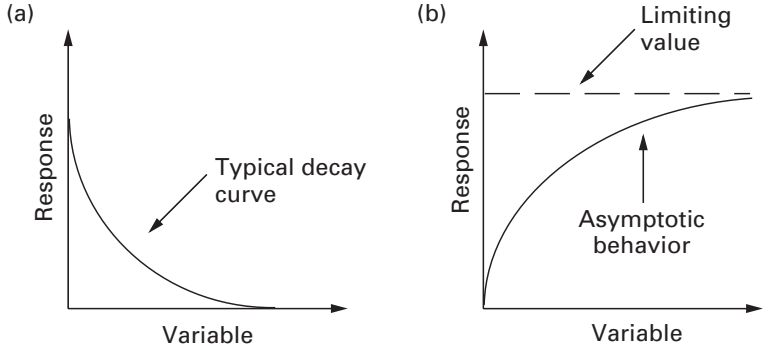


Figure 5.3 Illustrations of the kinds of exponential behavior that result when the constant of proportionality is less than zero (negative): (a) classic decay from a given initial value; and (b) asymptotic growth toward a limiting value or asymptote as time becomes indefinitely large.

more complicated and more interesting than simple, unrestrained exponential growth. We will then see that such exponential behavior is an important part of very practical and useful modeling in many disciplines. The foregoing discussion should, therefore, be taken as a cautionary "word to the wise" about some of the dangers in exponential modeling, not as a reason to dismiss or ignore it.

Problem 5.1. If you were asked to conduct a population study, what would you be looking for, what would be known, and what would you want to know?

Problem 5.2. What sort of assumptions would you make if you were asked to conduct a population study? On what basis could those assumptions be justified?

Problem 5.3. What factors might restrain or otherwise influence the unrestrained growth seen in Figures 5.1 and 5.2?

Problem 5.4. Determine the radius of the earth in both meters and feet from the surface areas given in Section 5.1. Are these values consistent with the conversion factors given in Table 2.3?

Problem 5.5. Confirm that it would take almost two hundred thousand years to count a population of 5.63 million billion people at a rate of 1000 people/s.

5.2 Exponential Functions and Their Differential Equations

In this section we first describe some of the arithmetic that underlies all of the numbers and curves given in Section 5.1. We follow that with a very brief primer on the first-order differential equation from whence derives the exponential function. This primer is intended to serve as a reminder of—not a substitute for—comparable introductory material in differential equations.

5.2.1 Calculating and Displaying Exponential Functions

The exponential behavior discussed in Section 5.1 can be put in mathematical terms as follows. Let $N(t)$ be the number or population of a collection of objects, and let t be the independent variable on which N depends and with which it changes. For most of our applications, t will be associated with time, but that is a result of the models we exhibit, not due to any underlying mathematical requirement. As we indicated in Section 5.1, exponential growth results when *the rate of growth is proportional to a population or number.* If we introduce a *constant of proportionality*, λ, then exponential growth occurs when

$$\frac{dN(t)}{dt} = \lambda N(t). \tag{5.1}$$

We see from eq. (5.1) that the constant of proportionality λ can be written as

$$\lambda = \frac{dN/N}{dt}. \tag{5.2}$$

Thus, λ represents the fractional change dN/N of the population per unit change of the independent variable, *dt.* The dimensions of λ are seen to be

$$[\lambda] = \frac{1}{[t]} = [t^{-1}]. \tag{5.3}$$

If the independent variable, t, is a measure of time, then the dimensions of λ are 1/time.

Equation (5.1) is a first-order differential equation that is linear in the dependent variable $N(t)$ and has constant coefficients. As we show in the next section, eq. (5.1) has a solution that can be written as

$$N(t) = N_0 e^{\lambda t}, \tag{5.4}$$

where the constant N_0 is an *arbitrary constant* whose value remains to be determined. The dimensions (and units) of N_0 must be the same as those of $N(t)$. Further, the number e is the *base of the natural logarithm.* It has an approximate value $e \cong 2.71828$. Since $e^0 = 1$, it also follows that the number N_0 must be the *initial value* of the population, that is, the number of objects whose change we are modeling at $t = 0$, when the model "starts." Note, too, that $N(t)$ grows in time if λ is positive, much like the curves in Figures 5.1 and 5.2, and that it decreases in magnitude or decays if $\lambda < 0$, as does Figure 5.3(a).

Since e is the base of natural logarithms, we can take the (natural) logarithm of both sides of eq. (5.4) to show that

$$\lambda t = \ln N(t) - \ln N_0 = \ln \left(N(t)/N_0\right). \tag{5.5}$$

Equation (5.5) tells us that if we want to find a time, t_n, when the number $N(t_n) = nN_0$, that is, when the population size is a specified multiple of its initial value, all we need to do is calculate

$$t_n = \frac{\ln n}{\lambda}. \tag{5.6}$$

People frequently ask how long it takes something to double in size, in which case the answer is the *doubling time*, t_2, determined from eq. (5.6) with $n = 2$:

$$t_2 = \frac{\ln 2}{\lambda} \cong \frac{0.693}{\lambda}. \tag{5.7}$$

One immediate application of eq. (5.7) is to investment: Money grows as it earns interest. Suppose that we want to know how long it would take to double an amount of money with continuously compounded interest. We determine that by interpreting λ in terms of percentage, P, in which case $P = 100\lambda$. Then eq. (5.7) becomes

Why?

Find?

$$t_2 = \frac{69.3}{P}. \tag{5.8}$$

The approximate time it would take to double some money as a function of different percentage growth rates P is shown in Table 5.1.

There are two other interesting properties of exponential growth. The first is the *inversion* of the doubling time that occurs when we calculate the *half-life* of a population. That is, suppose we want to know how long it takes for a population that started at N_0 to decrease to a value of $N_0/2$. In this case, λ would represent a (negative) decay rate, and from eq. (5.6) we would get a formula for the half life $t_{1/2}$ that is formally identical to eq. (5.7) or eq. (5.8). Thus, we need only change the column headings in Table 5.1 to "Annual Decay ($P < 0, \%$)" and "Half-Life ($t_{1/2}$, years)," respectively, to obtain the variation of half-life as a function of decay rate.

Table 5.1 The time it takes to double one's money, measured in years, as a function of continuously compounded growth rates, measured in percentages.

Annual Growth (P, %)	Approximate Doubling Time (t_2, years)
1	69.3
2	34.6
5	13.9
10	6.93
20	3.46

The second interesting property is this. The time, t_n, it takes for a population, $N(t)$, to grow by a constant factor, n, remains unchanged throughout the growth. Thus, from time $t = 0$ to $t = t_2$, the population doubles; from $t = t_2$ to $t = 2t_2$, the population doubles again; and so on. Thus, we obtain the results shown in Table 5.2.

Finally, for this section, some remarks on the display of exponential functions are now in order. We saw in Section 5.1 that exponential growth can lead to some horrifically large numbers. However, in the same way that great strengths and great weaknesses are often intertwined, it is similarly the case that the logarithms of exponential growth provide the means of graphical (and representational) salvation. If we look back at eq. (5.5), we see that one representation of exponential behavior can be expressed in the form:

$$\ln N(t) = \lambda t + \ln N_0. \tag{5.9}$$

Equation (5.7) suggests that a *semi-logarithmic plot* of $\ln N(t)$ against λt (plus a constant) would produce results in which the ordinate values are

Table 5.2 The growth of the exponential function as gauged by multiples of the doubling time.

Time (units of t_2)	Population ($N(t)$)	
$t = 0$	$N = N_0$	$= 2^0 N_0$
$= t_2$	$= 2N_0$	$= 2^1 N_0$
$= 2t_2$	$= 4N_0$	$= 2^2 N_0$
$= 3t_2$	$= 8N_0$	$= 2^3 N_0$
$= 10t_2$	$= 1024N_0$	$= 2^{10} N_0$
$= nt_2$		$= 2^n N_0$

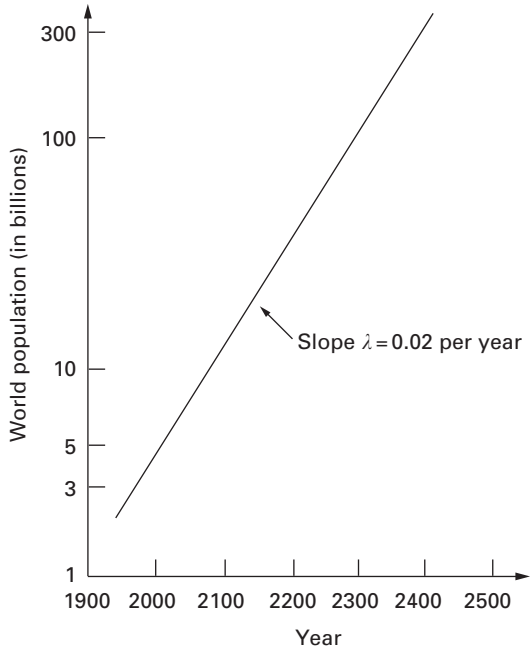


Figure 5.4 Population projections for the period 1960–2400 based on the data of Figures 5.1 and 5.2, presented herein in a *semi-logarithmic* plot. What had previously been displayed as a set of steeply rising exponential curves is now seen as a relatively benign straight line with ordinate values in particular that are much more manageable.

more commensurate with those of the abscissa. In fact, in such a semi-logarithmic plot, eq. (5.9) represents a straight line of slope λ and with intercept $\ln N_0$. In Figure 5.4 we show such a linear "semi-log" using the projected data of Figures 5.1 and 5.2.

Problem 5.6. Confirm by differentiating eq. (5.4) that the $N(t)$ given therein satisfies eq. (5.1).

Problem 5.7. How would the projections of Figures 5.1 and 5.2 change if the growth rate were, respectively, 1% per year and 3% per year?

Problem 5.8. What annual growth rate would be needed to double one's money in seven years?

5.2.2 The First-Order Differential Equation $dN/dt - \lambda N = 0$

There is another interesting property of exponential behavior that has been present but to which we have not paid much attention in our discussion thus far. This special property is the fact that there is only *one* arbitrary constant in the basic exponential model [see the discussion immediately after eq. (5.4)]. Why is that so? There is only one constant because, as we will now demonstrate, the exponential function (5.4) is the solution to a first-order differential equation, that is, a differential equation in which the highest-order derivative is of first order. The single arbitrary constant arises from the fact that a first-order differential equation needs to be integrated just once to obtain a solution.

Consider the differential equation governing population growth set out in eq. (5.1):

$$\frac{dN(t)}{dt} - \lambda N(t) = 0. \tag{5.10}$$

This differential equation has *constant coefficients*, that is, the multipliers of both $N(t)$ and its derivative are constants, namely, λ and 1, respectively. Equation (5.10) can also be written in the form

$$\frac{dN(t)}{N(t)} - \lambda dt = 0, \tag{5.11}$$

which can be integrated in exactly the same way that the Naperian logarithm was defined in Section 4.9 and then inverted to yield the solution (see Problem 5.9):

$$N(t) = Ce^{\lambda t}. \tag{5.12}$$

We can clearly identify C as the initial population by setting $t = 0$ in eq. (5.12). Equally clearly, we can identify that initial value in the notation introduced in eq. (5.4): $C = N_0$.

The initial value C need not be determined at the time $t = 0$. We could specify a starting condition that at some time t_0, $N(t_0) = N_0$. Equation (5.12) then dictates that

$$N(t_0) = N_0 = Ce^{\lambda t_0},$$

which means in turn that

$$C = N_0 e^{-\lambda t_0}. \tag{5.13}$$

If we substitute this form of our constant of integration C into the solution (5.12), we get

$$N(t) = N_0 e^{-\lambda t_0} e^{\lambda t} = N_0 e^{\lambda (t - t_0)}. \tag{5.14}$$

This obviously defines a population that for $\lambda > 0$ is increasing through N_0 at $t = t_0$, but that is less than N_0 for $t < t_0$.

Note that all of the foregoing manipulations are as valid for $\lambda < 0$ as they are for $\lambda > 0$. The interpretations would obviously be different, since we would be describing exponential decay ($\lambda < 0$) rather than exponential growth ($\lambda > 0$), but the underlying mathematics is unchanged. However, it is also true that the analysis to date is limited by the fact that our basic differential equation (5.10) is a *homogeneous equation*, that is, there is no *forcing function* on the right-hand side. When we discuss the charging of a capacitor in a simple electrical circuit in Section 5.4, we will see that the charge $q(t)$ in the capacitor in that circuit is described by an equation of the form

$$\frac{dq(t)}{dt} + \frac{1}{RC}q(t) = V_{in}(t). \tag{5.15}$$

Equation (5.15) looks very much like the differential equation (5.10) for exponentials, except that it has a *forcing function*, $V_{in}(t)$, on the right-hand side that forces or drives the change of the voltage in the circuit being modeled. Further, the coefficient in eq. (5.15) is equivalent to taking $\lambda = -(1/RC) < 0$ in eq. (5.10).

Problem 5.9. Verify that eq. (5.12) is the solution to the exponential differential equation as given in eq. (5.10) by using the result that

$$\int \frac{du}{u} = \ln u + \text{constant}.$$

Problem 5.10. Show that the solution (5.12) to the differential equation (5.10) can also be found by assuming the following trial solution for $N(t)$:

$$N(t) = Ce^{\alpha t}.$$

Problem 5.11. Why is the proportionality constant in eq. (5.15) equivalent to having $\lambda < 0$ in eq. (5.10)? What sort of behavior would we then expect?

5.3 Radioactive Decay

We now want to model the decay of radioactive isotopes as exponential behavior. As physicists and chemists began to study radioactivity at the **Why?**

end of the 19th century, they found that the activity of radioactive isotopes decreased with time at rates that varied with the material. When the emission of α and (primary) β particles was observed in the laboratory, it was found that the number of particles collected over time was unaffected by changes in pressure, temperature, chemical state, or the physical environment. Instead, the observed *half-life* of each isotope—the time it takes for the number of particles of the isotope to be reduced by half—was found to be a characteristic of the material itself. Thus, once half-life is identified as a material property, a measurement of a radioactive decay pattern can be used to identify a material by its characteristic half-life.

Find?

Given? In Figure 5.5 we show a generic, semi-logarithmic plot of the radioactive decay of an unspecified material. It strongly resembles Figure 5.4. In the radioactive decay model, however, the proportionality constant is negative (i.e., $\lambda < 0$). Further, in Figure 5.5, we have rendered the abscissa dimensionless by measuring it in terms of an (unknown) half-life, $t_{1/2}$. That is,

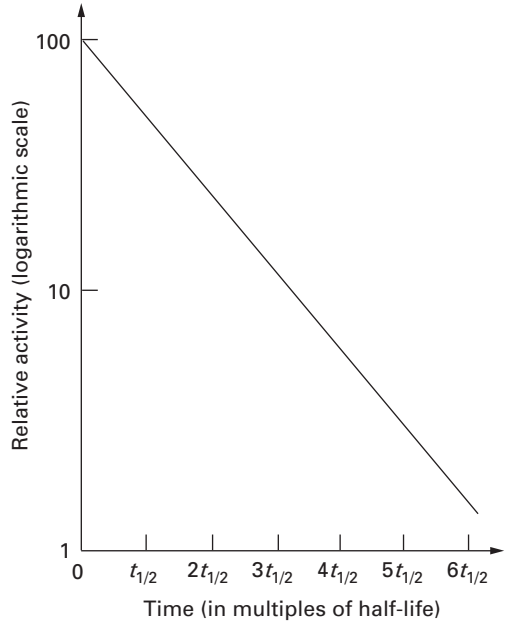


Figure 5.5 A generic plot of the decay of a radioactive isotope. Note that the data is presented in a *semi-logarithmic* plot. Note, too, that the abscissa has been made dimensionless by measuring it in terms of an (unknown) half-life, $t_{1/2}$.

time is measured here as a multiple of a parameter, $t_{1/2}$, whose dimensions (of time) are given but whose specific numerical value is not.

Decay rates are often used to characterize emitters as *short-lived* or *long-lived*. For example, consider the decay of the element thorium (Th in the atomic table of the elements). Thorium has a half-life of 16,500,000,000 yr, which does seem like quite a long time! If that is indeed true, can we calculate the effective decay constant, λ, and estimate how many thorium atoms will decay in a year? **Predict?**

We can calculate λ by applying eq. (5.6) with $n = 1/2$ and $t_{1/2} = 1.65 \times 10^{10}$ yr. Then the decay constant can be calculated as **How?**

$$\begin{aligned}\lambda &= \frac{-0.693}{t_{1/2}} = \frac{-0.693}{1.65 \times 10^{10}\,\text{yr}} = -4.20 \times 10^{-11}\,\text{yr}^{-1} \\ &= -4.20 \times 10^{-11}\frac{1}{\text{yr}} \times \frac{\text{yr}}{365\,\text{day}} \times \frac{\text{day}}{86{,}400\,\text{sec}} \\ &= -1.33 \times 10^{-18}\,\text{sec}^{-1},\end{aligned} \tag{5.16}$$

where *reciprocal seconds* are the units ordinarily used to express radioactive decay constants. In view of the definition (5.2) of decay rate in terms of fractional population change, eq. (5.16) suggests that only one thorium atom in every $(1.33 \times 10^{-18})^{-1} = 7.51 \times 10^{17}$ such atoms decays in one second. Indeed, even in a year, only one of every 2.38×10^{10} thorium atoms present initially will decay. Thus, it does seem that thorium can be safely characterized as a long-lived emitter.

It is worth touching on two related points here. One is that the characterization of a radioactive emitter as short- or long-lived seems, in the above context, a straightforward and neutral piece of scientific reasoning. However, similar calculations done in other contexts (e.g., the decay time for radioactive waste in a national storage facility for radioactive materials from nuclear power plants, or the remediation time for gasses to fully dissipate from a landfill) often turn these characterizations into political (and emotional) debates that try to define the meaning of "short (or long) enough for …" **Use?**

The second point is a deeply philosophical one about the very underpinnings of the models of physics. What does it mean for a fraction of a single isotope or atom to decay? Or, are the models really about averages calculated over a large number of particles? And, if that is the case, how are such averages calculated? And, further, what is the meaning of the various levels of models that are used to describe and predict these behaviors? **Valid?** **Improve?**

5.4 Charging and Discharging a Capacitor

Why? We will now model the behavior of a very simple electrical circuit, incorporating only two *passive* electrical elements, a *capacitor* defined by its capacitance, C, and a *resistor* defined by its resistance, R. These two ele-
Assume? ments are depicted in Figure 5.6. The first step in our circuit modeling is to identify a functional relationship for each element, called a *constitutive equation*, which expresses its behavior in terms of the voltage drop across the element and the current flowing through it.

How? The capacitor stores and discharges energy. This energy transfer occurs as charge is transferred from one side plate or electrode to the other (viz., Figure 5.6(a)) and, in this process produces a voltage drop across the capacitor given by:

$$[V_a(t) - V_b(t)]_C \equiv \Delta V_C(t) = \frac{q(t)}{C}, \tag{5.17}$$

where $\Delta V_C(t)$ represents the voltage drop across the capacitor while the *charge*, $q(t)$, flows through it. In SI units, C, the *capacitance*, is measured in coulombs (of charge) per volt or farads.

Keep in mind that while we are used to talking about current flowing through electrical devices in everyday life, here we are building our model in terms of the charge, $q(t)$, whose first derivative in time is the *current*,

$$i(t) \equiv \frac{dq(t)}{dt}. \tag{5.18}$$

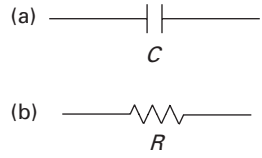


Figure 5.6 Simple conceptual drawings of the *icons* or symbols of two passive electrical elements: (a) the *capacitor*, denoted by C, stores energy by storing charge and discharges energy through the flow of charge (or current, the time rate of change of charge); and (b) the *resistor*, R, that allows charge or current to flow, but that in so doing dissipates some of the energy flow as wasted thermal energy.

We will return to and amplify this point in Section 8.7 wherein we model circuits more extensively and relate the electrical elements to analogous mechanical elements.

The second passive element, the resistor depicted in Figure 5.6(b), impedes or resists the flow of charge (or current) as the charge flows through the element. The resistor thus dissipates energy, usually perceived as wasted heat. The voltage drop across a resistor is usually expressed in terms of voltage and current as *Ohm's law*:

$$[V_a(t) - V_b(t)]_R \equiv \Delta V_R(t) = Ri(t), \tag{5.19}$$

where $\Delta V_R(t)$ represents the voltage drop across the resistor while the current, $i(t)$, flows through it. In SI units, R, the *resistance*, is measured in volts per ampere or ohms. Since we are interested in expressing our current model in terms of charge, we make use of eq. (5.18) to eliminate the current from eq. (5.19) and rewrite Ohm's law as:

$$\Delta V_R(t) = R\frac{dq(t)}{dt}. \tag{5.20}$$

5.4.1 A Capacitor Discharges

Having modeled our two circuit elements, we now model the simple electrical circuit shown in Figure 5.7. That picture shows a resistor in series with a capacitor, and with an (externally) applied voltage across the circuit's two "free" endpoints or nodes. Suppose first that no voltage is applied across the free endpoints. In that case, it seems quite logical to stipulate that the sum of the voltage drops across the capacitor and the resistor must simply vanish because nothing is being put into the system, that is,

$$\Delta V_C(t) + \Delta V_R(t) = 0. \tag{5.21}$$

If we substitute eqs. (5.17) and (5.20) into eq. (5.21), we can replace its voltage terms and express it entirely in terms of the charge $q(t)$ flowing around this simple circuit:

$$\frac{dq(t)}{dt} + \frac{1}{RC}q(t) = 0. \tag{5.22}$$

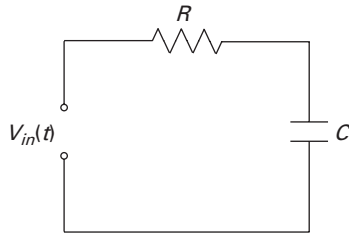


Figure 5.7 A very simple electrical circuit that connects a *capacitor*, C, in series with a *resistor*, R, and an externally applied input voltage, $V_{in}(t)$. Here we have directly connected the two elements rather than showing their individual nodes, but we have shown the two "free" nodes or endpoints that normally would serve as the terminals to which we would attach a battery or some other voltage supply.

Predict? The resemblance between eqs. (5.22) and (5.10) is unmistakable, so it follows immediately that the solution to eq. (5.22) can be written as [see eq. (5.12)]

$$q(t) = C_1 e^{-t/RC}, \tag{5.23}$$

where C_1 is an arbitrary constant that can be taken as the initial charge: $C_1 = q_0 = q(t = 0)$. Equation (5.23) shows that the capacitor's initial charge, q_0, left on its own, theoretically vanishes as $t \to \infty$. (In practice, the initial charge becomes so small that we can say it has vanished.) Described in Section 5.1, this behavior was shown in Figure 5.3(a).

Valid? We also see from eq. (5.23) that the behavior of a simple *RC circuit* occurs in times that we can express and measure in terms of a characteristic constant, namely, RC. This means, first of all, that the decay of the charge is inversely proportional to both the resistance and the capacitance. It is intuitively satisfying to see that the decay will be slowed if either the capacitor is large, in which case it can hold a larger charge that will take longer to dissipate, or if the resistance is large, in which case the discharge of current through the resistor will be slowed down. Second, it is not surprising that one widely used measure of the decay rate of such a circuit is a *time constant*, τ, defined as:

$$\tau = RC. \tag{5.24}$$

The time constant, τ, is the time it takes for an initial charge, $q_{0,}$ to be reduced to the value, q_0/e. With the definition (5.24), the RC circuit's governing equation can be written as

$$\frac{dq(t)}{dt} + \frac{1}{\tau}q(t) = 0. \tag{5.25}$$

Note also that with the governing equation written this way, dimensional consistency is much easier to discern and to verify.

Problem 5.12. Verify that each term in eq. (5.22) has the same physical dimensions.

Problem 5.13. Confirm that eq. (5.23) is the correct solution to eq. (5.22).

Problem 5.14. Use the definitions of resistance and capacitance to verify that the product RC has the physical dimensions of time.

5.4.2 A Capacitor Is Charged

Find? Can we extend the foregoing circuit model to charge the capacitor? We can, by inserting a voltage source across the two free endpoints of the RC circuit in Figure 5.7. (We should not confuse this with the familiar experience of charging a dead car battery by connecting it with jumper cables to a good battery because that charging results from a relatively rapid conversion of electrical energy to chemical energy.) How do we incorporate a voltage source to revise our circuit model?

How? There are two (at least) ways to answer this question. First, we would extend the reasoning behind eq. (5.21) by simply adding to that equation a term representing the input voltage $V_{in}(t)$ supplied by a battery or an equivalent device:

$$\Delta V_C(t) + \Delta V_R(t) = V_{in}(t). \tag{5.26}$$

Then, with the appropriate constitutive equations and the definition of the circuit's time constant, eq. (5.26) can be rewritten as

$$\frac{dq(t)}{dt} + \frac{1}{\tau}q(t) = \frac{V_{in}(t)}{R}. \tag{5.27}$$

The differential equation (5.27) is called *inhomogeneous* in the terms of mathematics because the voltage input makes its right-hand side take on a non-zero value.

Valid? We can also demonstrate that eq. (5.27) is a correct model by applying a classical result of electrical circuit analysis, *Kirchhoff's Voltage Law* (KVL), named after the German physicist Gustav Robert Kirchhoff (1824–1887). Kirchhoff observed that *the algebraic sum of the voltage drops across all of the elements connected in a closed circuit loop is zero.* Written in symbolic terms, the KVL looks like the following:

$$\sum_{k=1}^{K} [V_a(t) - V_b(t)]_k = \sum_{k=1}^{K} \Delta V_k(t) = 0, \tag{5.28}$$

where K is the total number of elements in the closed circuit loop. Note that we must pay close attention to the sign conventions built into the constitutive laws of the circuit elements when we apply the KVL because it calls for calculating an "algebraic sum" of the voltage drops. The KVL can be applied to the circuit in Figure 5.7 (see Problem 5.16) to find once again the result in eq. (5.27).

To return to our stated modeling task of charging a capacitor, let us apply eq. (5.27) under the simple assumption of a constant input voltage, $V_{in}(t) = V_0 =$ constant:

$$\frac{dq(t)}{dt} + \frac{1}{\tau} q(t) = \frac{V_0}{R}. \tag{5.29}$$

Remembering that the derivative of a constant is zero, it is not very hard to show (see Problem 5.17) that we can construct a solution to eq. (5.29) in the form

$$q(t) = V_0 C + C_1 e^{-t/\tau} = \tau V_0/R + C_1 e^{-t/\tau}. \tag{5.30}$$

Once again the single arbitrary constant, C_1, can be determined from the circuit's given initial conditions. In the simpler case where the initial charge is supplied only by the voltage input, it follows from eq. (5.30) that

$$q(0) = 0 = V_0 C + C_1.$$

The arbitrary constant is now determined and the complete correct solution becomes:

$$q(t) = V_0 C(1 - e^{-t/\tau}). \tag{5.31}$$

Predict? Equation (5.33) is plotted in Figure 5.8, which is a more detailed version of the sketch given in Figure 5.3(b). We see that the charge increases exponentially from its initially given value of zero. Here, however, the amount of

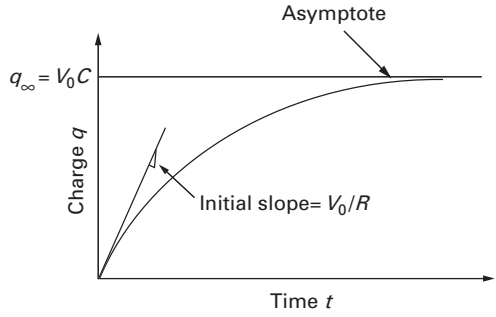


Figure 5.8 The charge in the capacitor when an input voltage, V_0, is applied across the terminals shown in Figure 5.7. Note the unmistakable resemblance of this drawing to the sketch in Figure 5.3(b). Here the initial slope, V_0/R, and the asymptotic value of the charge, $q_\infty = V_0 C$, are called out.

charge does not increase to infinity. Instead, it *asymptotically* approaches a maximum value given by

$$q_\infty \equiv q(t \to \infty) = V_0 C. \tag{5.32}$$

This asymptotic value of the charge, $q_\infty = V_0 C$, is the maximum value that the capacitor can hold for the given capacitance, C, and applied voltage, V_0. We also can calculate that the initial slope of the charging curve is V_0/R. If this slope were zero, then no charging would be possible and the charge would remain at its initial value of zero. Of course, this circumstance could only arise if no voltage were applied or if the resistance to that applied voltage was infinitely large. Thus, once again we have found results that are intuitively pleasing.

Improve?

One last point here. We have charged a capacitor even though the circuit's decay constant is negative, that is, $\lambda = -1/\tau$. We have thus imposed growth on an exponential system that otherwise would have decayed. This serves to point out that external conditions, such as the input voltage applied here, can influence the behavior of an exponential system to an extent not anticipated by the sign of the constant λ.

Problem 5.15. Verify that eq. (5.29) was properly derived from eq. (5.28).

Problem 5.16. Apply the KVL of eq. (5.28) to the circuit in Figure 5.7 to confirm the validity of eq. (5.27).

Problem 5.17. Confirm by differentiating eq. (5.31) that it satisfies eq. (5.29).

Problem 5.18. Calculate the initial slope of the charging capacitor from the solution given in eq. (5.31).

5.5 Exponential Models in Money Matters

Why? We will now talk about money, that is, we will model elementary exponential behavior as seen in two important aspects of our financial lives. First, we will talk about *interest* and *compound interest*, the repeated calculation of interest over shorter periods of time that produces higher effective rates of interest than may be evident. Then we will describe *inflation*, the phenomenon we see when prices rise rapidly and dramatically.

5.5.1 Compound Interest

It is hard to listen to the news these days without hearing reports on the stock and bond markets. When the markets and their underlying economies are not doing well, we hear about whether or not the Federal Reserve Bank will adjust the interest rates that the banks, including "The Fed," charge each other on interbank loans. We are besieged by advertisements promising high interest returns on various savings instruments and low interest rates on credit card balances and mortgage loans. For all this talk of interest, however, few understand that interest is an exponential phenomenon, which is one reason that economists speak of the *time value of money*, and that very serious consequences follow inattention to interest rates and compounding practices.

Find? Predict? Consider first the latest unsolicited offer of a credit card promising an interest rate of only 0.75% per month, which is advertised as "only" 9% per year and sounds cheap in this context. If the monthly interest was compounded on a monthly basis, the effective annual interest rate is found from the 12-fold multiplication

$$(1.0075)(1.0075)\cdots(1.0075) = (1.0075)^{12} \cong 1.0938. \tag{5.33}$$

Thus, monthly compounding produces an effective annual interest rate of about 9.38% per year. If these rates were *continuously compounded*, we would use eq. (5.4) to find:

$$\frac{N(t)}{N_0} = e^{(0.0075)(12\text{mos})} \cong 1.0942, \tag{5.34}$$

which represents an effective rate of 9.42% per year. Thus, depending on how interest is applied or compounded, the *effective interest rate* charged on a 9% (nominal) card would be 9.39% for monthly compounding and 9.42% for continuous compounding. (United States law requires that advertisements and transaction documents list the nominal, uncompounded APR or *Annual Percentage Rate*, with compounding details and effective rates often left to the fine print.) If these effective interest rates don't seem like a very big deal, consider that they add noticeable surcharges to the nominal rates.

We can also see the effects of compounding by looking at returns on investment. Suppose that interest is promised at a nominal rate of 10% per year. That interest could be calculated and distributed in discrete amounts of 10% annually, 5% semiannually, 2.5% quarterly, and so on. For m compounding periods per year, the initial investment would grow to:

How?

$$\left(\frac{N}{N_0}\right)_m = \left(1 + \frac{0.10}{m}\right)^m. \tag{5.35}$$

We have shown some results for various compounding intervals in Table 5.3. Note that the investment promises larger returns as the number of compounding periods, m, is increased. Thus, it seems interesting to consider what will happen to the value of the unit investment as the number of compounding periods becomes infinitely large.

Find?

Table 5.3 The growth of a unit investment (i.e., $N_0 = 1$) at a nominal rate of 10% with returns compounded and payable m times per year. Equation (5.35) is used to calculate that growth.

Number of Compounding Periods per Year (m)	Value of a Unit Investment ($N_0 = 1$)
0	1.0000
1	1.1000
2	1.1025
4	1.1038
12	1.1047
365	1.1051559

We can easily answer this question by first recasting eq. (5.35) in terms of a new variable $x = m/0.10$. Then eq. (5.35) becomes:

How?

$$\left(\frac{N}{N_0}\right)_m = \left(1 + \frac{1}{x}\right)^{0.10x}. \tag{5.36}$$

We now take the limit of eq. (5.36) as x becomes infinitely large:

$$\left(\frac{N}{N_0}\right)_\infty = \lim_{x\to\infty}\left[\left(1+\frac{1}{x}\right)^x\right]^{0.10}. \tag{5.37}$$

Within this limit lies, in fact, the formal definition of the base e of the natural logarithm:

$$e \equiv \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x. \tag{5.38}$$

Find? We thus see that our unit investment, continuously compounded, attains in one year a value only slightly larger than the daily compounding shown in the last line of Table 5.3:

$$\left(\frac{N}{N_0}\right)_\infty = e^{0.10} \cong 1.1051709. \tag{5.39}$$

Use? It is worth noting that, for economists, interest represents the price of money. What does that mean? Putting money into a savings account means giving up an opportunity to buy something *now* in exchange for the promise of being able to spend a larger amount of money—the initial investment plus earned interest income—*at a future date.* This means trading the opportunity to spend \$1.00 now for the opportunity to spend \$1.10 a year from now. The bank has "purchased" money for its own investment purposes at a price of \$0.10 for the year, and the saver bought the chance to spend still more money, \$1.10, one year later. This means that money has both a price and, again, a *time value* because investors make decisions about what their money will be worth in the future. This brings us to a second money issue, inflation, in which exponential behavior significantly affects the price of money.

Problem 5.19. Are eq. (5.35) and (5.38) related? How?

Problem 5.20. Verify all of the steps that lead from eq. (5.35) to eq. (5.39).

Problem 5.21. Construct a version of Table 5.3 for an annual interest rate of 18%.

5.5.2 Inflation

Why? Inflation has been a major economic and political problem in the United States at various times in the 20th century, and it has troubled and even destabilized the economies of many other countries in just the last few

years. Asian economies suffered major bouts of inflation in the late 1990s, and at the very end of 2001 Argentina had street riots and five (!) presidents in less than two weeks because of economic problems triggered in part by serious inflation. Inflation occurs when the value of money declines and the prices of goods and services rise accordingly. Countries suffering from bouts of inflation see the value of their currencies drop against those of other countries, and the consequences of such economic imbalances may include unemployment, trade embargoes and trade wars, and severe, spreading economic dislocation. These topics are the province of economics, "the dismal science," but they are interesting to us because inflation is an exponential phenomenon and the mathematics of inflation is provocative.

Find? Consider first simple price inflation as measured by the purchase price of gasoline. Gasoline cost a nickel a gallon in 1933, while at the end of 2001 it cost \$1.00 per gallon. We can calculate the annual price inflation rate for gasoline with eq. (5.5):

$$\lambda_{price} = \frac{\ln(1.00/0.05)}{68\,\text{yr}} \cong 0.0440/\text{yr}, \tag{5.40}$$

Predict? which corresponds to a price inflation rate of 4.40% per year. This price inflation rate caused gasoline's price at the pump to go up by a factor of 20 in 68 years.

Valid? As appealing as this simple calculation may be, it would be quite misleading to say that the *real* price of gasoline went up twentyfold during the time 1933–2001 because, while the *nominal* or apparent price of gasoline was going up, the value of the dollar itself was going down. That is, inflation **Improve?** affects not only the price of goods and services; it also affects the price of money. During the 68 years included in the previous calculation, the value of the dollar declined substantially, because a 1933-dollar and a 2001-dollar are only *nominally* the same. If we assume that the dollar was losing its purchasing power at only 2% per year, we could calculate the value of a single dollar after t years, $v(t)$, from eq. (5.14) with $\lambda_\$ = 0.02/\text{yr}$:

$$\frac{v(t)}{v(1933)} = e^{-\lambda_\$(t-1933)} = e^{-0.02(t-1933)}. \tag{5.41}$$

Thus, after 1, 10, and 68 years, the purchasing power or value of a 1933-dollar would be \$0.98, \$0.82, and \$0.26, respectively. So, after almost 70 years, the 2001-dollar has turned out to be worth little more than one-quarter of the 1933-dollar!

Use? However, an economist would view this differently. Recall from Section 5.5.1 that we noted that interest is the price of money bought in a forward-looking transaction.Thus, we can rephrase the question about the loss of value in the dollar into a purchasing question: How much would

one have to pay in 1933 to have \$1.00 available in 2001? That is a price question answered simply by inverting eq. (5.41):

$$\frac{v(1933)}{v(t)} = e^{+\lambda_{\$}(t-1933)} = e^{+0.02(t-1933)}. \tag{5.42}$$

So, repeating the calculation just done in this different form, a purchaser would have to invest \$0.98, \$0.82, and \$0.26, respectively, in order to have \$1.00 available to spend in the years 1934, 1943, and 2001. Equation (5.42) thus can be said to represent the currency inflation rate.

Predict? Purchases can then be assessed either in terms of their current sales prices or in terms of *inflation-adjusted dollars* that support the calculation of a real economic price that reflects changes in a currency's purchasing power. We would calculate that *real* price by subtracting the currency inflation rate from the price inflation rate, that is,

$$\lambda_{real} = \lambda_{price} - \lambda_{\$}. \tag{5.43}$$

Equation (5.43) then states that the real inflation rate over the time interval 1933–2001 is then, from the example data given above, $\lambda_{real} = 4.40 - 2.00 = 2.40\%$.

We do not mean to suggest that inflation is an easy problem because it can be modeled with exponential mathematics. The foregoing analyses have truly simplified the world of economics. Economics has become in recent times a mathematically-oriented social science, as evidenced in part by the sophisticated mathematical models that led to the prizes won by most recent Nobel laureates. However, we do want to point out that the

Improve? cumulative effects of percentages in economics can be enormous. We have ignored some measures that have been developed to deal with inflation, such as *indexing*, in which intended benefits are linked to a cost or price index, such as the oft-cited CPI, the *consumer price index*. We have also completely ignored the effects of technical innovation, productivity changes, new sources of energy, and many other factors that affect prices. Suffice it to say that the economics and politics of exponential growth in monetary affairs merit attention.

Problem 5.22. If gasoline cost \$0.70/gallon in 1978, calculate and compare the price inflation rates for the intervals 1933–1978 and 1978–2001.

Problem 5.23. If the cost of money exceeds the cost of goods, what happens to λ_{real}?

Problem 5.24. Speculate on the potential effects of λ_{real} staying negative for long periods of time.

5.6 A Nonlinear Model of Population Growth

In Sections 5.1 and 5.2 we discussed population growth and projections based on an elementary exponential model in which the population growth rate is *linearly* or directly proportional to the current size of the population. While we focused exclusively on growth rates, we could extend such linear models to account for mortality or death rates simply by taking the growth rate in eq. (5.1) as an effective or *net* rate that reflects the difference between birth and death rates: **Improve?**

$$\lambda_{effective} = \lambda_{birth} - \lambda_{death}. \tag{5.44}$$

In fact, we could also account for immigration and emigration in the analysis of the population changes of a particular country by writing a balance law much like eqs. (1.1) and (1.2) and accounting for the various growth and decay rates as: **How?**

$$\frac{dN(t)}{dt} = (\lambda_{birth} - \lambda_{death} + \lambda_{immigration} - \lambda_{emigration})N(t). \tag{5.45}$$

However, it is certain that these models either grow or decay monotonically, as simple exponentials, no matter how much we refine the details of these linear growth and decay rates. The fundamental behavior is unchanged, so that if we find the classic model inadequate, we need to change that model in a different way.

We would like to expand the notion of exponential growth to incorporate the idea of *limited* growth. There are many factors that do limit growth and that modelers have tried to incorporate into population projections, including resources, both renewable and nonrenewable, energy, capital (money), food supplies and distribution mechanisms, education, and family planning. These models were very popular in the late 1970s, but they were also both complicated and, by some, derided as unrealistic. Much of the complexity of those models stemmed from the fact that many of the growth variables are coupled, that is, the amount of capital formulation may depend on pollution indices and on energy availability, as well as on the instantaneous supply of money. Further, the right-hand side of eq. (5.1) may be more complex because the relationships among single or coupled variables may not be linear. How could that be? **Why?**

It could be more complex if we were to think of the right-hand side of eq. (5.1) as a Taylor series of a nonlinear function of $N(t)$ that is not yet defined. Thus, we would start by replacing eq. (5.1) by a more general formulation **Assume?**

$$\frac{dN(t)}{dt} = f(N(t)), \tag{5.46}$$

which states that the rate of growth of a population $N(t)$ is equal to some
How? undefined function of the population, $f(N(t))$. Then, as we did for series representations of functions in Chapter 4, we could expand that function into a Taylor series such that:

$$\frac{dN(t)}{dt} = f(N(t)) = C_0 + C_1 N(t) + C_2 N^2(t) + \cdots . \tag{5.47}$$

We would have to say first that $C_0 = 0$ simply because the growth rate of a population should be zero whenever the population size is zero. The constant C_1 must be our traditional growth rate, say λ_1. Then there are other constants, C_i, to evaluate, depending on how many terms we choose to keep in this series representation of $f(N(t))$. How do we evaluate these other constants?

Predict? We illustrate that by narrowing our focus to a particular quadratic approximation in which eq. (5.47) takes the form:

$$\frac{dN(t)}{dt} = \lambda_1 N(t) - \lambda_2 N^2(t), \tag{5.48}$$

wherein both of the parameters λ_1 and λ_2 are taken as positive: $\lambda_1, \lambda_2 > 0$. In eq. (5.48) λ_1 corresponds to the population's uninhibited or net growth rate. The meaning of λ_2 emerges from noting that the rate of growth vanishes when $N(t) = N_{\max}$:

$$\lambda_1 N_{\max} - \lambda_2 N_{\max}^2 = 0,$$

or when

$$\frac{1}{\lambda_2} = \frac{N_{\max}}{\lambda_1}. \tag{5.49}$$

Thus, the reciprocal of λ_2 is the time needed for the maximum obtainable population to be achieved by uninhibited growth. On the other hand, with the aid of eq. (5.49), we can eliminate λ_2 from that the nonlinear equation and write it as:

$$\frac{dN(t)}{dt} = \lambda_1 N(t) \left(1 - \frac{N(t)}{N_{\max}}\right). \tag{5.50}$$

Equation (5.50) shows a modification of the elementary exponential model where the growth rate is reduced by a factor representing the proportion of *unrealized population growth*, that is, the population represented by the difference between the maximum and instantaneous population values:

$$\frac{dN(t)}{dt} = \lambda_1 N(t) \left(\frac{N_{\max} - N(t)}{N_{\max}}\right). \tag{5.51}$$

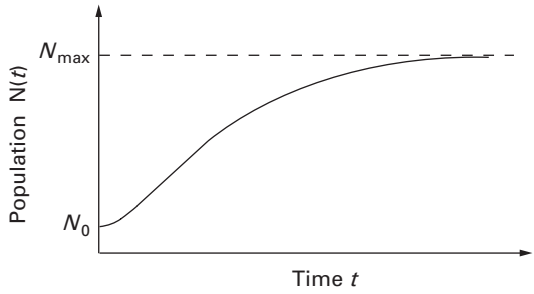


Figure 5.9 The *logistic growth curve*, shown as a model of limited or bounded growth as the population $N(t)$ moves from an initial population value, N_0, to its maximum value, N_{max}. It is plotted for the values $N_0 = 1$, $N_{max} = 10$, and $\lambda_1 = 1$.

There is a closed-form solution to eq. (5.51), despite the nonlinearity, and it is (see Problem 5.28):

$$N(t) = \frac{N_{max}}{1 + \left(\dfrac{N_{max} - N_0}{N_0}\right) e^{-\lambda_1 t}}, \tag{5.52}$$

where $N(t = 0) = N_0$ is the initial population. We have plotted eq. (5.52), known as the *logistic growth curve*, in Figure 5.9. Note that we can recover both the initial value of the population at $t = 0$, as well as the maximum value as time becomes indefinitely large.

We should observe again that we have not exhausted by any means the spectrum of exponential growth models. Nevertheless, we have shown here that models can lead to restricted or limited growth, which should provide some interest in exploring different exponential models in greater detail.

Problem 5.25. Look up the U.S. birth, death, immigration and emigration figures for the 10 decades of the 20th century and use the balance equation (5.45) to calculate the population changes that these rates predict.

Problem 5.26. How do the predictions of Problem 5.25 compare with the actual U.S. census data?

Problem 5.27. What are the implications for the model of eq. (5.48) of loosening the restriction that $\lambda_1, \lambda_2 > 0$?

Problem 5.28. Confirm by differentiating eq. (5.52) that it satisfies eq. (5.51).

5.7 A Coupled Model of Fighting Armies

We will now examine another exponential model wherein the complication of coupled populations is addressed. This model and the resulting *Lanchester's law* are named after Frederick William Lanchester (1868–1946), a remarkable British aeronautical engineer who wrote serious works on economics and fiscal policy, the theory of relativity, military strategy, as well as aerodynamics. Lanchester wanted to describe the attrition of opposing forces at war. Following this attrition required modeling the changes of two army populations whose respective rates of attrition depend on the size of the opposing army. Thus, there are two armies whose attrition or decay rates are of interest, each of whose decay rates are proportional to the size of the other force. We will identify the two army populations as friendly forces, $F(t)$, and enemy forces, $E(t)$. Since the rate of change of $F(t)$ depends on $E(t)$ and vice versa, we say that these variables are *coupled*, or that we are solving a *coupled problem*. This model also has the nice feature, encapsulated in Lanchester's law, that we can obtain a great deal of information with a *qualitative* approach to the governing differential equations. We will use qualitative analyses to describe energy conservation and dissipation for a vibrating pendulum in Sections 7.1.5 and 7.1.6 and for the interaction of predators and prey in Section 7.6.

Why? **Find?** **Given?**

Consider that at some time, t, we have populations $F(t)$ of friendly troops and $E(t)$ of enemy troops. Further, as we intended, let us assume that the rates of change of their respective populations are proportional to the opposing combat force's size:

Assume?

$$\begin{aligned} \frac{dF(t)}{dt} &= -\lambda_E E(t), \\ \frac{dE(t)}{dt} &= -\lambda_F F(t). \end{aligned} \tag{5.53}$$

The parameters, λ_E and λ_F, respectively, represent the effectiveness of the enemy and friendly forces, with interesting units: λ_E is the loss rate per unit time of friendly troops *per enemy troop*. Thus, if λ_E is larger than λ_F, the enemy troops are more effective because more friendly troops are lost per unit time *and* per unit of enemy forces.

How?

Equation (5.53) also shows more explicitly the meaning of coupling in a set of equations. Simply put, dF/dt depends on $E(t)$, and dE/dt depends on $F(t)$. That is why the pair of eqs. (5.53) are called *coupled equations*. It can be shown that this coupled pair of first-order equations is equivalent to a single second-order equation by, for example, simply treating the first of eq. (5.53) as an equation that defines $E(t)$, and then substituting it into

Use?

the second of eq. (5.53):

$$\frac{dE(t)}{dt} = -\frac{1}{\lambda_E}\frac{dF^2(t)}{dt^2} = -\lambda_F F(t), \tag{5.54a}$$

which is easily rearranged into the form:

$$\frac{dF^2(t)}{dt^2} - \lambda_E\lambda_F F(t) = 0. \tag{5.54b}$$

Once this uncoupled second-order equation is solved for its single dependent variable, $F(t)$, the second dependent variable, $E(t)$, can be found without further integration (see Problem 5.29).

The formal solution to eqs. (5.53) or (5.54) can be found in terms of hyperbolic sines and cosines, which are also exponential functions. We will not do that here, although the form of eq. (5.54b) should be recalled when we discuss the vibration of pendulums in Chapter 7. Instead, we will show here how we can obtain a lot of information without formally solving the governing equations. We do this by first multiplying the first of eq. (5.53) by $\lambda_F F(t)$ and the second of eq. (5.53) by $\lambda_E E(t)$, after which we find:

How?
Predict?

$$\begin{aligned} \lambda_F F(t)\frac{dF(t)}{dt} &= -\lambda_E\lambda_F F(t)E(t), \\ \lambda_E E(t)\frac{dE(t)}{dt} &= -\lambda_F\lambda_E E(t)F(t). \end{aligned} \tag{5.55}$$

Since the right-hand sides of eq. (5.55) are the same, it follows that:

$$\lambda_F F(t)\frac{dF(t)}{dt} = \lambda_E E(t)\frac{dE(t)}{dt}. \tag{5.56}$$

It is easy to show that eq. (5.56) is equivalent to the statement that:

$$\frac{d}{dt}(\lambda_F F^2(t) - \lambda_E E^2(t)) = 0,$$

or

$$\lambda_F F^2(t) - \lambda_E E^2(t) = \text{constant}. \tag{5.57}$$

The constant in eq. (5.57) must have the same value it had at the beginning of the combat being modeled. With $E_0 = E(t = 0)$ and $F_0 = F(t = 0)$, it follows that:

$$\lambda_F F^2(t) - \lambda_E E^2(t) = \lambda_F F_0^2(t) - \lambda_E E_0^2(t),$$

or

$$\lambda_F(F^2(t) - F_0^2(t)) = \lambda_E(E^2(t) - E_0^2(t)). \tag{5.58}$$

Use? Equation (5.58) is called *Lanchester's square law*. We can use the square law to calculate the final size of the winning army when the enemy forces have been annihilated *without solving any differential equations*. Assume that victory is declared when all of the enemy forces are gone from the scene. In this case, $E_{final} = 0$. The number of friendly troops remaining then follows from eq. (5.58) as:

$$F_{final}^2(t) = F_0^2(t) - \frac{\lambda_E}{\lambda_F} E_0^2(t). \tag{5.59}$$

Thus, even in victory the number of (surviving) friendly troops is reduced by an amount proportional to the square of the initial size of the enemy force.

Valid? The dependence of the friendly and enemy force sizes on the respective squares has intriguing consequences. Suppose that two equally effective armies oppose each other. This means $\lambda_E = \lambda_F$, and that Lanchester's law [eq. (5.58)] becomes:

$$F^2(t) - E^2(t) = F_0^2(t) - E_0^2(t). \tag{5.60}$$

Suppose further that two combat scenarios were being considered by military planners. In the first scenario, a friendly army of 50,000 soldiers faces an enemy force of 40,000 and then meets a second enemy force of 30,000 soldiers. In the second scenario, the same friendly army meets an enemy force of 70,000, that is, it meets the same number of enemy troops assembled for a single fight. In the sequential scenario, the friendly army prevails in the first of its two battles with a surviving forces of 30,000 because $[(50{,}000)^2 - (40{,}000)^2] = (30{,}000)^2$, which is just enough to force a draw with the enemy in the second battle. If the armies meet in the second scenario, however, the friendly forces lose by a significant margin because $(50{,}000)^2$ is less than $(70{,}000)^2$. This clearly shows that strategy is important, especially that well-known precept of *divide and conquer*!

Improve? **Verified?** Of course, all of the Lanchester results are predicated on the rate equations (5.53), an assumption that must be kept in mind when the model is exercised. Suitably modified to include other effects (e.g., introducing reinforcements), the Lanchester model has modeled the outcomes of famous battles such as Iwo Jima (see Problems 5.47 and 5.48).

Problem 5.29. Assuming that eq. (5.54b) can be solved for $F(t)$, show that $E(t)$ can be determined without further integration.

Problem 5.30. Confirm that eq. (5.57) does follow from eq. (5.55) via eq. (5.56).

Problem 5.31. Suppose you are given the solution for the enemy population that satisfies eq. (5.53) as

$$E(t) = E_0 \cosh \alpha t - \sqrt{\frac{\lambda_F}{\lambda_E}} F_0 \sinh \alpha t,$$

where $\alpha^2 = \lambda_E \lambda_F$. How much time does it take for the enemy forces to be completely annihilated?

Problem 5.32. Would the strategy of divide and conquer work for a "linear attrition law" that for equally effective armies replaces eq. (5.57) with

$$F(t) - E(t) = F_0(t) - E_0(t)?$$

5.8 Summary

We dealt with a wide variety of exponential behavior models in this chapter, including population growth, radioactive decay, charging and discharging of capacitors, inflation and interest, and armies at war. While some of the behavior was about decay, it is the cases of exponential growth that really draw our attention. We saw the importance of scale in presenting and assessing various growth phenomena. We noted that decay effects can be modified by external inputs, such as the charge in a capacitor responding to an applied voltage. We also explored the nonlinear logistic growth model and the coupled Lanchester square law.

It is worth noting that we have touched on some very timely issues. At the same time, we have not "solved" any of these very real "problems." But we have shown that the models chosen can influence our projections and perceptions of these problems, as well as the ways we might approach them in the "real world".

5.9 References

S. E. Ambrose, *Nothing Like It in the World: The Men Who Built the Transcontinental Railroad 1863–1869*, Simon & Schuster, New York, 2000.

A. A. Bartlett, "The Exponential Function, Parts I–VIII," *The Physics Teacher*: I, October 1976 (p. 393); II, November 1976 (p. 485); III,

January 1977 (p. 37); IV, February 1977 (p. 98); V, April 1977 (p. 225); VI, January 1978 (p. 23); VII, February 1978 (p. 92); VIII, March 1978 (p. 158).

M. Braun, *Differential Equations and Their Applications: Shorter Version*, Springer-Verlag, New York, 1978.

P. D. Cha, J. J. Rosenberg, and C. L. Dym, *Fundamentals of Modeling and Analyzing Engineering Systems*, Cambridge University Press, New York, 2000.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

J. H. Engel, "A Verification of Lanchester's Law," *Operations Research*, **2**(2), 163–171, 1954.

J. W. Forrester, *World Dynamics*, Wright-Allen, Cambridge, MA, 1971.

F. W. Lanchester, *Aircraft in Warfare; The Dawn of the Fourth Arm*, Constable, London, 1916. See also J. R. Newman (Ed.), *The World of Mathematics*, Simon & Schuster, New York, 1956.

R. E. Lapp and H. L. Andrews, *Nuclear Radiation Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1954.

R. B. Lindsay and H. Margenau, *Foundations of Physics*, Dover Publications, New York, 1957.

D. H. Meadows, D. L. Meadows, J. Randers, and W. W. Behrens III, *The Limits to Growth*, Universe Books, New York, 1972.

P. A. Samuelson, *Economics*, 17th Edition, McGraw-Hill, New York, 2001.

P. A. Samuelson, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, MA, 1947.

J. E. Stiglitz, *Economics*, 2nd Edition, W. W. Norton, New York, 1993.

E. R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT, 1983.

5.10 Problems

5.33. Show that if it takes time, t_c, to count a population, $P(t)$, that has a growth rate of λ, the population will increase by an amount equal to $\lambda t_c P(t)$.

5.34. (a) If the population counting rate is c, how long does it take to count the population at time, t?

(b) How much time does it take to count the increase in population that occurred while it was being counted at time t?

5.35. Find the actual world population figures for 1970, 1980, 1990, and 2000. Use these data to update the projections shown in Figures 5.1 and 5.2.

5.36. Using the 2% growth rate, plot the world population from 1960 to 2060 with an ordinate scale of 10 billion people per 1.50 in. Does this curve look like "reasonable" exponential growth?

5.37. The ordinate scales of Figures 5.1 and 5.2 are, respectively, 1.5 in = 100 billion people and 1.5 in = 3 million billion people. How much paper is needed to plot the 1960 world population of 3 billion people and the projected 2692 world population of 5.63×10^{15} people using each of those scales?

5.38. Plot the growth of world population from a 1960 value of 3 billion people at growth rates of 1, 2, and 3% per year through 2700 using semi-logarithmic paper. What shapes are these curves? What are their slopes and intercepts?

5.39. What is the time constant, comparable to that for an RC circuit, for a population decaying at a rate per unit time λ?

5.40. How much should be set aside in 2002 in a savings account earning 5.5% per year to accumulate \$1,000,000 by 2022? By 2042?

5.41. Suppose that there was a steady inflation rate of 3% per year, what would the investments of Problem 5.40 have to be to accumulate \$1,000,000 in 2042 measured in 2002 dollars?

5.42. The noted (and recently deceased) historian Stephen Ambrose has chronicled the growth of American railroads by listing the following amounts of total track by decade: 726 mi (1834), 4311 (1844), 15,675 (1854), and 33,860 (1864). Determine:
(a) the decade-by-decade growth rate; and
(b) the growth rate for exponential growth across all the data given.

5.43. Verify by differentiation and substitution that the following solution satisfies eqs. (5.53):

$$F(t) = F_0 \cosh \alpha t - \sqrt{\frac{\lambda_E}{\lambda_F}} E_0 \sinh \alpha t,$$

$$E(t) = E_0 \cosh \alpha t - \sqrt{\frac{\lambda_F}{\lambda_E}} F_0 \sinh \alpha t.$$

5.44. Confirm that the solution verified in Problem 5.43 satisfies Lanchester's square law of eq. (5.58).

5.45. The initial strengths of two opposing armies are $F_0 = 10{,}000$ and $E_0 = 5000$ troops, with equal loss rates of 0.1 per day. Who will win? How long will the battle take? (*Hint*: See Problem 5.31.) How many troops will the victor have when the enemy is vanquished? Graph the army populations until the enemy is completely annihilated.

5.46. The initial strengths of two opposing armies are $F_0 = 10{,}000$ and $E_0 = 5000$ troops, and $\lambda_F = 0.1$ per day. Who will win and with

what remaining forces if $\lambda_E = 0.2$, 0.5, and 1.0 per day? What value of λ_E would produce a draw?

5.47. The landmark World War II battle of Iwo Jima began with troop sizes of $F_0 = 54{,}000$ and $E_0 = 21{,}500$ troops, with $\lambda_F = 0.0106$ per day and $\lambda_E = 0.0544$ per day. Absent any reinforcements, how long would this battle have lasted? How many troops would the victor have when the loser's forces were totally exhausted?

5.48. In order to end the fight for Iwo Jima in 28 days, how many troops would the United States have had to have initially? How do the U.S. losses in this scenario compare to those found in the scenario of Problem 5.47?

6
Traffic Flow Models

People like to drive, especially in the United States. In fact, we can often tell where people come from by how they refer to highways: people on America's east coast talk about taking *the turnpike* (or *the 'pike*) or *the interstate*, while on the west coast we get on *the freeway* or we take *the 5* or *the 101*, referring to a particular highway by its number. In order to design the roads and the cars that enable and facilitate such personal transportation, we model both the behavior of *individual* cars with their drivers in a (single) line of autos, and that of *groups* of cars in one or more lanes of traffic. However, our concern is not with modeling the ergonomics of operating a car. Rather, we focus on the interactions of autos on single highway lanes, both individually and in dense lines.

6.1 Can We Really Make Sense of Freeway Traffic?

No matter how we refer to traffic arteries, the flow of traffic on them is modeled, analyzed, and predicted with *traffic flow theory*, which we now detail at two levels. The *macroscopic modeling* of traffic assumes a sufficiently large number of cars in a lane or on a road such that each stream of autos can be treated as we would treat fluid flowing in a tube or stream. Thus, to maintain the biological metaphor, traffic flow is treated as a flow of a fluid *field* in an artery. Macroscopic models are expressed in terms of

three gross or *average* variables for a whole line of traffic: the number of cars passing a fixed point per unit of time, called the *rate of flow*; the distance covered per unit time, the *speed of the traffic flow*; and the number of cars in a traffic line or column of given length, which we identify as the *traffic density*. The relationship between the speed and the density is embodied by macroscopic modelers in a plot of these two variables called the *fundamental diagram*. We also invoke the *continuum hypothesis* (viz. Section 4.7.2) to confirm that it is appropriate to (mathematically) treat the traffic as a field.

The second level of traffic modeling, *microscopic modeling*, addresses the interaction of individual cars in a line of traffic. Microscopic models describe how an individual *follower* car *responds* to an individual *leader* car by modeling its acceleration as a function of various perceived *stimuli*, which might be the distance between the leader and follower cars, the relative speeds of the two cars, or the reaction time of the operator of the follower car. Car-following models come in several varieties, and they can be used to construct the speed-density curves that are the underpinning of macroscopic modeling. Such speed-density plots, supported by data taken from real traffic arteries, enable traffic experts to model and understand road or freeway capacity as a function of traffic speed and density—even if everyday drivers feel they do not fully "understand" what is happening around them. (The microscopic models are also used to support the modeling of vehicular *control*, that is, to implement control strategies that enable lines of traffic to maintain high flow rates at high speeds. However, we will not delve into control theory and its applications here.)

We will start our brief overview of traffic modeling at the macroscopic level, applying conservation principles for cars aggregated into a field (or sufficiently large collection of cars) to define the fundamental diagram for the flow of traffic on a highway populated with multiple vehicles. Then we will examine how the continuum hypothesis influences our view of individual cars (and drivers), as a guide to developing car-follower models that model the interaction between a single car as its driver reacts to another auto immediately ahead. These car-follower models are then used to derive the speed-density relationships that allow us to put specific models and numbers into the more general macroscopic traffic flow theory.

6.2 Macroscopic Traffic Flow Models

Why? We start by asserting the validity of an analogy, namely, that the flow of a stream of cars can be modeled as a field, much as we would model the flow of a fluid. Thus, the collection of cars taking the 10 east out of Los Angeles on any given evening is mathematically similar to the flow of blood in an

artery or water in a home piping system. We want to relate the speed of a line of traffic to the amount of traffic in that line (or lane). We use three variables to describe such traffic flows: **Find?**

- the rate of flow, $q(x, t)$, measured in the number of cars per unit time;
- the density of the flow, $\rho(x, t)$, which is the number of vehicles per unit length of road; and
- the speed of the flow, $v(x, t)$.

How are these three variables related?

6.2.1 Conservation of Cars

We can provide one answer to the foregoing question by applying the conservation principle embodied in eqs. (1.1) and (1.2) to traffic moving (in one direction) along an arbitrary stretch of a road. The conservation principle states that the change in the number of cars within that stretch of road results from the flow of traffic into and out of that road interval, and from the generation or consumption of cars within the interval. Notwithstanding the occasional pictures we have all seen of horrific mega-accidents that occur during severe fogs or major storms, we will (safely) assume that cars are neither generated nor consumed within that road interval. **How?** **Assume?**

Thus, imagine a coordinate, x, along a particular stretch or interval of road under consideration that has endpoints defined by $x = x$ and $x = x + \Delta x$. The number of cars within this road interval of length Δx is given by $\Delta N(x, t)$. Given our assumption that we will neither generate or consume cars, the conservation principle of eq. (1.2) states that the change in the number of cars within the interval $\Delta N(x, t)$ during a time interval Δt is, in the limit, equal to the *rate of traffic flow*, $q(x, t)$:

$$q(x, t) \equiv \lim_{\Delta t \to 0} \frac{\Delta N(x, t)}{\Delta t}. \tag{6.1}$$

The change in the number of cars within the road interval, $\Delta N(x, t)$, is simply the difference between the number of cars going in and out of that stretch of road at each end, $N(x, t)$ and $N(x + \Delta x, t)$, respectively:

$$\Delta N(x, t) = N(x, t) - N(x + \Delta x, t), \tag{6.2}$$

If Δx denotes the length of road interval that is traveled during the time, Δt, the statement of conservation of cars (6.1) can also be written as

$$q(x, t) = \lim_{\Delta t \to 0} \frac{\Delta N(x, t)}{\Delta x} \left(\frac{\Delta x}{\Delta t} \right), \tag{6.3}$$

where the fraction introduced in eq. (6.3) is the speed of the traffic, $v(x, t)$, in the interval:

$$v(x, t) = \left(\frac{\Delta x}{\Delta t} \right). \tag{6.4}$$

Equations (6.2) and (6.4) are now substituted into the conservation of cars (6.3) to yield

$$q(x, t) = \left(\lim_{\Delta x \to 0} \frac{N(x, t) - N(x + \Delta x, t)}{\Delta x} \right) v(x, t). \tag{6.5}$$

Note that the limit in eq. (6.5) is now taken as $\Delta x \to 0$, and that its dimensions correspond to the number of vehicles per unit length of road, which we define as the *density of the traffic flow*:

$$\rho(x, t) \equiv \lim_{\Delta x \to 0} \frac{N(x, t) - N(x + \Delta x, t)}{\Delta x}. \tag{6.6}$$

Thus, eq. (6.5) can be rewritten for the last time to cast the *principle of conservation of cars* in the form

$$q(x, t) = \rho(x, t)\, v(x, t). \tag{6.7}$$

Beyond preserving the notion that "what goes in must go out," what does eq. (6.7) mean? First, we note that the equation is dimensionally consistent and correct (see Problem 6.1). Second, we note that eq. (6.7) can be shown to make "physical" sense by a rather simple argument derived by looking at two different ways of counting the number of cars passing a (specified) point on the road during a very small time interval.

One measure of the traffic count is that the number of cars, ΔN, passing a point during a time interval, Δt, is simply the product of the flow rate, q, and the time interval: $\Delta N = q\Delta t$. The second measure count assumes that during the same small interval of time a car moving with a speed, v, will cover a distance, $\Delta x = v\Delta t$. The number of vehicles passing through that distance is found from another simple product: of density, ρ, times distance: $\Delta N = \rho\Delta x$. Hence, equating the two measures of the number of cars passing a point yields the result

$$q\Delta t = \rho\Delta x, \tag{6.8}$$

which is clearly an averaged version of eq. (6.7) that accords well with this elementary physical reasoning (see Problem 6.2).

We also observe that the single equation (6.7) is expressed in three variables: q, ρ, and v. Therefore, it is of very limited use in this form without substantial further information. However, it is clear that traffic density, ρ, and speed, v, are the *two fundamental traffic variables* because we can determine the rate, q, at which traffic flows by inserting them into eq. (6.7).

Further, if we could relate speed directly to density, i.e., $v = v(\rho)$, then we could write a direct relationship between the traffic flow rate, q, and the density, ρ:

$$q(\rho) = \rho\, v(\rho). \tag{6.9}$$

As we will see in Section 6.2.3, plots of traffic flow rate, q, against density, ρ, are so widely used in modeling traffic flow that they are identified under the rubric of the *fundamental diagram of road traffic*.

Speed-density relationships (e.g., $v = v(\rho)$) are clearly central to our understanding of traffic flow, so we turn to them next.

Problem 6.1. Confirm that eq. (6.7) is dimensionally correct.

Problem 6.2. Explain which variables were averaged, and how, over the intervals of distance (Δx) and time (Δt) in the heuristic derivation of eq. (6.8)?

6.2.2 Relating Traffic Speed to Traffic Density

Even inexperienced drivers would agree that traffic speed and traffic density are related. Drivers speed up when traffic is sparse, and they slow down (perhaps involuntarily!) to clog up arteries when traffic is thick. Thus, we are tempted to postulate that there is a direct relationship between traffic speed and traffic density: **Assume?**

$$v = v(\rho). \tag{6.10}$$

Let us now reason a bit further about this relationship to determine any conditions that need to be applied to any particular functional form, $v(\rho)$, that might be proposed.

Building on the intuition just mentioned, we expect that a driver will drive fastest, $v_{\max}$, when the density is at its smallest value, $\rho \to 0$. The speed decreases as the density increases, which is a statement about the slope of the v versus ρ curve. Finally, traffic grinds to a halt, $v = 0$, at some maximum or *jam* density, ρ_{jam}, presumably when the traffic is bumper-to-bumper. We can summarize these experience-born intuitions in mathematical requirements on the function, $v(\rho)$: **Assume?** **Assume?**

$$v(\rho = 0) = v_{\max}, \tag{6.11a}$$

$$\frac{dv}{d\rho} \leq 0, \tag{6.11b}$$

$$v(\rho = \rho_{\text{jam}}) = 0. \tag{6.11c}$$

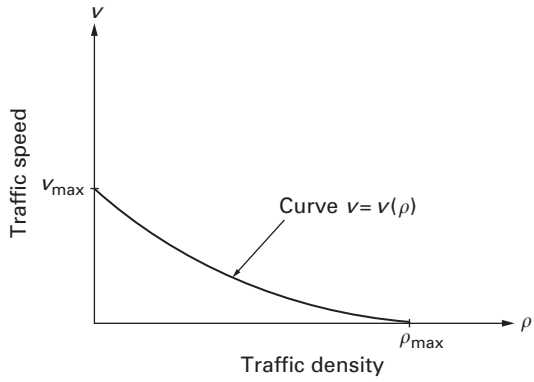


Figure 6.1 A generic schematic of the variation of traffic velocity with density. It displays the endpoints, [(0, v_{max}) and (ρ_{max}, 0), respectively], and shows that the slope is always non-positive, $dv/d\rho < 0$, which results from our experience that traffic speed drops off as traffic density increases.

We can also display these results graphically, in the generic curve shown in Figure 6.1. Note that the precise shape of the curve is unknown; only the endpoint values and the sign of the slope are specified at this point.

The elementary modeling assumptions just outlined do not exhaust all of the possibilities, although experience suggests that eqs. (6.10) and (6.11) adequately reflect the behavior of traffic that is accelerating or decelerating. Models behind traffic speed-density relations will reflect human behavior—rather than mechanical laws—because they reflect how drivers respond to stimuli. That is, drivers can respond to perceived distances between cars, to relative speeds, to the perceived density further down the road, and so on. In fact, speed-density relations such as eq. (6.10) are found both from empirical data and from the very stuff of the modeling of car-following interactions that we address in Section 6.3.

6.2.3 Relating Traffic Flow to Traffic Density: The Fundamental Diagram

Why? From the viewpoint of the traffic engineer who is designing a road and all of its facilities (including entrance and exit ramps, traffic signs and signals, toll booths, etc.), the most relevant variable is the *capacity* (or maximum flow rate) that the road system must accommodate, as reflected in its traffic
Given? flow rate, $q(x, t)$. For macroscopic models we can take the speed to be

homogeneous, which means that it does not explicitly depend on the road coordinate, x, or on time, t. Then, we can write $v = v(\rho)$, anticipating as in eq. (6.9), that traffic flow ultimately depends only on the density, ρ.

We can now extend our qualitative analysis of the speed-density relationship (of Section 6.2.2) to the relationship between the traffic flow rate and the density. Thus, because a driver's fastest speed, $v_{\max}$, occurs when the density is at its smallest, $\rho = 0$, eq. (6.9) tells us that $q(\rho = 0) = 0$, that is, that the flow rate is zero. Similarly, when traffic slows to a halt at its maximum density, $v(\rho_{\text{jam}}) = 0$, eq. (6.9) tells us once again that the traffic flow rate is zero: $q(\rho_{\text{jam}}) = \rho_{\text{jam}} v(\rho_{\text{jam}}) = 0$. The traffic flow rate must be positive for all values of the density ($0 < \rho < \rho_{\text{jam}}$), and must attain its maximum value $q_{\max}$ somewhere in that interval. Further, the slope of the traffic flow rate is given by (see Problem 6.3):

$$\frac{dq}{d\rho} = v(\rho) + \rho \frac{dv}{d\rho}. \tag{6.12}$$

The qualitative results just found are embodied in the generic curve shown in Figure 6.2, which is called the *fundamental diagram of traffic flow*. As with Figure 6.1, the precise shape of the curve is unknown: the endpoint values are specified and the variation of the slope can be inferred (see Problem 6.4).

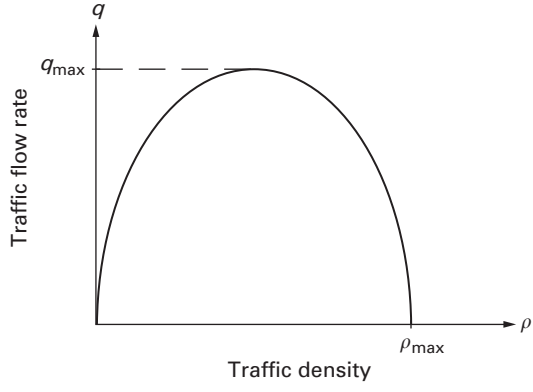


Figure 6.2 A generic schematic of the variation of the traffic flow rate with density. It displays the endpoints, [(0, 0) and ($\rho_{\max}$, 0), respectively], and shows that the slope is positive until the maximum flow rate or *capacity*, $q_{\max}$, is reached, and negative thereafter.

To make some of these qualitative ideas more specific, consider the following linear speed-density relationship:

$$v(\rho) = v_{\max}\left(1 - \frac{\rho}{\rho_{\text{jam}}}\right). \tag{6.13}$$

This relationship clearly satisfies (see Problem 6.5) all of the conditions required by eqs. (6.11a–c). Moreover, as the simplest (linear) mathematical expression that satisfies these conditions, it is particularly attractive as a "building block" for further modeling, provided that it adequately models reality. When substituted into eq. (6.9), it produces a relationship for the traffic flow rate as a function of density that is *parabolic*:

$$q(\rho) = v_{\max}\left(\rho - \frac{\rho^2}{\rho_{\text{jam}}}\right). \tag{6.14}$$

The maximum flow rate occurs when its slope vanishes:

$$\frac{dq(\rho)}{d\rho} = v_{\max}\left(1 - \frac{2\rho}{\rho_{\text{jam}}}\right) = 0. \tag{6.15}$$

Equation (6.15) shows that the maximum traffic flow rate under these assumptions occurs at the mid-point of the fundamental diagram, when $\rho = \rho_{\text{jam}}/2$, and that its value is

$$q_{\max} = \frac{1}{4}\rho_{\text{jam}}\, v_{\max}. \tag{6.16}$$

So, is the linear speed-density relationship of eq. (6.13) just a nice demonstration model, or does it have any real utility or validity in modeling traffic flow? As a matter of fact, it is useful. In studies conducted for the Lincoln, Holland, and Queens-Midtown Tunnels leading into New York's Manhattan island, for example, the linear speed-density relationship has been shown to be a very good approximation to the central (and dominant) part of the speed-density data gathered empirically. Such a curve is shown in Figure 6.3. We will return to this point in Section 6.3 because car-following models are expressly used to derive speed-density relationships.

Problem 6.3. Demonstrate that eq. (6.12) is correct.

Problem 6.4. Confirm qualitatively that eq. (6.12) produces the shape of the fundamental diagram of road traffic shown in Figure 6.2.

Problem 6.5. Show that the relationship (6.13) satisfies the conditions defined in eqs. (6.11a–c).

Problem 6.6. Derive and sketch the fundamental diagram for the speed-density relationship

$$v(\rho) = v_{\max}\left(1 - \left(\frac{\rho}{\rho_{\text{jam}}}\right)^2\right).$$

Problem 6.7. Derive and sketch the fundamental diagram for the speed-density relationship

$$v(\rho) = v_{\max}\left(1 - \left(\frac{\rho}{\rho_{\text{jam}}}\right)^m\right).$$

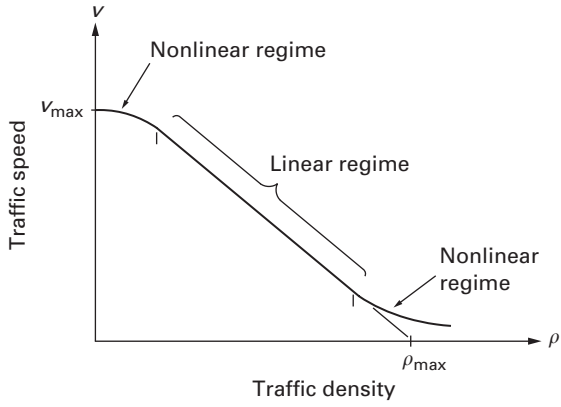


Figure 6.3 Another generic view of the variation of traffic velocity with density, based on the results often obtained when data is gathered for particular traffic systems. In addition to displaying the endpoints, $[(0, v_{\max}), (\rho_{\max}, 0)]$, and the non-positive $(dv/d\rho < 0)$ slope behavior, it shows that a significant portion of the curve can be modeled by a linear speed-density relationship.

6.2.4 The Continuum Hypothesis in Macroscopic Traffic Modeling

The macroscopic traffic flow analysis we have done so far has been predicated on the proposition that we could treat a line of traffic in the same way that we would model the flow of a fluid through an artery or tube, that is,

as a field. This means that the traffic line contains enough cars that instead of worrying about the speed of the ith car, $v_i(x, t)$, we choose to deal with a speed *field* in which every point along the x axis is assigned a unique speed $v(x, t)$. Thus, we have replaced the line of discrete cars at coordinates, $x = x_i$, by an infinite sequence of points, each having a unique speed expressed by the continuous function, $v(x, t)$. This is an application of the *continuum hypothesis* that we discussed in Section 4.7.2. Taking advantage of the continuum hypothesis allows us to deal with continuous fields (e.g., smooth curves) instead of discrete elements (e.g., histograms), which often makes the mathematics of model building much nicer. However, it carries drawbacks: in the present model, for example, we could not include cars overtaking and passing each other because that would require some points on the x axis to have two different speeds!

How many cars do we need for a macroscopic analysis? The answer depends on how we characterize the number of cars. We saw in Section 6.2.1 that we could measure the number of cars in two ways. One way is to stand at a fixed point and count the number of cars passing by during a fixed time interval, thus finding the traffic flow rate, $q(x, t)$, with units of cars per unit of time. The second way requires counting the number of cars in a given length of road and so determining the traffic density, $\rho(x, t)$, with units of cars per unit of distance. (As a practical matter, the density would be determined from aerial photographs of a given length of road.) In both instances we must ask whether our counting intervals are sufficiently long, that is, have we taken enough time to measure the traffic flow or enough distance to measure the density?

To measure the density, we must choose a length of road that is (1) not so short that we too often see fractions of cars or intervals with no cars at all, and (2) not so long that the meaningful fluctuations would simply cancel out. For example, a spatial count over the length of Interstate 5 between Los Angeles and San Francisco—about 350 miles—would miss both the buildup at cities along the way and the long stretches through farm country with sparse amounts of traffic. Figure 6.4 shows a conceptual sketch for just such a measurement, showing the variation of traffic density with the length of the measurement interval. (Note how similar it is to its cousin in Figure 4.9!) It illustrates the discontinuities arising from the fluctuations when the measuring interval is too short, and it shows the decline in the density when the measuring length becomes so long that the meaningful variations disappear. The central portion shows a regime where the *local* density is relatively constant. It is for this region that we can model our traffic density with a continuous field $\rho(x, t)$, in much the same way we replaced the speeds of individual cars with the continuous speed field, $v(x, t)$.

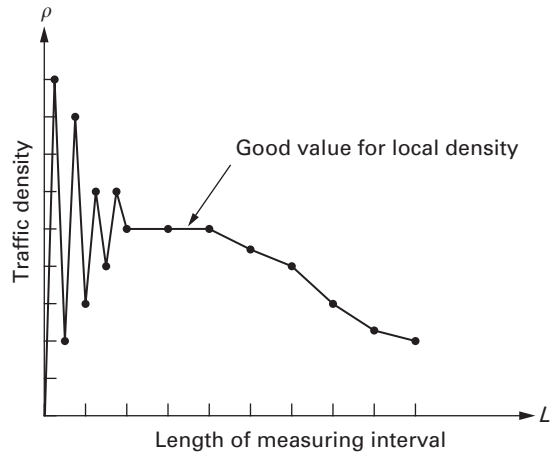


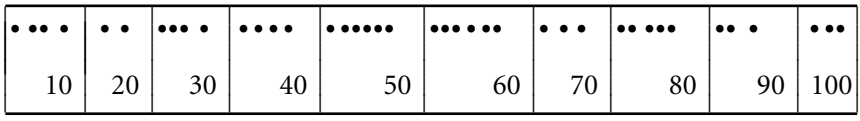
Figure 6.4 A conceptual plot of the variation of traffic density, ρ, against the length of the measuring interval. It shows that the central portion of the curve defines a useful approximation of the *local* traffic density that is (1) preceded by a regime where the density fluctuates too much because the measuring interval is too short, and (2) followed by a regime where the density progressively falls off because the measuring interval is too long.

A comparable situation obtains if the traffic flow rate, $q(x, t)$, is the measurement of choice. Here it is the length of the time interval that must be "just right." Short intervals before and after the change of a traffic light, say from red to green, would show no cars before and a sudden burst after. Similarly, counting by days would almost certainly cover up the peaks generated by morning and evening rush hours. Thus, again, there is a balancing act that must be performed in order to get the time measurement interval properly set.

How?

To sum up, the continuum hypothesis enables us to deal with *averaged* or gross variables of traffic speed, density, and flow rate that do *not* pertain to individual cars or vehicles, but to the fields that represent them. And, these fields are good models, or good representations of reality, if we have done our *scaling* properly in choosing the proper measurement intervals, that is, if we have properly set the measurement scales.

Problem 6.8. Consider a road segment 0.5 mi long that is divided into 10 equally-spaced intervals. There are 40 cars on the road, spaced as shown below, where the density of the dots represents the traffic density. Find a "good" value for the local density at point A in terms of the number of cars per mile, assuming for simplicity that each car has zero length.



6.3 Microscopic Traffic Models

Why? We now turn from macroscopic models that use averaged variables to *microscopic* models that look at individual cars. Our interest is in using the microscopic models to develop the traffic speed-density relations that we need to do macroscopic evaluations of capacity, which we require if Find? we're going to design highway systems. As we noted in Section 6.2.2, we are looking for models that describe how drivers *respond* to the *stimuli* of their traffic situations. The driver will *perceive* a variety of stimuli, including the distance between vehicles, their relative speed, and their perceived relative acceleration. We thus seek psychological, not mechanical, models in order to model human behavior. The driver's response will depend on the responder's *sensitivity* to the given stimuli, as well as on the speed with which the response is undertaken. Thus, some time delay should also be incorporated into such models.

6.3.1 An Elementary, Linear Car-following Model

Imagine a line of cars traversing a given road, as shown in Figure 6.5. Each car is identified by a discrete coordinate that varies in time, so that the Assume? location of the nth car is given by $x_n(t)$. We also assume that the line has a reasonable value of local density and does not permit passing or overtaking. How? Then the basic "equation" of car-following for such a single lane of traffic is the psychological one:

$$\text{response} = \text{sensitivity} \bullet \text{stimulus}. \tag{6.17}$$

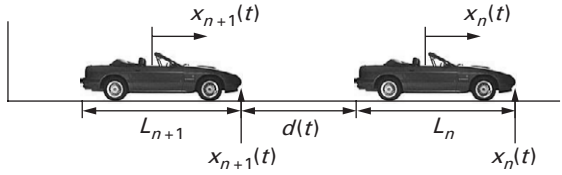


Figure 6.5 The nomenclature for a line (or lane) of cars on a highway of total length, L_R. Each car has the same length, L, and is separated from its neighbors by a common distance, $d(t)$. The discrete functions, $x_{n+1}(t)$ and $x_n(t)$, represent, respectively, the coordinates of the *follower* and *leader* cars.

The response will generally be modeled as the acceleration of the $(n+1)$st *follower* car, $\ddot{x}_{n+1}(t)$, as it moves behind the nth *leader* car. The stimulus will be modeled in terms of the coordinate of the follower car *relative to the leader car*, which can in turn be written in terms of the traffic density, ρ. The acceleration is then integrated to determine the speed of that car as a function of the traffic density, which is the input we require for our macroscopic modeling. **Predict?**

Consider a simple linear car-following model in which the driver of the follower car responds to the speed of the leader car relative to the follower car:

$$\frac{d^2x_{n+1}(t)}{dt^2} = -K_p\left(\frac{dx_{n+1}(t)}{dt} - \frac{dx_n(t)}{dt}\right). \tag{6.18}$$

The coefficient, K_p, introduced here is a sensitivity parameter that has dimensions of per unit time. Note, that with $K_p > 0$, the follower car will decelerate to avoid hitting the car in front if it is slowing down, relatively speaking. We will discuss this in further detail later.

We can model the time it takes the following driver to respond to events by building in a *reaction* time that slows the follower's acceleration by the *delay* time T:

$$\frac{d^2x_{n+1}(t+T)}{dt^2} = -K_p\left(\frac{dx_{n+1}(t)}{dt} - \frac{dx_n(t)}{dt}\right). \tag{6.19}$$

Assuming that the sensitivity parameter, K_p, is a constant, eq. (6.19) is a linear ordinary differential equation with constant coefficients that can be integrated once to yield

$$\frac{dx_{n+1}(t+T)}{dt} = -K_p(x_{n+1}(t) - x_n(t)) + C_{n+1}, \tag{6.20}$$

where C_{n+1} is the arbitrary constant, with dimensions of speed, that results from the integration just performed. Note that eq. (6.20) clearly relates the speed of the follower car to the distance maintained between the follower and leader cars. Thus, it is a natural precursor of the speed-density relationship that we seek.

Assume? Let us further assume that all of the cars have the same length, L, and that the spacing between common points on any pair of cars (see Figure 6.5) is given by $d(t)$:

$$d(t) = x_n(t) - L - x_{n+1}(t). \tag{6.21}$$

It then follows that the number of cars, N_R, found in a stretch of road of length, L_R, is

$$N_R = \frac{L_R}{L + d(t)}, \tag{6.22}$$

which means that the density of cars on that road is

$$\rho = \frac{L_R}{N_R} = \frac{1}{L + d(t)} = \frac{1}{x_n(t) - x_{n+1}(t)}, \tag{6.23}$$

where we have used the spacing defined in eq. (6.21) to obtain the final form of eq. (6.23). Thus, we have in eq. (6.23) a relationship between the (macroscopic) traffic density, ρ, and the (microscopic) coordinates of the leader and follower cars.

There is an important point about the units of eq. (6.23) that should be kept in mind. With particular reference to the units still used by American traffic engineers, both car lengths and inter-vehicle distances are typically measured in feet, while density is expressed in units of vehicles per mile. Thus, for numerical calculations, eq. (6.23) should be written in consistent numerical units:

$$\rho = \frac{5280}{L + d(t)} \left(\frac{\text{vehicles}}{\text{mile}} \right). \tag{6.24}$$

Assume? Let us still further assume, for now at least, that the traffic flow is in a *steady state*, by which we mean that all of the cars are traveling at the same speed. Then

$$\frac{dx_{n+1}(t+T)}{dt} = \frac{dx_{n+1}(t)}{dt} \equiv v. \tag{6.25}$$

Equation (6.24) shows a relationship between the (macroscopic) speed, v, and the (microscopic) speeds of any of the follower cars. Additionally, for this steady state, the arbitrary constant C_{n+1} is the same for any adjacent pair of cars. Thus, we can now substitute eqs. (6.23) and (6.25) into eq. (6.20) to find

$$v = \frac{K_p}{\rho} + C. \tag{6.26}$$

The constant, C, can be determined from the condition cited in eq. (6.11c), namely, that the speed is zero when the density is at its maximum or jam value. Hence it follows that

$$v = K_p \left(\frac{1}{\rho} - \frac{1}{\rho_{\text{jam}}} \right). \tag{6.27}$$

The speed-density relationship of eq. (6.27) is sketched in Figure 6.6.

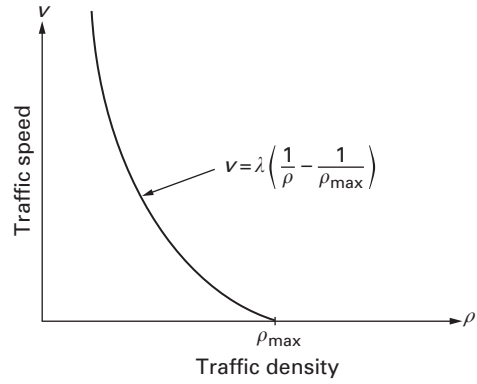


Figure 6.6 A schematic curve illustrating the traffic speed-density relationship [see eq. (6.27)] corresponding to a linear car-following model in which the driver responds to the relative speed of the car ahead.

Valid?

The curve shown in Figure 6.6 seems reasonable enough (see Problem 6.10), except for the fact that it shows an infinite speed as the density goes to zero, a result that hardly seems credible. This is an almost classical modeling dilemma: we have a model that seems reasonable and credible over a good portion of the relevant domain, but that crashes in some region. Can this model be improved or fixed? It can be fixed, or improved; it depends on what we want from this model.

Improve?

Fixing the high (infinite at $\rho = 0$) speed at small values of the density is straightforward enough. All we need do is stipulate that a maximum speed applies for all values of density below some (specified) critical density. This seems like a reasonable fix that roughly accords with our everyday driving experience. This fix is shown in Figure 6.7 and in eqs. (6.28a–b):

$$v(\rho) = \begin{cases} v_{\max} & \rho < \rho_{\text{crit}} \\ K_p \left(\dfrac{1}{\rho} - \dfrac{1}{\rho_{\text{jam}}} \right) & \rho \geq \rho_{\text{crit}} \end{cases} \tag{6.28a}$$

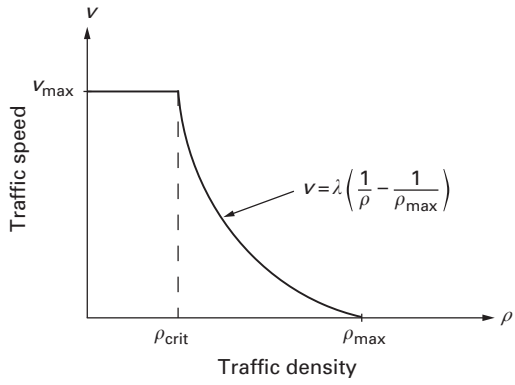


Figure 6.7 A schematic curve illustrating the traffic speed-density relationship [eqs. (6.28a–b)] corresponding to the *fixed* linear car-following model in which the driver responds to the relative speed of the car ahead—except at small values of the density, $\rho < \rho_{crit}$, for which the maximum speed has a fixed upper limit of $v = v_{max}$.

and

$$\rho_{crit} = \left(\frac{v_{max}}{K_p} + \frac{1}{\rho_{jam}}\right)^{-1}. \tag{6.28b}$$

The traffic flow rate corresponding to this fixed speed-density relationship is found as:

$$q(\rho) = \begin{cases} \rho v_{max} & \rho < \rho_{crit} \\ K_p\left(1 - \dfrac{\rho}{\rho_{jam}}\right) & \rho \geq \rho_{crit} \end{cases} \tag{6.29}$$

The traffic flow rate, pictured in Figure 6.8, increases linearly with density (from zero), and reaches its maximum value, the capacity, when $\rho = \rho_{crit}$:

$$q_{max} = q(\rho_{crit}) = \rho_{crit} v_{max} = K_p\left(1 - \frac{\rho_{crit}}{\rho_{jam}}\right). \tag{6.30}$$

For density values $\rho \geq \rho_{crit}$, the traffic flow rate decreases linearly with ρ from its maximum value at $\rho = \rho_{crit}$ until it vanishes altogether at $\rho = \rho_{jam}$.

Verified? How good is this model? As luck would perhaps have it, having just fixed a model that is incredible (literally!), we are still left with one that does compare well with some available data. In Figures 6.9 and 6.10 we show measurement data made in Orange County, California, on the I–405 freeway. It yields reasonable values of the jam (or maximum) density and, as shown in Figure 6.10, the shape of the resulting traffic flow rate curve

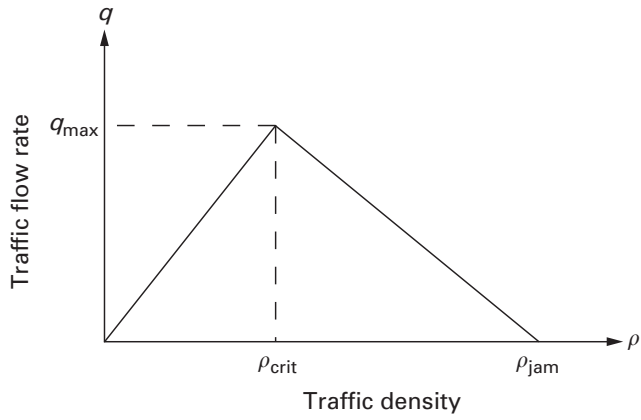


Figure 6.8 A schematic curve illustrating the relationship between the traffic flow rate and the density [eq. (6.29)] for the *fixed* linear car-following model in which the driver responds to the relative speed of the car ahead. Note that the maximum traffic flow rate $q = q_{max}$ occurs when $\rho = \rho_{crit}$.

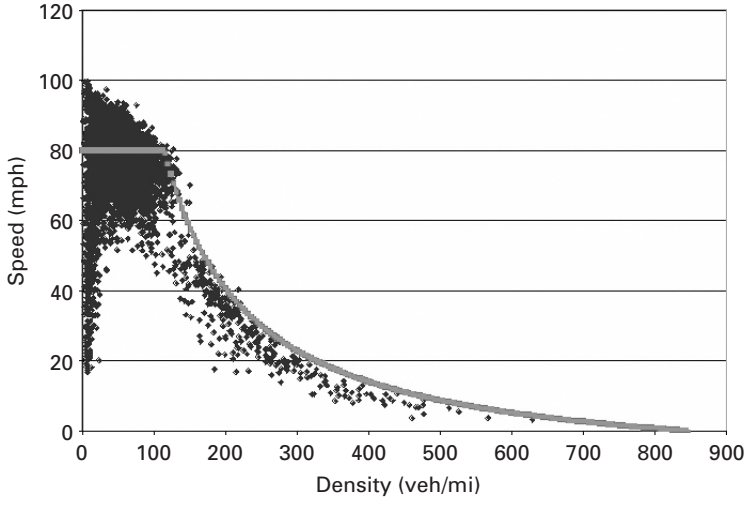


Figure 6.9 Some traffic speed-density data measured for the I–405 freeway in Orange County, California, plotted along with corresponding results from the *piecewise linear* or *triangular* car-following model [eq. (6.38)] (Recker, 2003). The corresponding parameter values are $S_f = 80$ mph, $q_{crit} = 2300$ cars/hr, and $\rho_{jam} = 211$ cars/mi.

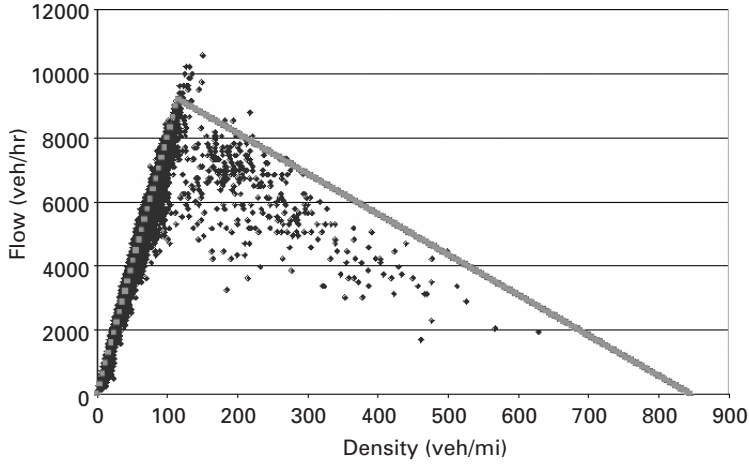


Figure 6.10 Some traffic flow rate data measured for the I–405 freeway in Orange County, California, plotted along with the *piecewise linear* or *triangular* car-following model [eq. (6.35)] (Recker, 2003). The corresponding parameter values are $S_f = 80$ mph, $q_{\text{crit}} = 2300$ cars/hr, and $\rho_{\text{jam}} = 211$ cars/mi.

follows the data for traffic parameter values that are not uncommon on California freeways, including speeds up to 80 mph and jam densities of 211 veh/mi/lane that correspond to vehicles stopped at 25 ft separation.

Another aspect of this model is worth noting. One of the heuristics or rules of thumb offered by state Departments of Motor Vehicles (DMV) is that drivers should maintain a distance behind the car immediately in front that is equal to one car length, L (ft.), for each increment of 10 mph of the car's speed. Thus, the DMV heuristic would require that

$$d(t) = \left(\frac{L}{10}\right) v. \tag{6.31}$$

If eq. (6.31) is substituted into our previous, units-corrected definition of the traffic density (6.24), we immediately obtain a speed-density relationship

$$\rho = \frac{5280}{L + (L/10)v},$$

that can be recast in the form:

$$v = \frac{5280(10)}{L}\left(\frac{1}{\rho}\right) - 10. \tag{6.32}$$

Equation (6.32) bears an unmistakable resemblance to the result (6.27) derived just above (see Problems 6.13–6.15).

Problem 6.9. Derive eq. (6.21) from Figure 6.5.

Problem 6.10. Determine whether or not eq. (6.26) satisfies each of the three conditions in eqs. (6.11a–c).

Problem 6.11. Derive the result presented in eq. (6.28b). Is it dimensionally correct?

Problem 6.12. Confirm the traffic flow rate results shown in eqs. (6.29).

Problem 6.13. Determine the values of the constants, K_p and ρ_{jam}, that make eqs. (6.27) and (6.32) identical.

Problem 6.14. Why does the DMV model produce the same form (and numbers) as the speed-sensitive car-following model?

Problem 6.15. What is the physical interpretation of ρ_{jam} for the DMV model?

6.3.2 An Alternate Derivation of the Same Model

Improve?
Assume?

Suppose we want to derive the above model using an empirical, yet "mechanical" approach. We know that flow rate increases with density until it reaches a critical value, and then it decreases to zero at the jam density. Thus—without benefit of the car-following model (6.25) or the data we have already seen in Figure 6.10!—we assume a priori that the traffic flow rate will behave in a piecewise linear fashion, in the following *triangular* traffic flow rate:

$$q(\rho) = \begin{cases} A\rho & \rho < \rho_{\text{crit}} \\ B\left(1 - \dfrac{\rho - \rho_{\text{crit}}}{\rho_{\text{jam}} - \rho_{\text{crit}}}\right) & \rho \geq \rho_{\text{crit}} \end{cases} \tag{6.33}$$

where the constants A and B are determined from the requirement that $q(\rho)$ be continuous at $\rho = \rho_{\text{crit}}$, that is,

$$q(\rho = \rho_{\text{crit}}) = q_{\text{crit}}. \tag{6.34}$$

Thus, eq. (6.33) becomes

$$q(\rho) = \begin{cases} q_{\text{crit}}\left(\dfrac{\rho}{\rho_{\text{crit}}}\right) & \rho < \rho_{\text{crit}} \\ q_{\text{crit}}\left(1 - \dfrac{\rho - \rho_{\text{crit}}}{\rho_{\text{jam}} - \rho_{\text{crit}}}\right) & \rho \geq \rho_{\text{crit}} \end{cases} \tag{6.35}$$

The speed-density relationship corresponding to the traffic flow rate (6.35) is then found by applying the relationship (6.9) between the traffic flow rate and the speed, so that

$$v(\rho) = \begin{cases} \dfrac{q_{\text{crit}}}{\rho_{\text{crit}}} & \rho < \rho_{\text{crit}} \\ \dfrac{q_{\text{crit}}}{\rho_{\text{crit}}} \left(\dfrac{\dfrac{\rho_{\text{jam}}}{\rho} - 1}{\dfrac{\rho_{\text{jam}}}{\rho_{\text{crit}}} - 1} \right) & \rho \geq \rho_{\text{crit}} \end{cases} \tag{6.36}$$

While the speed-density relationship in eq. (6.36) does not have the nice, linear properties of the speed-density of eq. (6.13), we have maintained the corresponding piecewise linear flow-density relationship. Equations (6.35) and (6.36) have the same form as, respectively, eqs. (6.29) and (6.28), although they were derived by very different means!

One interesting version of the results in eq. (6.36) is their presentation in terms of a parameter called the *free-flow speed*, S_f, which is the speed at which a driver would travel if all alone on the road, that is, if the density were zero. From the first of eq. (6.36) we find that

$$S_f = \frac{q_{\text{crit}}}{\rho_{\text{crit}}}, \tag{6.37}$$

from which it follows that eqs. (6.36) now become:

$$v(\rho) = S_f \begin{cases} 1 & \rho < \rho_{\text{crit}} \\ \left(\dfrac{\dfrac{\rho_{\text{jam}}}{\rho} - 1}{\dfrac{\rho_{\text{jam}}}{\rho_{\text{crit}}} - 1} \right) & \rho \geq \rho_{\text{crit}} \end{cases} \tag{6.38}$$

Verified? Equations (6.38) and (6.35), with parameter values of $S_f = 80$ mph, $q_{\text{crit}} = 2300$ cars/hr, and $\rho_{\text{jam}} = 211$ cars/mi, are shown in Figures 6.9 and 6.10, together with data taken from the I–405 freeway measurements. We see that the agreement is quite good over most of the range of density for both the speed and the traffic flow.

6.3.3 Comments on Car-following Models

It is worth noting that the two models just presented were found in very different ways. The elementary and *fixed* car-following models of Section 6.3.1 were derived from a stimulus-response model that was re-worked into a speed-density relationship, from which we then obtained the traffic flow rate. The revised model presented in Section 6.3.2 was found by starting

with traffic flow rate data and trying to create a model to match that data. Indeed, we have not gone so far as to find a matching stimulus model for the improved model. Does that matter?

The answer is a familiar one: it depends. If our principal goal is the one we claimed earlier, that of modeling capacity, then it matters less which of the two approaches we use as long as we can validate and verify the results. On the other hand, in an emerging area of transportation engineering, efforts are being made to model the *control* of vehicles, with the aim of trying to maximize the flow of traffic by more effectively controlling how each vehicle is driven. This area encompasses a number of exciting prospects that are, unfortunately, beyond our present scope. Achieving results in the latter case means that stimulus-response control modeling will be required, while "only" good modeling of traffic speed and traffic flow rate is required for capacity-based engineering to move forward.

6.4 Summary

This chapter has introduced some of the most fundamental ideas of traffic modeling as they are applied in the engineering of traffic systems. We described macroscopic models that predict the average variables of traffic density and traffic flow rates because they are very important for calculating the *capacity* of roads and highways. We then pointed out the role of scaling and of the continuum hypothesis in moving from macroscopic models to microscopic and in beneficially integrating the two. We introduced microscopic models that predict how speed varies with driver sensitivities and responses to various traffic stimuli because they provide a basis for obtaining the gross traffic density and flow rates needed in macroscopic models. Finally, we also noted in passing that the microscopic models are increasingly used to investigate the control of individual vehicles, as well as lines (or lanes) of vehicles.

6.5 References

R. E. Chandler, R. Herman, and E. W. Montroll, "Traffic Dynamics: Studies in Car Following," *Operations Research*, **6**(2), 165–184, 1958.

R. E. Chandler, E. W. Montroll, R. B. Potts, and R. W. Rothery, "Traffic Dynamics: Analysis of Stability in Car Following," *Operations Research*, **7**(1), 86–106, 1959.

B. D. Greenshields, "A Study in Highway Capacity," *Highway Research Board Proceedings*, **14**, 1935.

R. Haberman, *Mathematical Models*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

G. Newell, "A Simplified Theory of Kinematic Waves: I General Theory; II Queuing at Freeway Bottlenecks; III Multi-destination Flows," *Transportation Research-B*, **27B**, 281–313, 1991.

W. W. Recker, "Understanding the Nature of Traffic," *Notes for CE122–Transportation Systems II: Operations and Control*, University of California, Irvine, CA, Winter 2003.

J. A. Wattleworth, "Traffic Flow Theory," in J. E. Baerwald (Ed.), *Transportation and Traffic Engineering Handbook*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

M. Wohl and B. V. Martin, *Traffic Systems Analysis*, McGraw-Hill, New York, 1967.

6.6 Problems

6.16. What is the meaning and physical significance of the statement, $\partial q/\partial x > 0$, (i.e., that the macroscopic traffic flow rate, $q(x,t)$, increases with the distance, x, along the line of traffic)?

6.17. If the average length of a car (in pre-*Expedition* days) is 5 m, what is the density of traffic in a line when its cars are maintaining a distance of two car lengths between themselves. What is the traffic flow if the line is moving at 80 km/hr (50 mph)? (*Hint*: You may ignore the fact that the data given ignores both AAA recommendations and your own experience on a freeway or turnpike.)

6.18. (a) Assume that velocity depends linearly on density, such that $v(\rho) = a + b\rho$. Determine the values of a and b in terms of the maximum values of the speed and the density, assuming that the assumptions of eqs. (6.11a–c) hold.

(b) How does the flow depend on the density?

6.19. (a) Sketch the fundamental diagram of road traffic for the model developed in Problem 6.18 if $a = 80$ km/hr and $b = -10^5$ m^2/car·hr.

(b) Determine the values of the density and the speed when the flow is a maximum.

(c) What is the capacity of the road being modeled?

6.20. Consider a flow-density relationship of the form $q(\rho) = \rho(\alpha - \beta\rho)$. The best fit (i.e., least squares) of this relationship to some real traffic data occurred when $\alpha = 91.33$ km/hr and $\beta = 1.4$ km^2/car·hr.

(a) What is the maximum density?

(b) What is the maximum speed?

(c) What is the capacity of the road?
(d) Identify the type of road being modeled and explain your identification.

6.21. Find the speed of traffic on a line of traffic for which there are three car lengths between the leader and follower cars. (*Hint*: Use macroscopic traffic theory with a linear speed-density relation.)

6.22. Determine the capacity of the road described in Problem 6.21 if cars are assumed to be 5 m long, $v_{max} = 88.5$ km/hr and $\rho_{max} = 0.22^{-1}$.

6.23. The data in the table shown below were obtained by recording the indicated parameters along a busy stretch of highway.

(a) Sketch the fundamental diagram for this traffic flow.
(b) What is the maximum traffic flow?
(c) What are the density and speed at the maximum flow rate?

Speed (mph)	Density (cars/mi)
42	44
40	49
37	53
35	58
32	64
28	67
26	69
23	74
20	80
19	85
18	90
17	95
16	101
15	106
14	112
13	120
12	128
11	139
10	151
9	166

6.24. Plot traffic speed against traffic density for the data given in Problem 6.23. Draw an approximate curve through this data and estimate the maximum values of the speed and the density on this road.

7
Modeling Free Vibration

We now turn to modeling vibration, the behavior of something moving back and forth, to and fro, usually in an evident rhythmic pattern. Vibration not only occurs all around us, but within us as well, as noted in 1965 by a well-known British mechanical engineer, R. E. D. Bishop:

After all, our hearts beat, our lungs oscillate, we shiver when we are cold, we sometimes snore, we can hear and speak because our eardrums and our larynges vibrate. The light waves which permit us to see entail vibration. We move by oscillating our legs. We cannot even say 'vibration' properly without the tip of the tongue oscillating. And the matter does not end there—far from it. Even the atoms of which we are constituted vibrate.

Other vibratory phenomena that come to mind are pendulums, clocks, conveyor belts, machines and engines, buildings subjected to a broad array of moving forces (e.g., pedestrians, air conditioners, elevators, wind, earthquakes), as well as tides and seasons. Clearly, we could go on. But the more interesting questions for us are: Do these diverse instances of vibration have anything in common? If so, what? How do we model their common features?

Why?

We devote most of this chapter to modeling a well-known "golden oldie," the swinging or vibrating pendulum. It provides a familiar platform upon which we can lay out a number of modeling strategies. Then we will provide a few examples of freely vibrating phenomena. We will also illustrate how the mathematics of free vibration can be used to model *stability* phenomena. In Chapter 8 we will provide some more examples and then go on to model forced vibration.

7.1 The Freely-Vibrating Pendulum–I: Formulating a Model

Given? We will now model the free vibration of a pendulum, starting with some experimental results and using dimensional analysis, some basic physics, and some basic mathematics (e.g., linearity, second-order differential equations) to model that motion.

7.1.1 Some Experimental Results

How? We started by building some very simple pendulums in the laboratory, each consisting of a lead-filled wooden ball suspended from a stand by an ordinary piece of string. A basic schematic of the laboratory set-up is shown in Figure 7.1. The balls were initially held at rest at some angle, θ_0, and then they were let go to swing back and forth until they all stopped moving. As each pendulum swung, we measured its *period of free vibration*, the time T_0 it takes to swing through two complete arcs (from $\theta = \theta_0$ to $\theta = -\theta_0$ and back again). The periods of vibration were measured with photoelectric cells that were placed at the lowest point on the pendulum arc ($\theta = 0$) and were in turn connected to digital counters operating with a gated pulse. The counters were turned on by the first passing of the pendulum and then off again at the second passing, thus providing a direct read of one-half of the period T_0.

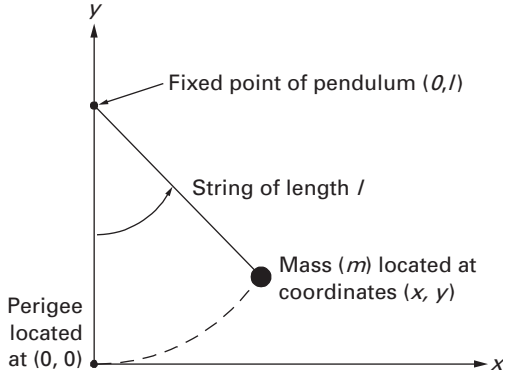


Figure 7.1 The geometry of a planar pendulum. Note that the origin of the coordinate system is located at the pendulum's *perigee*, the lowest point of its arc.

Table 7.1 The dependence of the period, T_0, of a freely-vibrating pendulum on its initial amplitude of vibration, θ_0. The mass is 390 gm and the string length is 276 cm.

θ_0(deg)	θ_0(rad)	T_0 measured (sec)	(T_0 measured)/(3.372)
8.34	0.1456	3.368	1.00
13.18	0.2300	3.368	1.00
18.17	0.3171	3.372	1.00
23.31	0.4068	3.372	1.00
28.71	0.5011	3.390	1.01
33.92	0.5920	3.400	1.01
39.99	0.6980	3.434	1.02
46.62	0.8137	3.462	1.03

The experiments were done with two different masses (237 gm and 390 gm), each of which was hung from strings of two different lengths (276 cm and 226 cm). The experimental data thus obtained are shown in Tables 7.1 and 7.2; note that each data point shown represents the average of five measured values. Thus, the data presented result from a consistent, repeatable experiment. The data in Table 7.1, for the larger mass (390 gm) and the shorter string (276 cm), show how the period, T_0, varies with different starting values of θ_0. We see that the period varies with the initial starting angle, θ_0, but the dependence is very weak and exceeds 1% only when $\theta_0 \geq 40^\circ$.

Table 7.2 The dependence of the period, T_0, of a freely-vibrating pendulum on its length and on its mass. The data show a marked change with length, but virtually no change with mass.

	$m = 237$ gm	$m = 390$ gm
$l = 226$ cm	3.044 sec	3.058 sec
$l = 276$ cm	3.350 sec	3.372 sec

The data in Table 7.2 summarize the periods across the four possible combinations of mass and length that were available for the pendulums used in this experiment. This data suggest that the period varies very little, if at all, with mass: increasing the mass by some 65% from 237 gm to 390 gm changes the period by a fraction of 1%. On the other hand, increasing the length by 22% from 226 cm to 276 cm increases the period by approximately 10%. Thus, the data suggest that the free motion of a vibrating

pendulum is *periodic*, and that the period of vibration does not depend on the pendulum's mass, but that it does depend on the pendulum's length.

Problem 7.1. Assume a hypothetical relationship, $T_0 = am^b$, for the dependence of the period of a pendulum on its mass. Determine the unknown parameters, a and b, using the data in Table 7.2. (*Hint*: Logarithms may be useful here.)

Problem 7.2. Assume a hypothetical relationship, $T_0 = cl^d$, for the dependence of the period of a pendulum on its length. Determine the unknown parameters c and d using the data in Table 7.2. (*Hint*: Logarithms may be useful here.)

7.1.2 Dimensional Analysis

We will now apply some dimensional analysis results to formalize the results we obtained in the laboratory. In Section 2.4.2 we used the Buckingham Pi theorem to determine that the period of vibration, T_0, of a pendulum was related to its length, l, and the gravitational acceleration, g [see the first of eq. (2.30)]:

$$T_0 = \Pi_1 \sqrt{\frac{l}{g}}. \tag{7.1}$$

Valid? Note that the pendulum's period does not depend on mass, a result supported by the data in Table 7.2, and that the constant, Π_1 is dimensionless. We can determine the value of Π_1 from the data given in Table 7.2. For the pendulum of length $l = 276$ cm, one measured value of the period is $T_0 = 3.372$ sec, so that with $g = 980$ cm/sec/sec,

$$\Pi_1 = \frac{3.372}{\sqrt{276/980}} \cong 6.35. \tag{7.2}$$

Is the number "6.35" in eq. (7.2) some new universal constant? Actually, no. Rather, it is an approximation of another well-known constant: $2\pi \cong 6.28$. Thus, substituting this judgment call about the constant into eq. (7.2) yields the final result,

$$T_0 = 2\pi \sqrt{\frac{l}{g}}. \tag{7.3}$$

Table 7.3 Calculated values of the period, T_0, of a freely-vibrating pendulum that provide support for the experimental data presented in Table 7.2.

l (cm)	T_0 (sec)
226	3.02
276	3.33

Predict?
Verified?

We can use eq. (7.3) to predict values of the period to match the remaining values displayed in Table 7.2, as shown in Table 7.3. The calculated predictions and the experimental data agree to within less than 1.5%. Thus, it seems that we have a pretty good model—determined from dimensional analysis and use of some experimental data—that works quite well and predicts the remaining experimental data, including both the period's dependence *on* length and its independence *of* mass. We will confirm the model (7.3) again before we're done with the pendulum.

7.1.3 Equations of Motion

How?

We formulate the problem by writing the mathematical expression of a balance or conservation principle (see Section 1.3.3) from physics. The principle is Newton's second law: *The time rate of change of momentum is equal to the net force producing it; that momentum change is in the same direction as the net force.* Newton's second law is both a balance principle and a conservation principle: it reflects a balance of the forces acting on a particle or system, and it also reflects the conservation of momentum. Written as a balance principle (see Problems 7.3 and 7.4), Newton's second law in a plane is:

$$\sum F_x = m\frac{d^2x}{dt^2}, \tag{7.4a}$$

and

$$\sum F_y = m\frac{d^2y}{dt^2}, \tag{7.4b}$$

where $x(t)$ and $y(t)$ are the time-dependent coordinates of a mass, m, acted on by net forces $\sum F_x$ and $\sum F_y$, respectively.

We want to apply Newton's second law, commonly referred to as *equations of equilibrium*, to the pendulum depicted in Figure 7.1. The pendulum is simply a mass, m, attached to the end of a string of length, l. It swings in

a plane from an attachment point with coordinates $(0, l)$ so that the origin of the coordinates coincides with the *perigee* or low point of the pendulum's arc. The coordinates (x, y) of the pendulum mass can be written in terms of the string length and the angle θ between the string and the y-axis:

$$x(t) = l \sin \theta(t), \tag{7.5a}$$

and

$$y(t) = l(1 - \cos \theta(t)), \tag{7.5b}$$

In Figure 7.2 we show a *free-body diagram* (FBD) of the two forces that act on the mass: the tension in the string, T, and the weight, mg, which acts due to the pull of gravity. Then we can identify the net forces along the coordinates from the FBD, so that eqs. (7.4) can then be written as *equations of motion*:

$$m\frac{d^2x}{dt^2} = \sum F_x = -T \sin \theta, \tag{7.6a}$$

and

$$m\frac{d^2y}{dt^2} = \sum F_y = T \cos \theta - mg. \tag{7.6b}$$

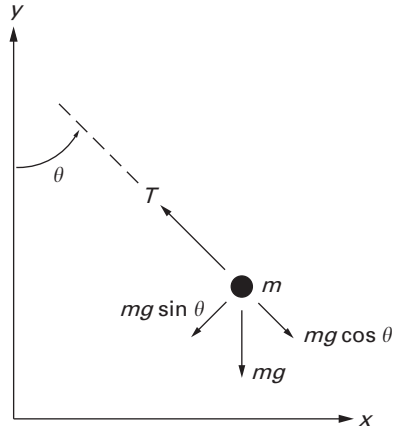


Figure 7.2 A *free-body diagram* (FBD) of the oscillating planar pendulum. It shows the two forces acting on the pendulum's mass, m, the string tension, T, and the weight, mg, and their components in the radial and tangential directions.

In principle, all we need to do now is integrate eqs. (7.6a–b) to find how the pendulum's coordinates vary with time, from which we can then find out whatever else we might want to know about the pendulum. However, life's not that easy, for a number of reasons. First, we don't know the tension in the string, T, so that the right-hand sides of both of eqs. (7.6a–b) are unknown. Second, since we have two equations with *three* unknowns— **Improve?** $x(t)$, $y(t)$, T—we are prompted to wonder how Newton's second law would look if rewritten in *radial* (along the string) and *tangential* (to the pendulum's arc) coordinates. In fact, those equations are

$$\sum F_{\text{radial}} = ml\left(\frac{d\theta}{dt}\right)^2, \tag{7.7a}$$

and

$$\sum F_{\text{tangential}} = ml\frac{d^2\theta}{dt^2}. \tag{7.7b}$$

Equation (7.7a) clearly displays the familiar centripetal acceleration. If we sum the forces in the FBD of Figure 7.2 in the radial and tangential directions, we would find that

$$T = ml\left(\frac{d\theta}{dt}\right)^2 + mg\cos\theta, \tag{7.8a}$$

and

$$ml\frac{d^2\theta}{dt^2} + mg\sin\theta = 0. \tag{7.8b}$$

Equations (7.8a–b) show two equations for two dependent variables, the tension, T, and the angle, θ. Equation (7.8b) is a single equation with a single unknown, θ, so it can in principle be solved on its own, which thus determines the location of the mass [see also eqs. (7.5a–b)]. Then the tension, T, can be obtained directly by substituting the newly-found θ into eq. (7.8a). We also note that eqs. (7.8a–b) are equivalent to eqs. (7.6a–b): both are representations of Newton's second law, eqs. (7.8a–b) written in radial and tangential coordinates (l, θ), eqs. (7.6a–b) in rectangular coordinates (x, y).

We further note that eqs. (7.8a–b) are decidedly nonlinear because the dependent variable $\theta(t)$ or its derivatives have an exponent different than 1. This is most obvious in eq. (7.8a) because of the centripetal acceleration (see Problem 7.5), but it is equally true of eq. (7.8b) because

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots. \tag{7.9}$$

As we noted in Section 1.3.4, the presence of such nonlinear terms means that superposition, one of the most powerful weapons in the arsenal of

mathematics, is no longer available. We will return to this point in greater detail in Section 7.3.

Problem 7.3. Why do eqs. (7.4a–b) represent Newton's second law as a balance principle?

Problem 7.4. How would eqs. (7.4a–b) be written as a conservation principle?

Problem 7.5. Identify and explain *all* of the nonlinearities in eq. (7.8a).

7.1.4 More Dimensional Analysis

Valid? Are the dimensions of eqs. (7.8a–b) correct and consistent? Can we use dimensional information to further our understanding? In Table 7.4, we show (again, see Table 2.2) the pendulum variables expressed in terms of the fundamental dimensions of mass, length, and time. With this data, we can confirm (see Problem 7.6) that each of the terms in eqs. (7.8a–b) has the physical dimensions of *force*, or in terms of fundamentals, $(\mathsf{M} \times \mathsf{L})/\mathsf{T}^2$, which is appropriate for an equation of equilibrium. Further, we have satisfied the test that every stand-alone term in an equation has the same dimensions.

We now introduce a *scaling factor*, ω_0, that has, by definition, the dimensions of $1/\mathsf{T}$. The scaling factor also allows us to introduce a *dimensionless* time variable, τ, defined as

$$\tau = \omega_0 t. \tag{7.10}$$

Table 7.4 The fundamental dimensions of the six derived quantities chosen to model the oscillating pendulum.

Derived Quantities	Dimensions
Length (l)	L
Gravitational acceleration (g)	L/T^2
Mass (m)	M
Period (T_0)	T
Angle (θ)	$\mathsf{1}$
String tension (T)	$(\mathsf{M} \times \mathsf{L})/\mathsf{T}^2$

Then the tangential equation of motion (7.8b) can be written as (see Problem 7.7)

$$l\omega_0^2 \frac{d^2\theta(\tau)}{d\tau^2} + g \sin\theta(\tau) = 0. \tag{7.11}$$

Hence, if we choose the scaling factor, ω_0, to be

$$\omega_0 = \sqrt{g/l}, \tag{7.12}$$

we can write the tangential equation of motion (7.11) in a rather elegant form that is *completely dimensionless*:

$$\frac{d^2\theta(\tau)}{d\tau^2} + \sin\theta(\tau) = 0. \tag{7.13}$$

Note that the dimensions of the scaling factor are reciprocal to the dimensions of the period of free vibration, T_0, and that eqs. (7.3) and (7.12) can be combined to eliminate the common radicand, thus yielding:

$$T_0 = \frac{2\pi}{\omega_0}. \tag{7.14}$$

Equation (7.14) strongly suggests that we should recognize that the scaling factor, ω_0, is actually the circular frequency of the pendulum, that is, the measure of the pendulum's periodicity expressed in radians per unit of time. **Use?**

Now that we have confirmed dimensional consistency and cast at least one of our equilibrium equations in an elegant, dimensionless form, can we learn anything else? We can. We start by observing that $|\sin\theta| \leq 1$. This means that the acceleration term in eq. (7.11) must also exhibit similar behavior: $|d^2\theta/d\tau^2| \leq 1$, which provides a time scale for the problem. To demonstrate this, consider the function:

$$\theta(\tau) = \theta_0 \cos\tau, \tag{7.15}$$

for which it follows that

$$\frac{d\theta(\tau)}{d\tau} = -\theta_0 \sin\tau \quad \text{and} \quad \frac{d^2\theta(\tau)}{d\tau^2} = -\theta_0 \cos\tau. \tag{7.16}$$

which means that $\theta(\tau)$ and all of its derivatives with respect to τ have the same maximum amplitude θ_0.

If we choose to make our independent variable, t, dimensionless as we just did, are there any restrictions we need to place on its dimensionless counterpart, τ? No. Equations (7.10) and (7.12) tell us that

$$\tau = t\omega_0 = \frac{t}{1/\omega_0}, \tag{7.17}$$

which can be seen as a "verbal" or "conceptual equation":

$$\tau = \frac{\text{actual physical time}}{\text{a constant with dimensions of time}}. \tag{7.18}$$

Use? Equation (7.18) tells us that we get to choose how we make our equations dimensionless by choosing "a constant with dimensions of time" to match the problem of interest. If we are modeling something that takes years, the "constant" should be expressed in years. Then, small values of the dimensionless time, τ, would mean times of weeks, days, or even hours. Large values of the dimensionless time, τ, would mean times of decades, centuries, or even millennia.

Sometimes the "constant" is determined or dictated by the physics of the problem being investigated. For example, for a pendulum that is 1 m long, $\omega_0 = \sqrt{g/l} \cong 3.13\ \text{sec}^{-1}$, we would say that the system has a characteristic time of about one-third of a second—implying that the pendulum is moving rather fast. For a rather long pendulum, say $l = 98$ m, $\omega_0 = \sqrt{g/l} \cong 0.31\ \text{sec}^{-1}$ the system has a characteristic time of about 3 sec.

Problem 7.6. Identify the fundamental dimensions of each free-standing term in eqs. (7.8a–b) and confirm that each has net dimensions of force.

Problem 7.7. Substitute the dimensionless variable of eq. (7.10) into eq. (7.8b) to verify eq. (7.11).

7.1.5 Conserving Energy as the Pendulum Moves

Why? We now turn to a qualitative analysis of the behavior of solutions to the differential equations of motion (7.6) or (7.8). But we start not with the differential equations themselves, but with considerations of energy rooted in the basic physics. When the pendulum is swinging through its arc, it possesses kinetic energy and potential energy. As we will see, each of these energies may vary with position, but both are present and their sum will be a constant.

The kinetic energy, *KE*, is found from a familiar calculation:

$$KE = \frac{1}{2} m(\text{speed})^2. \tag{7.19}$$

The speed can be calculated in the usual way by differentiating the coordinates of mass [eqs. (7.5a–b)] with respect to time, to find that (see Problem 7.8)

$$KE = \frac{1}{2}m\left(l\frac{d\theta(t)}{dt}\right)^2 = \frac{1}{2}mgl\left(\frac{d\theta(\tau)}{d\tau}\right)^2. \tag{7.20}$$

The potential energy of the swinging mass, PE, is measured with respect to a datum through the origin of the coordinates ($x = 0, y = 0$) in another familiar calculation:

$$PE = mgy(t) = mgl(1 - \cos\theta(\tau)). \tag{7.21}$$

Then the total energy, $E(\tau)$, is found by adding eqs. (7.20) and (7.21):

$$E(\tau) = KE + PE = mgl\left[\frac{1}{2}\left(\frac{d\theta(\tau)}{d\tau}\right)^2 + (1 - \cos\theta(\tau))\right]. \tag{7.22}$$

How does the total energy vary with time? A straightforward differentiation shows that

$$\frac{dE(\tau)}{d\tau} = mgl\left[\frac{d^2\theta(\tau)}{d\tau^2} + \sin\theta(\tau)\right]\left(\frac{d\theta(\tau)}{d\tau}\right). \tag{7.23}$$

Equation (7.23) is a remarkable result! The term in the brackets is identical to the tangential equation of motion (7.8b). Thus, two lessons emerge. First, we recover the equation of motion of a system by differentiating its total energy. Second, if $\theta(t)$ is such that the equation of motion is satisfied, then *the total energy is conserved*:

$$\frac{dE(\tau)}{d\tau} = 0 \quad \text{and} \quad E(\tau) = E_0 = \text{constant}. \tag{7.24}$$

Can we determine this constant value of energy, E_0? We can by recognizing that we imparted some energy to the pendulum when we let it start swinging from a rest position θ_0. Thus, the initial potential energy is, in fact, the initial total energy:

$$PE(0) = mgy(0) = mgl(1 - \cos\theta_0) = E_0. \tag{7.25}$$

Problem 7.8. What is the speed of the pendulum mass expressed in polar coordinates? How does that relate to eq. (7.20)?

Problem 7.9. Can eq. (7.22) be simplified for small angles of oscillation? If so, how?

Problem 7.10. How would eq. (7.23) appear after the simplifications of Problem 7.9?

7.1.6 Dissipating Energy as the Pendulum Moves

Our discussion of the pendulum has thus far assumed it to be *ideal* in that no energy was lost. We now extend our model to include the effects of the *damping forces* that arise when motion is resisted by friction or air resistance. Damping or friction forces are generally assumed to be the result of *viscous damping* that is proportional to the speed of the object being analyzed (and slowed by the damping), with a constant of proportionality, c, called the *damping coefficient*. For a viscous damping force we have

Why?

How?

$$F_{\text{damping}} = -c(\text{velocity}), \tag{7.26}$$

where c is a positive constant with dimensions of force per unit velocity or M/T. The minus sign in eq. (7.26) reflects the fact that the viscous damping slows or retards the pendulum motion by opposing it. For the swinging pendulum, the retarding force would act tangentially, so that the friction force would appear in a suitably modified version of the tangential equation of motion (7.8b):

$$ml\frac{d^2\theta}{dt^2} + cl\frac{d\theta}{dt} + mg\sin\theta = 0. \tag{7.27}$$

How does the inclusion of the damping force affect the energy of the pendulum? The forms of the kinetic and potential energies are unchanged by the damping force, so that the total energy can be written as before [eq. (7.22)], except in terms of real time, t:

$$E(t) = \frac{1}{2}ml^2\left(\frac{d\theta(t)}{dt}\right)^2 + mgl(1-\cos\theta(t)). \tag{7.28}$$

The time rate of change of the energy is [again, much as before in eq. (7.23)]

$$\frac{dE(t)}{dt} = \left[ml^2\frac{d^2\theta(t)}{dt^2} + mgl\sin\theta(t)\right]\left(\frac{d\theta(t)}{dt}\right),$$

which in view of eq. (7.27) can be cast as

$$\frac{dE(t)}{dt} = -cl^2 \left(\frac{d\theta(t)}{dt} \right)^2. \tag{7.29}$$

Equation (7.29) shows that the pendulum's total energy steadily decreases with time.

We can take this a step further with the following argument. The energy of an ideal pendulum as it begins from rest is entirely potential energy, and its energy is entirely kinetic when the pendulum swings through its perigee (because the origin of our coordinate system is located at the perigee). Thus, on average, the kinetic and potential energies are approximately the same, even in the presence of all but the most severe damping. To the extent this argument is reasonable, we can approximate the total energy of the pendulum—whether damped or not—as twice the kinetic energy:

$$E(t) \cong ml^2 \left(\frac{d\theta(t)}{dt} \right)^2 \tag{7.30}$$

Now we can eliminate the term $(d\theta/dt)^2$ between eqs. (7.29) and (7.30) to obtain a differential equation for the energy $E(t)$:

$$\frac{dE(t)}{dt} = -(c/m)E(t). \tag{7.31}$$

Note that the dimensions of (c/m) are force per unit velocity per unit mass or $1/\mathsf{T}$. Thus, eq. (7.31) is dimensionally consistent. **Verify?**

Equation (7.31) is also a first-order differential equation with constant coefficients, like the models developed in Chapter 5. Thus, the solution to eq. (7.31) is

$$E(t) = E_0 e^{-(c/m)t}. \tag{7.32}$$

Equation (7.32) shows that the total energy decays exponentially from its initial maximum value, E_0, imparted by the pendulum's initial position. The rate at which the energy decays depends on a *characteristic decay time*, m/c. The characteristic decay time has the proper dimensions, and its precise value (measured in seconds, days, or centuries) will depend on the particular pendulum being modeled. However, we can calculate the energy decay as a function of time measured as a multiple of the characteristic decay time. Table 7.5 shows us that the energy of a damped pendulum is halved in a time equal to $0.69(m/c)$—which is a useful indicator of energy decay time. **Use?**

We note in closing this part of the discussion that we have already learned a lot about the swinging pendulum—and we have determined that information without knowing the specific form of $\theta(t)$ and without

Table 7.5 The decay of the total energy of an oscillating pendulum expressed in multiples of the characteristic decay time, m/c.

Time	Energy
$t = 0$	$E(t) = E_0$
$t = 0.10(m/c)$	$E(t) = 0.905E_0$
$t = 0.69(m/c)$	$E(t) = 0.500E_0$
$t = 1.00(m/c)$	$E(t) = 0.368E_0$
$t = 5.00(m/c)$	$E(t) = 0.007E_0$

solving the differential equations of motion that describe the pendulum's arc. Note, too, that we have not had to distinguish between linear and nonlinear models of the pendulum's behavior, so that the results already obtained—and the methods used to obtain them—are valid for a relatively large class of problems. We will go on to solve the differential equations for the linear model of the pendulum in Section 7.2 and for its nonlinear model in Section 7.5.

7.2 The Freely-Vibrating Pendulum–II: The Linear Model

Why? In Section 7.3 we will come to know the linear model of the pendulum as the ubiquitous *spring-mass oscillator*. But now we want to know: How does a nonlinear model become linear? What do the solutions to linear models look like?

7.2.1 Linearizing the Nonlinear Model

How? We turn a nonlinear model into a linear model by the process of *linearization* in which magnitudes and behaviors are assumed to be sufficiently small in some sense that their products can be neglected. This may not always be possible, and it must be done carefully even when it is possible, because some phenomena are so inherently nonlinear that they can never be linearized. There are nonlinear terms in the pendulum's radial and tangential equations of motion (7.8), which we write here in terms of the dimensionless time, τ, defined in eq. (7.10) and with the nonlinear terms

underlined:

$$T = mg\left[\underline{\left(\frac{d\theta(\tau)}{d\tau}\right)^2} + \underline{\cos\theta(\tau)}\right], \tag{7.33a}$$

and

$$\frac{d^2\theta(\tau)}{d\tau^2} + \underline{\sin\theta(\tau)} = 0. \tag{7.33b}$$

Now let us assume that the angle of the pendulum can be written as

$$\theta(\tau) = \theta_0 f(\tau), \tag{7.34}$$

where $f(\tau)$ is a function whose absolute value is such that $|f(\tau)| \leq 1$. Then

$$\theta_0 = \max|\theta(\tau)|. \tag{7.35}$$

We can identify θ_0 as the *amplitude* of the pendulum's motion that indicates the magnitude of the pendulum's swings. We want to define just how large that amplitude may be, whether it is *small* or *large*, which means that we must provide a reference against which we can meaningfully measure *small* and *large*. We do that by referring back to the Taylor series for the trigonometric functions given in Section 4.1.2, now written in terms of the amplitude θ_0:

$$\sin\theta_0 = \theta_0 - \frac{\theta_0^3}{3!} + \frac{\theta_0^5}{5!} - \frac{\theta_0^7}{7!} + \cdots \cong \theta_0 + \mathrm{O}(\theta_0^3) \tag{7.36a}$$

$$\cos\theta_0 = 1 - \frac{\theta_0^2}{2!} + \frac{\theta_0^4}{4!} - \frac{\theta_0^6}{6!} + \cdots \cong 1 + \mathrm{O}(\theta_0^2) \tag{7.36b}$$

Valid?

In writing these results we have again (see the last two paragraphs of Section 4.1.2) assumed that the angle θ_0, expressed in radians, is a number that is small compared to 1. In eqs. (7.36) we have also introduced the *order* notation, $\mathrm{O}(\theta_0^2)$, that indicates the lowest exponent on the remaining, unwritten terms in the series that represent the difference between a linear approximation, the first term in each series, and the function being approximated. The question of how many terms need to be retained in these series is answered simply: What level of precision is required of the model we are building? It is easy enough to show (see Problems 7.11–7.14) that we can approximate the sine and cosine functions by their *linear approximations* for angles $|\theta_0| \leq \pi/6 = 30°$ as follows:

$$\sin\theta_0 \cong \theta_0 \quad \text{percent error } \sim 5\% \tag{7.37a}$$

$$\cos\theta_0 \cong 1 \quad \text{percent error } \sim 15\% \tag{7.37b}$$

With the approximation (7.37a), we can immediately linearize the tangential equation of motion (7.33b):

$$\frac{d^2\theta(\tau)}{d\tau^2} + \theta(\tau) = 0. \tag{7.38}$$

A similar linearization of the cosine in the radial equation of motion (7.33a) produces the result that

$$T = mg\left[\underline{\left(\frac{d\theta(\tau)}{d\tau}\right)^2} + 1\right], \tag{7.39}$$

which still retains a nonlinear term. However, in the light of eq. (7.34) and the discussion of Section 7.1.4, it is easy enough to show (see Problem 7.15) that the values of $\theta(\tau)$ and its derivatives with respect to τ are all of the same order of magnitude or size. The underlined quadratic term in eq. (7.39) can then be neglected compared to 1, so the linearized model of the pendulum produces a *constant tension*:

$$T \cong mg. \tag{7.40}$$

We close this discussion of linearization by noting that notwithstanding the argument just made about the derivatives of $\theta(\tau)$ with respect to τ, we cannot assume that $\theta(t)$ and its derivatives with respect to the real time, t, are of the same order of magnitude. That assumption is valid *only* with respect to the dimensionless forms discussed.

Problem 7.11. How many terms of the series (7.36a) are needed to calculate $\sin\theta_0$ to a precision of 1% for angles $|\theta_0| \leq \pi/6 = 30°$? To 2%? To 5%?

Problem 7.12. How many terms of the series (7.36b) are needed to calculate $\cos\theta_0$ to a precision of 1% for angles $|\theta_0| \leq \pi/6 = 30°$? To 2%? To 5%?

Problem 7.13. Explain any differences between the answers to Problems 7.11 and 7.12.

Problem 7.14. How does a computer produce values of the "trig" and other transcendental functions?

Problem 7.15. Show (and explain) why the derivatives of eq. (7.34) with respect to τ are all of the same magnitude or size.

7.2.2 The Differential Equation $md^2x/dt^2 + kx = 0$

How do we determine the function $\theta(\tau)$ that satisfies and thus solves eq. (7.38)? First, to be more general, let us return that equation to its dimensional form, **How?**

$$ml\frac{d^2\theta(t)}{dt^2} + mg\theta(t) = 0. \tag{7.41}$$

To be still more general, we write eq. (7.41) in the equivalent form (see Problem 7.16) of

$$m\frac{d^2x(t)}{dt^2} + kx(t) = 0, \tag{7.42}$$

which is the classical equation for a simple *spring-mass oscillator*, which we will begin to discuss in some detail in Section 7.3 and with great generality in Chapter 8. In the meantime, we can safely refer to m as the (constant) *mass* of the oscillator, k as its (constant) *stiffness*, and $x(t)$ as its displacement (or movement or deflection). It is clear that if we can solve eq. (7.42) we obtain a solution to eq. (7.38).

Equation (7.42) is a *homogeneous, second-order, linear* differential equation that has *constant coefficients*, k and m. Guided by the discussion of Section 5.2.2, we assume a solution to eq. (7.42) in the form

$$x(t) = Ce^{\lambda t}, \tag{7.43}$$

which when substituted into eq. (7.42) leads to the *characteristic equation* that defines the constant, λ,

$$m\lambda^2 + k = 0. \tag{7.44}$$

Equation (7.44) has two solutions,

$$\lambda_{1,2} = \pm\sqrt{-1}\sqrt{\frac{k}{m}} \equiv \pm j\omega_0, \tag{7.45}$$

where we have now noted that $j = \sqrt{-1}$ and have redefined the scaling factor, ω_0, as

$$\omega_0 \equiv \sqrt{\frac{k}{m}} \tag{7.46}$$

Since eq. (7.42) is of second order, we expect that it will have two solutions, each corresponding to the two values of λ defined by eq. (7.45):

$$x(t) = C_1e^{j\omega_0 t} + C_2e^{-j\omega_0 t}. \tag{7.47}$$

These general forms of the homogeneous solutions are quite valid. However, guided by the "most remarkable formula" presented in Section 4.9,

we can (see Problem 7.17) rewrite the solution (7.47) in terms of the standard trigonometric functions:

$$x(t) = B_1 \cos \omega_0 t + B_2 \sin \omega_0 t, \tag{7.48}$$

where B_1 and B_2 are two arbitrary constants that are entirely equivalent to the constants in eq. (7.47). It is also easily verified by direct substitution (Problem 7.18) that eq. (7.48) is a solution to eq. (7.42).

Equation (7.48) is called the *homogeneous solution* of eq. (7.42) because it solves a differential equation that has no forcing function on its right-hand
Use? side. Equation (7.48) is also called the *transient solution* because it actually represents the initial conditions that initiate the pendulum's motion. Thus, if $x(0) = x_0$ and $dx(0)/dt = \dot{x}_0$, it is easily shown (Problem 7.19) that

$$x(t) = x_0 \cos \omega_0 t + \frac{\dot{x}_0}{\omega_0} \sin \omega_0 t, \tag{7.49}$$

As we will further describe in the next section, the motion described by eq. (7.49) is periodic and would go on indefinitely for an ideal pendulum that experiences no damping. However, for a damped pendulum, this initial motion will be damped out, which is why it is called the "transient solution."

Problem 7.16. What is the effective spring stiffness, k, for the simple pendulum? Are its dimensions proper, for the pendulum itself and as a stiffness?

Problem 7.17. Use "the most remarkable formula" in mathematics to show how eq. (7.47) becomes eq. (7.48).

Problem 7.18. Substitute the solution (7.48) into eq. (7.42) and confirm that it is a correct solution.

Problem 7.19. Determine the constants, B_1 and B_2, in eq. (7.48) for the initial conditions $x(0) = x_0$ and $dx(0)/dt = \dot{x}_0$.

7.2.3 The Linear Model

Use? Returning now to the linear model of the pendulum, we can straightforwardly cast eq. (7.49) into the dimensionless notation of the pendulum (see Problem 7.20):

$$\theta(\tau) = \theta_0 \cos \tau + \dot{\theta}_0 \sin \tau, \tag{7.50}$$

where we can now identify θ_0 and $\dot{\theta}_0$ as, respectively, the initial location and the initial speed with which the pendulum is set in motion. These initial parameters are entirely independent, so that they can be specified separately. Thus, to drop a pendulum from a fixed angle, θ_0, but with no initial speed, the transient solution would be

$$\theta(\tau) = \theta_0 \cos \tau. \tag{7.51}$$

On the other hand, to launch the pendulum from the origin, $\theta_0 = 0$, with a specified initial speed, $\dot{\theta}_0$, the transient solution would take the form

$$\theta(\tau) = \dot{\theta}_0 \sin \tau. \tag{7.52}$$

Since we are solving a linear problem, superposition applies (see Section 1.3.4), and the general solution (7.50) is simply the sum of the two solutions (7.51) and (7.52).

In Section 4.9 we noted that the elementary trigonometric functions are periodic: the functions $\sin \tau$ and $\cos \tau$ have the same value when their arguments are increased by 2π, that is,

$$\cos(\tau + 2\pi) = \cos \tau \quad \text{and} \quad \sin(\tau + 2\pi) = \sin \tau. \tag{7.53}$$

In physical time t, then, the value of $\theta(t)$ repeats at time intervals such that

$$t = \frac{2\pi n}{\omega_0} = nT_0, \quad n = 1, 2, 3, \ldots . \tag{7.54}$$

Hence, T_0 is (again) the period of the pendulum motion and ω_0 its circular frequency, measured in radians per unit time. We can also define a frequency f_0 with units of $(\text{time})^{-1}$ or hertz (Hz), named after a famous acoustician, Heinrich Rudolf Hertz (1857–1894):

$$f_0 = \frac{1}{T_0} = \frac{\omega_0}{2\pi} \tag{7.55}$$

One last observation about the results just described: the period of the vibrating pendulum, T_0, depends *only* on the physical properties of the pendulum and *not at all* on the amplitude of the oscillation. The uncoupling of the amplitude from the period, like the applicability of the principle of superposition, is another defining characteristic of linear models of vibration.

Problem 7.20. Show how the solution (7.49) becomes the solution (7.50) for initial conditions $\theta(0) = \theta_0$ and $d\theta(0)/dt = \dot{\theta}_0$.

7.3 The Spring-Mass Oscillator–I: Physical Interpretations

Why? How? We now explore some physical interpretations of the linear model just developed. The more general form, eq. (7.42), is an equation of equilibrium, which means that its physical dimensions are of force or $\mathsf{F} = \mathsf{ML}/\mathsf{T}^2$. Since $x(t)$ is the oscillator displacement and has the dimensions of length or L, the stiffness, k, must have the dimensions of force per unit length or F/L. Thus, the equation (7.42) represents a balance of an inertial force with a spring force. Further, our everyday experience with springs confirms Hooke's law, which states that a spring exerts a restoring force that is directly proportional to the amount that it is stretched or compressed, that is,

$$F_{\text{spring}} = kx(t). \tag{7.56}$$

Note that the sign of the spring force changes with the sign of the displacement, so that extending a spring ($x > 0$) produces a positive, tensile force that tends to return it to its original length, while compressing the spring ($x < 0$) produces a negative, compressive force that also tends to restore the spring to its original length.

How does this work for the pendulum? A slight rewriting of eq. (7.41) shows that

$$m\frac{d^2\theta(t)}{dt^2} + \left[k = \frac{mg}{l}\right]\theta(t) = 0. \tag{7.57}$$

Thus, we see that the pull of gravity acts just like a spring, exerting a larger restoring force as the pendulum angle increases.

Another reflection of this behavior can be seen if we examine the energy of the spring-mass oscillator. If we multiply eq. (7.42) by the oscillator speed, $dx(t)/dt$, we find

$$\left[m\frac{d^2x(t)}{dt^2} + kx(t)\right]\frac{dx(t)}{dt} = 0. \tag{7.58}$$

Now, both terms in eq. (7.58) are total derivatives. Therefore, we can integrate this equation to obtain

$$\frac{1}{2}m\left(\frac{dx(t)}{dt}\right)^2 + \frac{1}{2}k(x(t))^2 = E_0. \tag{7.59}$$

Thus, by inverting the process by which we identified the pendulum's total energy in Section 7.1.5, we have here derived the energy of the spring-mass oscillator and showed that it, too, is the sum of the kinetic and potential

energies. Further, as we know from the pendulum and can easily demonstrate (see Problems 7.21–7.23) with the solution (7.49), the energy moves back and forth from being entirely kinetic energy when the pendulum is at its perigee to a position when it is entirely potential energy, that is, at its maximum amplitude. This means that each of the two elements in the spring-mass system acts as an *energy-storage element*: the spring stores (and releases) potential energy, while the mass stores (and gives up) kinetic energy.

Predict?
Use?

Problem 7.21. Calculate the *kinetic* energy of a spring-mass oscillator released from a rest position $x(0) = x_0$ initially and at time intervals, $T_0/4$, $T_0/2$, $3T_0/4$, and T_0.

Problem 7.22. Calculate the *potential* energy of a spring-mass oscillator released from a rest position $x(0) = x_0$ initially and at time intervals, $T_0/4$, $T_0/2$, $3T_0/4$, and T_0.

Problem 7.23. What fractions of the total energy are the kinetic and potential energies at time intervals, $T_0/4$, $T_0/2$, $3T_0/4$, and T_0? (*Hint*: Use the results of Problems 7.21 and 7.22!)

7.4 Stability of a Two-Mass Pendulum

In our brief review of the elementary transcendental functions (in Section 4.9), we saw that trigonometric and hyperbolic functions are closely related. The arithmetic difference between the two is traceable to the j factor in the argument of the exponential function. Their behaviors differ as well, with the trigonometric functions showing bounded periodicity and the hyperbolic functions showing exponential growth or decay. The change from periodic to exponential arithmetic behavior typically signals a change in physical behavior from a stable, bounded configuration to unstable, unbounded exponential growth. The transition from bounded trigonometric behavior to unbounded exponential behavior occurs when a model parameter passes through a critical value. We will illustrate this transitional behavior for a two-mass pendulum.

Why?

Consider the vertically-arrayed dumbbell shown in Figure 7.3. If set absolutely still in a perfectly vertical alignment, it conceivably could remain in that precarious position. However, in the normal course of events, if the dumbbell is let go and starts to swing, we would expect that its final position—and its behavior in arriving at that position—will depend very

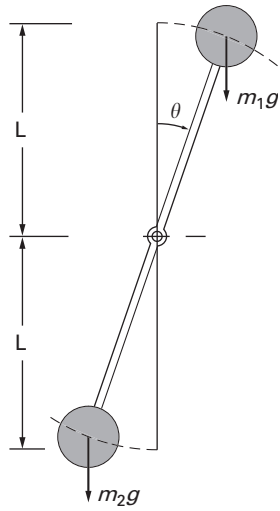


Figure 7.3 A schematic of a dumbbell, a two-mass pendulum. Its initial state has a mass, m_1 on top, and a mass, m_2 on the bottom. The stability of this state is dependent on the relative magnitudes of the two masses.

much on the relative sizes of the masses, m_1 and m_2. If $m_1 < m_2$, we would expect that the dumbbell would oscillate just like a simple pendulum, around its present position. On the other hand, if $m_1 > m_2$, we would expect that the two-mass pendulum would swing downward until the masses settled into an inverted position, with m_2 at the top and m_1 at the bottom. Thus, this is a stability problem, with the operative question being: Is the configuration shown in Figure 7.3 a stable configuration?

Predict?

How? To answer this question we must model the free vibration of the two-mass pendulum.We can build that model by extending the elementary pendulum model: First, we write the total energy for the dumbbell and then we differentiate that total energy to derive the equation of motion. Note that while there are two separate masses, only one degree of freedom, the angle, $\theta(t)$, is needed to specify the positions of both masses. Thus, taking our cue from eq. (7.20), the kinetic energy for the dumbbell is

$$KE_2 = \frac{1}{2}(m_1 + m_2)\left(l\frac{d\theta(t)}{dt}\right)^2. \tag{7.60}$$

The potential energy of the swinging mass, PE, is measured with respect to a datum through the origin of the coordinates ($x = 0, y = 0$) in another familiar calculation:

$$PE_2 = m_1 g y_1(t) - m_2 g y_2(t) = -(m_1 - m_2) g l (1 - \cos\theta(t)). \quad (7.61)$$

For a linear two-mass pendulum model, we can approximate the potential energy as

$$PE_2 \cong -\frac{1}{2}(m_1 - m_2) g l (\theta(t))^2. \quad (7.62)$$

The total energy, $E_2(t)$, is found by adding eqs. (7.60) and (7.62):

$$E_2(t) = \frac{1}{2}(m_1 + m_2)\left(l\frac{d\theta(t)}{dt}\right)^2 - \frac{1}{2}(m_1 - m_2) g l (\theta(t))^2. \quad (7.63)$$

Then we can derive the equation of motion for the dumbbell by differentiating eq. (7.63) with respect to time,

$$\frac{dE_2(t)}{dt} = \left[(m_1 + m_2) l^2 \frac{d\theta^2(t)}{dt^2} - (m_1 - m_2) g l \theta(t)\right]\left(\frac{d\theta(t)}{dt}\right), \quad (7.64)$$

from which it follows that

$$(m_1 + m_2) l \frac{d\theta^2(t)}{dt^2} - (m_1 - m_2) g \theta(t) = 0,$$

or

$$\frac{d\theta^2(t)}{dt^2} + \frac{(m_2 - m_1)}{(m_1 + m_2)}\left(\frac{g}{l}\right)\theta(t) = 0. \quad (7.65)$$

Equation (7.65) is the same homogeneous, second-order, linear differential equation with constant coefficients that we solved before [i.e., eq. (7.42)] with the solution

$$\theta(t) = Ce^{\lambda t}, \quad (7.66)$$

which leads to a characteristic equation for the constant, λ, that has two solutions,

$$\lambda_{1,2} = \pm j\sqrt{\frac{(m_2 - m_1)}{(m_1 + m_2)}\left(\frac{g}{l}\right)}. \quad (7.67)$$

Now the most interesting feature of eq. (7.67) is that the very nature of the roots,$\lambda_{1,2}$, changes according to the relative size of the two masses. For the case $m_2 > m_1$, the roots are purely imaginary, so the dumbbell will simply oscillate around its initial position (i.e., m_1 at the top and m_2 at the

Verified?

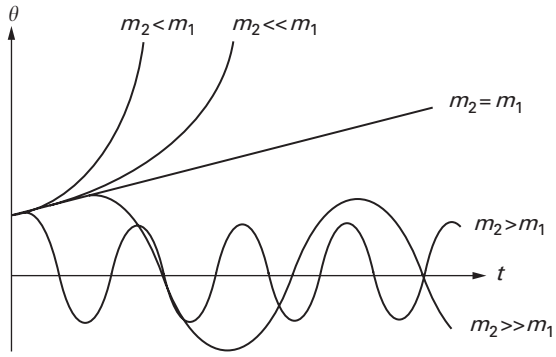


Figure 7.4 A sketch of the solutions to the linearized equations of motion of a dumbbell, a two-mass pendulum. These solutions are periodic when the initial configuration is stable ($m_2 > m_1$) and are exponential when the initial state is unstable ($m_2 < m_1$). The case $m_2 = m_1$ is a *critical point* that defines the border between the stable and unstable states.

bottom). On the other hand, if $m_2 < m_1$, the roots (7.67) become two *real* roots:

$$\lambda_{1,2} = \pm j\sqrt{\frac{-(m_1 - m_2)}{(m_1 + m_2)}\left(\frac{g}{l}\right)} = \mp\sqrt{\frac{(m_1 - m_2)}{(m_1 + m_2)}\left(\frac{g}{l}\right)}. \tag{7.68}$$

Equation (7.68) mean that the two homogeneous solutions for $m_2 < m_1$ are exponentials, one decaying to zero, the other growing without bound. Thus, the case $m_2 < m_1$ represents an instance where the initial configuration is unstable, a finding that accords with our intuition of what would happen if we tried to stand a top-heavy dumbbell on its lighter end. Figure 7.4 shows a plot of schematic solutions for both real and imaginary values of the roots, for both the periodic and exponential solutions. The case $m_2 = m_1$ is a *critical point* that defines the border between a stable initial configuration ($m_2 > m_1$) and an unstable initial state ($m_2 < m_1$).

Thus, we have seen here an instance where changes in the parameters produce changes in the mathematical behavior of the model, which is a signal that different physical behavior is to be expected. An often-asked question in engineering and the physical sciences is whether a system's parameters support its bounded oscillation about its equilibrium position, or whether its instability is possible or even certain. We will see an instance of the former in a nonlinear biological model in Section 7.6.

Use?

7.5 The Freely-Vibrating Pendulum–III: The Nonlinear Model

Why?

We now return to the classical single pendulum to illustrate one of the most elegant solutions in applied mathematics and to show how an approximation to the nonlinear results can be obtained with some of the series introduced in Chapter 4. We begin with eq. (7.22) for the total energy of the pendulum, while also noting that the energy is a constant (eq. (7.24)) for this conservative system:

How?

$$\frac{1}{2}\left(\frac{d\theta(\tau)}{d\tau}\right)^2 + (1 - \cos\theta(\tau)) = \frac{E_0}{mgl}. \tag{7.69}$$

Now for a pendulum released from the resting position, $\theta(0) = \theta_0$, we can determine (see Problem 7.24) the constant, E_0, so that

$$\left(\frac{d\theta(\tau)}{d\tau}\right)^2 + 2(1 - \cos\theta(\tau)) = 2(1 - \cos\theta_0). \tag{7.70}$$

With the aid of a standard double-angle formula, we can rewrite eq. (7.70) as

$$\left(\frac{d\theta(\tau)}{d\tau}\right)^2 = 4\sin^2\frac{\theta_0}{2} - 4\sin^2\frac{\theta(\tau)}{2}. \tag{7.71}$$

We now introduce a constant,

$$p \equiv \sin\frac{\theta_0}{2}, \tag{7.72}$$

and a change of variable to a new angle, ϕ,

$$\sin\frac{\theta(\tau)}{2} \equiv \sin\frac{\theta_0}{2}\sin\phi = p\sin\phi, \tag{7.73}$$

so that the energy equation (7.71) can be written as

$$\left(\frac{d\theta(\tau)}{d\tau}\right)^2 = 4p^2\cos^2\phi. \tag{7.74}$$

Equation (7.74) does look neater and more elegant, but it has two dependent variables, θ and ϕ. However, we can differentiate eq. (7.73) to show that

$$\frac{1}{2}\cos\frac{\theta}{2}d\theta = p\cos\phi\, d\phi,$$

or

$$d\theta = 2p\frac{\cos\phi}{\cos\frac{\theta}{2}}d\phi = 2p\frac{\cos\phi}{\sqrt{1-p^2\sin^2\phi}}d\phi, \tag{7.75}$$

which allows us to rewrite eq. (7.74) as

$$d\tau = -\frac{d\phi}{\sqrt{1-p^2\sin^2\phi}}, \tag{7.76}$$

with a minus sign [for the square root of eq. (7.74)] that arises because $\theta(\tau)$ is measured positive counter-clockwise from the pendulum's perigee. Thus, for $\theta(0) = \theta_0 > 0$, we have both $d\theta/d\tau$ and $d\phi/d\tau < 0$.

Equation (7.76) can be formally integrated, but we must exercise care in choosing the limits. The period of the nonlinear model, $\widetilde{T}_0$, differs from the linear period, $T_0 = 2\pi/\omega_0$. In terms of the dimensionless time variable, $\tau = t\omega_0$, an integration over the first quarter of the period means that $0 \le \tau \le (\widetilde{T}_0\omega_0/4 = \pi\widetilde{T}_0/2T_0)$, and that $\pi/2 \le \phi \le 0$:

$$\frac{\widetilde{T}_0}{T_0} = -\frac{2}{\pi}\int_{\pi/2}^{0}\frac{d\phi}{\sqrt{1-p^2\sin^2\phi}} = \frac{2}{\pi}\int_{0}^{\pi/2}\frac{d\phi}{\sqrt{1-p^2\sin^2\phi}}. \tag{7.77}$$

The integral on the right-hand side of eq. (7.77) is an *elliptic integral* (of the first kind), for which there are published tables of numerical values as a function of p. Thus, the tabulated values of the integral make it possible to calculate how the nonlinear period varies with p—which means how the nonlinear period, $\widetilde{T}_0$, varies with the initial amplitude of the pendulum, θ_0 (recall the definition of p in eq. 7.73)). This confirms what we said when we discussed the experimental data presented in Section 7.1.1: The period of oscillation of the pendulum does depend on its initial position or amplitude.

What happens with the linear model? The answer is that for very small values of θ_0, and thus of p, we make the same kind of approximation of the radicand in eq. (7.77) that we did in eqs. (7.37a–b): We say $1-p^2\sin^2\phi \cong 1$, in which case we recover the linear result, $\widetilde{T}_0 \cong T_0$.

The reduction to the linear case also suggests that we apply the binomial expansion (4.24) to the radicand in eq. (7.77) for small values of p:

$$\frac{\widetilde{T}_0}{T_0} = \frac{2}{\pi}\int_{0}^{\pi/2}\frac{d\phi}{\sqrt{1-p^2\sin^2\phi}} \cong \frac{2}{\pi}\int_{0}^{\pi/2}(1+\frac{p^2}{2}\sin^2\phi)d\phi, \tag{7.78}$$

which, after integration and another application of the small-angle approximation, yields

$$\frac{\widetilde{T}_0}{T_0} \cong 1 + \frac{p^2}{4} = 1 + \frac{1}{4}\sin^2\frac{\theta_0}{2} \cong 1 + \frac{\theta_0^2}{16}. \tag{7.79}$$

Once again we see here the dependence of the period on the amplitude, and the results predicted from eq. (7.79) can be compared both to the exact result given in eq. (7.77) and to the experimental data given in Table 7.1 (see Problems 7.25 and 7.26).

Problem 7.24. Determine the value of the constant energy, E_0, in eq. (7.69) for (a) a pendulum released from a resting position $\theta(0) = \theta_0$, and (b) for a pendulum given an initial speed $\dot{\theta}_0$ while hanging vertically ($\theta(0) = 0$).

Problem 7.25. Complete the integration of the last form of eq. (7.78) and confirm the first equality in eq. (7.79).

Problem 7.26. Use tabulated values of the elliptical integral of the first kind (eq. (7.77)) to determine the values of $\widetilde{T}_0/T_0$ for the values of θ_0 used in Table 7.1.

Problem 7.27. Compare and contrast the values found in the last column of Table 7.1 with the results found in Problem 7.26.

7.6 Modeling the Population Growth of Coupled Species

In Section 5.6 we introduced the logistic growth model that shows how, in a nonlinear fashion, the exponential growth of a single population or species can be bounded. What happens if there are *two* species that interact with each other? The Lotka-Volterra model of population growth provides an answer to this question, and in so doing it uses many of the modeling ideas developed above for the pendulum. The two-species model is of particular interest to biologists, with one species typically playing *host* to the second, *parasitic* population.

Why?

The bounding effect of the single-population logistic model is produced by the inclusion of the term $-\lambda^2 N^2$ in the population balance equation

How?

(see eqs. (5.48) and (5.50)). This term describes the *inhibition* of the popu-
Assume? lation's growth. We start with two populations, the host (or prey) $H(t)$ and the parasite (or predator) $P(t)$, and we assume that the growth of *each* population is inhibited by the size of the *other* population. Thus, in the place of eq. (5.50) for a single population, we start with

$$\frac{dH(t)}{dt} = \lambda_{\mathrm{H}} H(t) \left(1 - \frac{P(t)}{P_{\mathrm{e}}}\right), \tag{7.80a}$$

and

$$\frac{dP(t)}{dt} = -\lambda_{\mathrm{P}} P(t) \left(1 - \frac{H(t)}{H_{\mathrm{e}}}\right). \tag{7.80b}$$

The positive constants, λ_{H} and λ_{P}, represent the uninhibited growth and decay rates, respectively, of the host and parasite populations, and each has physical dimensions of $(\text{time})^{-1}$. The population values, H_{e} and P_{e}, correspond to the equilibrium values of the two populations, the point at which the population rates, dH/dt and dP/dt, both vanish and the two populations are in *static* equilibrium with each other.

Verified? Equation (7.80a) shows that the parasite population reduces the growth rate of the host population, which is what parasites or predators do. On the other hand, the presence of the hosts in eq. (7.80b) slows the decline of the parasite population (for $H(t) < H_{\mathrm{e}}$), since there are fewer sources of sustenance when there are fewer hosts or prey. Thus, eqs. (7.80a–b)—which are variously known as the *Lotka-Volterra* equations or the *predator-prey* or *parasite-host* equations—do seem to be intuitively correct.

Further, while eqs. (7.80) resemble the single-population logistical model (5.50), there is one interesting and important distinction. While the single-population model (5.50) incorporated a maximum population $N_{\max}$, the predator-prey model refers to equilibrium populations that may be exceeded, which means that there could be a *change in the arithmetic signs* of the right-hand sides of eqs. (7.80a–b). For example, when $H(t) > H_{\mathrm{e}}$, the parasite decay rate turns into a growth rate. This suggests that the population sizes might *oscillate* or vibrate about their equilibrium sizes.

Equations (7.80) are *coupled*, *nonlinear*, ordinary differential equations. They are *coupled* because the dependent variables, $H(t)$ and $P(t)$, appear in both equations, and *nonlinear* because of the products of $H(t)$ and $P(t)$. No explicit solutions for $H(t)$ and $P(t)$ are known to exist for these nonlinear equations. However, as with the pendulum, we can use other means to extract a great deal of information.

7.6.1 Qualitative Solution for the Nonlinear Model

How?

While we cannot explicitly integrate eqs. (7.80a–b), we can divide one by the other and obtain a form that is independent of the independent variable t:

$$\frac{dH}{dP} = -\frac{\lambda_H}{\lambda_P}\frac{(1 - P/P_e)H}{(1 - H/H_e)P}. \tag{7.81}$$

If the fractions in eq. (7.81) are cleared and the populations are rendered dimensionless with respect to their equilibrium populations, we find

$$\frac{1}{\lambda_H}\left(\frac{1}{H/H_e} - 1\right) d(H/H_e) + \frac{1}{\lambda_H}\left(\frac{1}{P/P_e} - 1\right) d(P/P_e) = 0. \tag{7.82}$$

Equation (7.82) can be straightforwardly integrated to yield

$$\frac{1}{\lambda_H}\left(\ln\frac{H}{H_e} - \frac{H}{H_e}\right) + \frac{1}{\lambda_P}\left(\ln\frac{P}{P_e} - \frac{P}{P_e}\right) = \text{constant}. \tag{7.83}$$

When plotted on the set of axes comprising the (H, P) space, eq. (7.83) represents a family of closed curves "centered" around the equilibrium point (H_e, P_e), as shown in Figure 7.5. Each member of the family of curves corresponds to a different value of the constant in eq. (7.83), with the area enclosed by the curve increasing with the value of the constant. We also

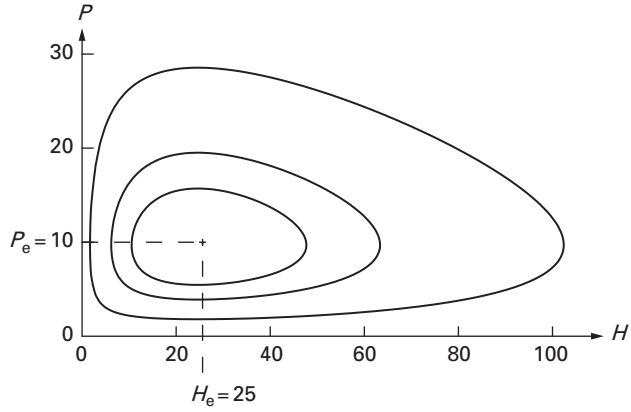


Figure 7.5 Three curves that illustrate the family of curves represented by eq. (7.83). Here $\lambda_H = 1.00$ per unit time, $\lambda_P = 0.50$ per unit time, $P_e = 10$ and $H_e = 25$. Note the equilibrium point, as well as the horizontal and vertical flat spots discussed previously, as well as the elliptical nature of the curves closest to the equilibrium point (Pielou, 1969).

note flat spots at abscissa values of $H = H_e$ that correspond to the vanishing of the slope dP/dH (or where $dP/dt = 0$ in eq. (7.80b)). Similarly, we note vertical tangents ("vertical flat spots") at ordinate values of $P = P_e$ that correspond to the slope dP/dH becoming infinite (or where $dH/dt = 0$ in eq. (7.80a)). More importantly, for given values of the constant, we can trace the magnitudes of the two populations and can thus examine how predator and prey or parasite and host interact.

7.6.2 Oscillatory Solution for the Linearized Model

How? A further examination of the curves in Figure 7.5 also shows that those nearest the equilibrium point are nearly elliptical in shape. Thus, let us write the values of $H(t)$ and $P(t)$ in the forms

$$\frac{H}{H_e} = 1 + \frac{h}{H_e} \quad \text{and} \quad \frac{P}{P_e} = 1 + \frac{p}{P_e}. \tag{7.84}$$

Assume? Let us further assume that the values of $h(t)$ and $p(t)$ are small compared to their respective equilibrium values of the populations:

$$\frac{h}{H_e} \ll 1 \quad \text{and} \quad \frac{p}{P_e} \ll 1. \tag{7.85}$$

Equations (7.84) and (7.85) provide a basis for generating binomial expansions of the natural logarithms in eq. (7.83). If that's done, the result is that to $O(h, p)^3$, eq. (7.83) becomes (see Problem 7.28):

$$\frac{1}{\lambda_H}\left(\frac{h}{H_e}\right)^2 + \frac{1}{\lambda_P}\left(\frac{p}{P_e}\right)^2 = \text{constant}. \tag{7.86}$$

Equation (7.86) is clearly that of an ellipse and so confirms the observation made above about the shapes of the closed curves near equilibrium.

What happens when we substitute eq. (7.84) into our original model equations (7.80a–b)? We would find that

$$\frac{dh(t)}{dt} = -\lambda_H H_e \left(1 + \frac{h(t)}{H_e}\right)\left(\frac{p(t)}{P_e}\right), \tag{7.87a}$$

and

$$\frac{dp(t)}{dt} = \lambda_P P_e \left(1 + \frac{p(t)}{P_e}\right)\left(\frac{h(t)}{H_e}\right). \tag{7.87b}$$

If we now linearize eqs. (7.87a–b) to keep only linear terms on their right-hand sides, we get

$$\frac{dh(t)}{dt} \cong -\lambda_H H_e \left(\frac{p(t)}{P_e}\right), \tag{7.88a}$$

and

$$\frac{dp(t)}{dt} \cong \lambda_P P_e \left(\frac{h(t)}{H_e}\right). \tag{7.88b}$$

We can now eliminate either of the functions $h(t)$ or $p(t)$ between eqs. (7.88a–b) to show that they each satisfy the same equation (see Problems 7.29 and 7.30):

$$\frac{d^2h(t)}{dt^2} + \lambda_H \lambda_P h(t) = 0, \tag{7.89a}$$

and

$$\frac{d^2p(t)}{dt^2} + \lambda_H \lambda_P p(t) = 0. \tag{7.89b}$$

Equations (7.89a–b) are the equations of simple harmonic oscillators! Thus, $h(t)$ or $p(t)$ represent small oscillations about the equilibrium position, a stable result. In fact, it is not hard to show (see Problems 7.31–7.33) that a solution to eqs. (7.88) or (7.89) is

$$\begin{aligned} p(t) &= p_0 \cos\sqrt{\lambda_H \lambda_P}\, t \\ h(t) &= -p_0 \sqrt{\frac{\lambda_H}{\lambda_P}} \left(\frac{H_e}{P_e}\right) \sin\sqrt{\lambda_H \lambda_P}\, t. \end{aligned} \tag{7.90}$$

where p_0 is a constant that will be determined by the initial conditions. In terms of the original host and parasite populations, the solution (7.90) appears as

$$\begin{aligned} P(t) &= P_e \left(1 + \frac{p_0}{P_e} \cos\sqrt{\lambda_H \lambda_P}\, t\right) \\ H(t) &= H_e \left(1 - \frac{p_0}{P_e} \sqrt{\frac{\lambda_H}{\lambda_P}} \sin\sqrt{\lambda_H \lambda_P}\, t\right). \end{aligned} \tag{7.91}$$

This result makes explicit the oscillation of the host and parasite populations around the equilibrium point. Moreover, the oscillations for both host and parasite occur at exactly the same natural frequency, $T_0 = 2\pi/\sqrt{\lambda_H \lambda_P}$. **Verified?**

It is worth noting that a potential instability phenomenon is embedded in the solutions (7.90) and (7.91). Recall that the uninhibited growth and decay rates, λ_H and λ_P, were assumed to be positive constants. If one of them were negative, that is, if the host population was declining or the parasite population growing, the outcome would be far different (see Problems 7.34 and 35). **Predict? Use?**

We close this discussion by noting that we have gained a great deal of information about host-parasite population systems without having

obtained explicit solutions. We used both energy and small perturbation formulations to derive considerable *qualitative* understanding of the behavior of prey and predator. These qualitative approaches allowed us to identify the equilibrium point, the family of closed-curve solutions, the elliptical shapes of those curves in the neighborhood of equilibrium, and the periodic vibration of the two populations about equilibrium.

Problem 7.28. Use eqs. (7.84) and (7.85) to generate binomial expansions of the natural logarithms in eq. (7.83) and to confirm eq. (7.86) to $O(h, p)^3$.

Problem 7.29. Substitute $p(t)$ from eq. (7.88a) into eq. (7.88b) to obtain eq. (7.89a).

Problem 7.30. Substitute $h(t)$ from eq. (7.88b) into eq. (7.88a) to obtain eq. (7.89b).

Problem 7.31. Guided by the general solution (7.48), determine the solutions to eqs. (7.88) or (7.89) that satisfy initial conditions $p(0) = p_0$ and $dp(0)/dt = 0$.

Problem 7.32. What initial conditions are satisfied by $h(t)$ in the solution of Problem 7.31? Could they have been specified differently or separately?

Problem 7.33. What are the initial values of the populations $H(t)$ and $P(t)$ corresponding to the solution (7.90)?

Problem 7.34. What does it mean for the rate λ_P to become a negative constant?

Problem 7.35. Show how the solution (7.90) changes if λ_P is a negative constant.

7.7 Summary

In this chapter we have used the classical pendulum to show a mathematical model was derived, how it was grounded in and verified against experimental results, and how we could obtain qualitative information about its behavior. We also demonstrated the behavior of linear oscillators in several domains, and drew some distinctions between the behaviors exhibited by linear and nonlinear models. In so doing, we used concepts of linearity, dimensional consistency, scaling, and some basic ideas of second-order differential equations.

In terms of the behavior of the pendulum itself, we have shown how the period of the linear model depends only on the pendulum's properties and not on its amplitude of vibration, as is the case for nonlinear models

wherein the amplitude is large. We also developed an elegant exact solution for the period of a pendulum and related it to the linear model. We also showed, for both the two-mass pendulum and a predator-prey population system, how the period of the vibrating system is sensitive to properties of that system—especially for the two-mass pendulum, for which instability occurs for certain combinations of masses.

7.8 References

M. Abramowitz and I. A. Stegun (Eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Applied Mathematical Series 55, National Bureau of Standards, Washington, D.C., 1964.

R. E. D. Bishop, *Vibration*, Cambridge University Press, Cambridge, UK, 1965.

R. E. D. Bishop and D. C. Johnson, *The Mechanics of Vibration*, Cambridge University Press, Cambridge, UK, 1960.

M. Braun, *Differential Equations and Their Applications: Shorter Version*, Springer-Verlag, New York, 1978.

R. H. Cannon, Jr., *Dynamics of Physical Systems*, McGraw-Hill, New York, 1967.

P. D. Cha, J. J. Rosenberg, and C. L. Dym, *Fundamentals of Modeling and Analyzing Engineering Systems*, Cambridge University Press, New York, 2000.

J. G. Croll and A. C. Walker, *Elements of Structural Stability*, John Wiley, New York, 1972.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

F. Fahy, *Sound and Structural Vibration*, Academic Press, London, 1985.

M. Farkas, *Dynamical Models in Biology*, Academic Press, San Diego, CA, 2001.

R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vols. I and II, Addison-Wesley, Reading, MA, 1963.

B. R. Gossick, *Hamilton's Principle and Physical Systems*, Academic Press, New York, 1967.

R. Haberman, *Mathematical Models*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

G. W. Housner and D. E. Hudson, *Applied Mechanics: Dynamics*, Von Nostrand-Reinhold, New York, 1959.

E. C. Pielou, *An Introduction to Mathematical Ecology*, Wiley Interscience, New York, 1969.

B. O. Pierce and R. M. Foster, *A Short Table of Integrals*, 4th Edition, Ginn and Company, Boston, 1956.
J. M. Smith, *Mathematical Ideas in Biology*, Cambridge University Press, London and New York, 1968.
G. W. Swenson, *Principles of Modern Acoustics*, Boston Technical Publishers, Cambridge, MA, 1965.
J. W. Tongue, *Principles of Vibration*, 2nd Edition, Oxford University Press, New York, 2001.
M. R. Wehr and J. A. Richards, Jr., *Physics of the Atom*, Addison-Wesley, Reading, MA, 1960.
R. M. Whitmer, *Electromagnetics*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

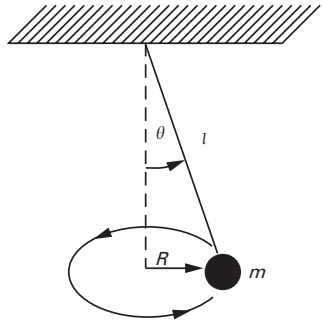
7.9 Problems

7.36. Use eq. (7.79) to determine the maximum angle, θ_0, such that the ratio, $\widetilde{T}_0/T_0$, does not exceed 1.005.

7.37. (a) Determine which variables affect the period of free vibration of the conical pendulum shown below from the accompanying table of data.

(b) Determine which variables affect the period of free vibration of the conical pendulum shown below using dimensional analysis.

		Period of Revolution (sec)					
θ	m	$l_1 = 1$ m			$l_2 = 3$ m		
θ_1	m_1	2.09	2.09	2.10	3.45	3.40	3.48
	m_2	2.07	2.08	2.08	3.46	3.44	3.44
θ_2	m_1	1.95	1.98	1.94	3.37	3.40	3.38
	m_2	1.96	1.93	1.95	3.36	3.38	3.35
θ_3	m_1	1.87	1.87	1.88	3.24	3.29	3.27
	m_2	1.86	1.85	1.87	3.22	3.25	3.21



7.38. Confirm the answer to Problem 7.37 (b) by deriving the equations of motion for a conical pendulum.

7.39. A uniform rod or stick is supported by and swings from a pivot at one end. The mass of this swinging rod is distributed over its length (unlike that of the classical pendulum introduced in Section 7.1). Use dimensional analysis to determine how the period of this pendulum depends on its mass per unit length, m, its length, l, and the gravitational constant, g.

7.40. Determine the period of the uniform rod or stick of Problem 7.39 by deriving its linearized (small angle) equation of motion. (*Hints*: Use Newton's laws of rotational motion, which then provide an analogy to the simple pendulum. The second moment of the rotational inertia is given as $I = ml^2/3$.)

7.41. Show that the total energy of the uniform rod or stick of Problem 7.40 is conserved. (*Hints*: The kinetic energy is given as $I(d\theta/dt)^2/2$. The potential energy is the pendulum's weight multiplied by the height of its mass center with respect to an appropriate datum.)

7.42. (a) Determine the rate at which energy is dissipated for a damped planar pendulum when the damping force is proportional to the square of the pendulum's speed.
(b) Confirm that the answer to part (a) is dimensionally correct.

7.43. (a) Write the equation for the total energy of an undamped linear spring-mass system in terms of its maximum displacement, A, and the spring stiffness, k.
(b) Confirm that the answer to part (a) is dimensionally correct.

7.44. Kepler's third law of planetary motion can be written as an equation for the square of a planet's period of motion around the sun,

$$T^2 = \frac{4\pi^2 a^3}{GM_s},$$

where a is the semi-major axis of the elliptical planetary orbit, M_s is the mass of the sun, and G is the universal gravitational constant. Further, Newton's first law states that the force of gravitation between the sun and a planet can be written as

$$F = \frac{GM_s(\text{mass of planet})}{(\text{distance from planet to sun})^2}.$$

(a) Starting with this form of Kepler's third law, find an equation for the frequency in the form $\omega = \omega(a, G, M_s)$.
(b) Determine the appropriate approximation of Newton's gravitational law to obtain Kepler's third law.

7.45. Explain whether or not energy is conserved in planetary motion. (*Hint*: The gravitational potential energy is GM_s(mass of planet)/(distance from planet to sun).)

7.46. Show from eq. (7.8b) that the mass of a simple pendulum attains its maximum speed when $\theta = 0°$. Is this physically reasonable?

7.47. Show that the result just obtained in Problem 7.46 is valid for both the linear and nonlinear models of the planar pendulum.

7.48. Would you expect to see energy conserved in laboratory experiments with pendulums? If not, how would the dissipation of energy make itself known?

8
Applying Vibration Models

As we noted in Chapter 7, vibration is omnipresent in our lives, both in people-made and living objects and devices. Vibration is also complex. For example, sound is modeled as a sum of *harmonics*, of vibrations with different periods or natural frequencies. Certainly buildings and cars and airplanes and dentists' drills vibrate in complex, *multi-modal* ways as well, with a lot of modes having different frequencies and different amplitudes. Given that life seems so complex, is it worth doing elementary vibration modeling? Yes, it is, as so eloquently said by one of the great pioneers of the field of vibration, Sir John William Strutt, third Baron Rayleigh, known quite widely as Lord Rayleigh:

The material systems, with whose vibrations Acoustics is concerned, are usually of considerable complication, and are susceptible of very various modes of vibration, any or all of which may coexist at any particular moment. Indeed in some of the most important musical instruments, as strings and organ-pipes, the number of independent modes is theoretically infinite, and the consideration of several of them is essential to the most practical questions relating to the nature of the consonant chords. Cases, however, often present themselves, in which one mode is of paramount importance; and even if this were not so, it would still be proper to commence the consideration of the general problem with the simplest case—that of one degree of freedom. It need not be supposed that the mode treated of is the only one possible, because so long as vibrations of other modes do not occur their possibility under other circumstances is of no moment.

Why?

Guided by Lord Rayleigh's insight, we will continue to limit our discussion of models of vibratory behavior to those having but a single degree of

freedom. We will focus on two important elements. First, we develop the *mechanical-electrical analogy*, wherein we make more explicit the several commonalities of vibration behavior that we had identified in Chapter 7. In our second focus, we note a dividing line that is extraordinarily powerful for modeling vibration: some phenomena seem to go on indefinitely, quite on their own, while others appear as responses to repetitive stimulation. Thus far, our models have been in the first category, called *free* or *unforced* vibration, referring to phenomena that continue after some initial jolt gets them going. It includes the vibration of struck piano strings and the tides of the seas. The second category that we take up in this chapter, *forced* vibration, occurs when there is a persistent, repetitive input, such as the kind an air conditioning system imparts to the building it cools or an engine imparts to the vehicle it powers.

8.1 The Spring–Mass Oscillator–II: Extensions and Analogies

How? In Section 7.3 we noted that the pendulum could be modeled as a *spring-mass oscillator*, a model we now develop by applying once again the force balance embodied in Newton's second law. We show such a spring-mass system in Figure 8.1. Newton's law states that (see Section 7.3.1) the motion of the oscillator's mass, m, is governed by

$$\text{net force} = m\frac{d^2x(t)}{dt^2}. \tag{8.1}$$

Given? Two forces are shown acting on the mass: a specified *applied force*, $F(t)$, and a force exerted by the spring. The spring is an ideal elastic spring that has no mass and dissipates no energy. Its attachment points at each end

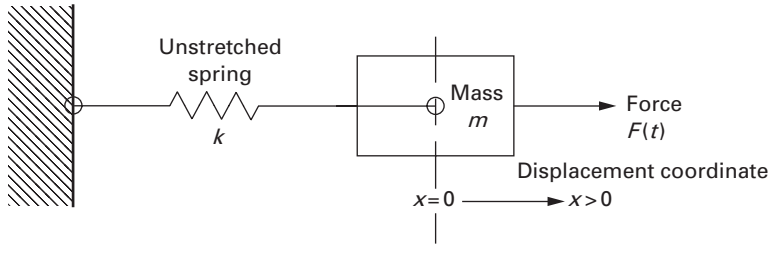


Figure 8.1 An elementary *spring-mass system* the shows an ideal spring exerting a restoring force on a mass, m, as does a specified applied force, $F(t)$. The spring's stiffness is k, and the displacement or movement of the mass to which the spring's right end is attached is $x(t)$.

are called *nodes*. The left node of the spring in Figure 8.1 is attached to a fixed point, say on a wall, while the right node is attached to a mass whose movement, $x(t)$, is the system's single degree of freedom. Moreover, the spring always exerts a *restoring force* on the node or mass that returns the spring to its original, unextended position. Thus, if moved a positive distance to the right, $x(t)$, the spring pulls the node back to the left; if the spring is compressed a distance to the left, $-x(t)$, it pushes the node back to the right. The magnitude of the spring force is given by

Assume?

$$F_{\text{spring}} = kx(t). \tag{8.2}$$

The net force on the mass is the difference between the applied and the spring forces,

$$\text{net force} = F(t) - F_{\text{spring}}. \tag{8.3}$$

so that the equation of motion is found by combining eqs. (8.1), (8.2), and (8.3):

$$m\frac{d^2x(t)}{dt^2} + kx(t) = F(t). \tag{8.4}$$

Equation (8.4) was already introduced as an analog of the pendulum in Section 7.3, where we made the argument that the gravitational pull on the pendulum mass exerted a spring-like force on the pendulum (see Problem 8.1). For free, unforced vibration, there is no applied force, and the governing equation is

$$m\frac{d^2x(t)}{dt^2} + kx(t) = 0. \tag{8.5}$$

If we introduce a scaling factor, ω_0, to make the time dimensionless, as we did in eq. (7.10), the oscillator equation (8.5) becomes

$$m\omega_0^2\frac{d^2x(\tau)}{d\tau^2} + kx(\tau) = 0, \tag{8.6}$$

which suggests that the scaling factor for the spring-mass system is

$$\omega_0 = \sqrt{\frac{k}{m}}. \tag{8.7}$$

Equation (8.7) can be confirmed to be dimensionally correct (see Problem 8.2) and, as for the pendulum, ω_0 can be identified as the *circular frequency* of the spring-mass oscillator. The circular frequency can be related to the frequency and the period:

$$f_0 = \frac{1}{T_0} = \frac{\omega_0}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{8.8}$$

Again, both f_0 and ω_0 have the physical dimensions of $(\text{time})^{-1}$, but the units of f_0 are number of cycles per unit time, while those of ω_0 are radians per unit time.

Use? Predict? Equation (8.7) is actually far more important than its simple appearance suggests. It provides a fundamental paradigm for thinking about the vibration of systems: The natural frequency of the oscillator is proportional to the square root of the *stiffness-to-mass* ratio. Thus, natural frequencies increase (and periods decrease) with increasing stiffness, k, while natural frequencies decrease (and periods increase) with increasing mass, m. We will refer back to this paradigm often, and we will also see that it captures a very useful design objective.

Why? How? We now extend the spring-mass model to incorporate non-ideal, dissipative behavior. We do this by attaching to the mass a *damping* or *dissipative element*, sometimes called a *dashpot* or *damper*, which exerts a restoring force proportional to the speed at which the element is extended or compressed:

$$F_{\text{damper}} = c\dot{x}(t). \tag{8.9}$$

The damper acts *in parallel* with the spring, as shown in Figure 8.2, so that the net force exerted on the mass is

$$\text{net force} = F(t) - F_{\text{spring}} - F_{\text{damper}}, \tag{8.10}$$

and the corresponding equation of motion for a *spring-mass-damper system* is

$$m\frac{d^2x(t)}{dt^2} + c\dot{x}(t) + kx(t) = F(t). \tag{8.11}$$

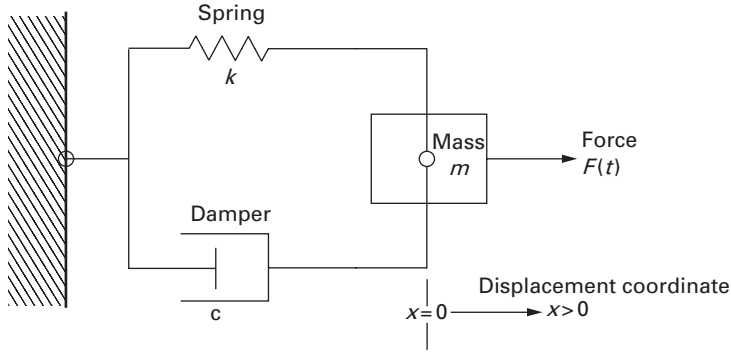


Figure 8.2 An elementary *spring-mass-damper system* that shows the ideal spring (of stiffness, k) exerting a force on a mass, m, the specified applied force, $F(t)$, and a viscous damping element that exerts a restoring force that is proportional to the speed, $\dot{x}(t)$, at which the mass moves.

This result is very similar to the corresponding result for the damped pendulum, eq. (7.27), save for the facts that the present result includes a forcing function, $F(t)$, and its spring term is (already) linear.

Problem 8.1. We experience the pull of gravity as constant and not dependent on position. How does it come to be interpreted as exerting a spring force that is linearly proportional to position? (*Hint*: Think about the equation of motion in which the relevant term appears.)

Problem 8.2. Identify the fundamental physical dimensions of the spring stiffness, k, and the mass, m, and use them to determine the physical dimensions of ω_0 for a spring-mass oscillator.

8.1.1 Restoring and Dissipative Forces and Elements

Equation (8.11) offers the prospect of generalizing the energy ideas of Sections 7.1.5 and 7.1.6 in rather broad terms. The spring-mass-damper system is itself a paradigm for a very broad range of vibration models—physical, biological, chemical, and so on. Thus, we will not only be able to identify a system's mass, but we will also be able to identify a spring-like element with a stiffness, such as the gravitational pull of the pendulum, and a dissipative element with a damping constant, much like the shock absorber of an auto suspension (see Section 8.3). There is one salient feature common to each of these elements that will be true no matter what physical, biological, chemical or other model we are analyzing: Each element either *stores* energy or *dissipates* energy. Two elements store energy in the spring-mass-damper: the mass, which stores kinetic energy,

$$KE = \frac{1}{2}m(\dot{x}(t))^2, \tag{8.12}$$

and the spring, which stores potential energy,

$$PE = \frac{1}{2}k(x(t))^2. \tag{8.13}$$

In an ideal system, where there is no damping, the spring and the mass exchange energy from potential to kinetic to potential, and so on indefinitely. Thus, the two storage elements exchange their forms of energy repetitively as the ideal spring-mass system vibrates.

The damping element dissipates energy according to (see eq. (7.29))

$$\frac{dE(t)}{dt} = -\frac{1}{2}c(\dot{x}(t))^2. \tag{8.14}$$

As a spring-mass-damper vibrates or oscillates, energy is no longer simply passed back and forth between the spring and the mass. Rather, the damping element draws energy out of the system and dissipates it as wasted power or energy, typically through the heat transfer we associate with frictional devices.

Again, these characterizations turn out to be useful for helping us analyze systems or phenomena as we try to build models of their behavior.

8.1.2 Electric Circuits and the Electrical-Mechanical Analogy

Electric circuits and their elements offer a parallel paradigm for analyzing oscillatory behavior. Consider the elementary, *parallel RLC circuit* shown in Figure 8.3. It has three ideal elements connected in parallel that are driven by a *current source* that produces a current $i_{\text{source}}(t)$. The three elements are idealized in the same way that the mass of a spring-mass system is perfectly rigid and that its spring is mass-less. The first element we introduce is the ideal *capacitor* that, when discharged, transmits a voltage drop, $V(t)$, that is proportional to the electric charge, $q(t)$, stored on two plates separated

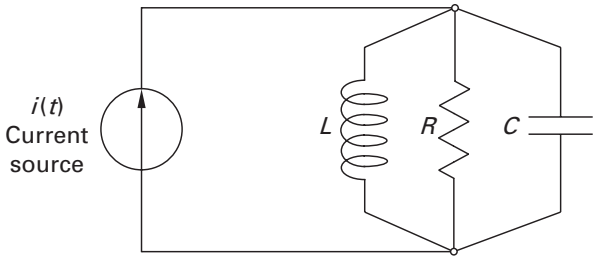


Figure 8.3 A *parallel RLC circuit* that has a current source as its driver. The elements are the *capacitor* of capacitance, C, the *inductor* with inductance, L, and the *resistor* with resistance, R. The current source provides a current of magnitude, $i_{\text{source}}(t)$.

by an insulator:

$$V(t) = \frac{q(t)}{C}. \tag{8.15}$$

The constant, C, is the *capacitance* of the capacitor and its units are farads, named after the British chemist and physicist Michael Faraday (1791–1867). The capacitor stores energy in an amount proportional to the square of the voltage across it:

$$E_C = \frac{1}{2} C \left(V(t)\right)^2. \tag{8.16}$$

Notwithstanding the elegant simplicity of eqs. (8.15) and (8.16), electrical circuit models are generally cast in terms of the time rate of change of charge, called the *current*, because it is hard to measure charge:

$$i(t) = \frac{dq(t)}{dt}. \tag{8.17}$$

This form of the capacitor model is an element that carries a current, $i_C(t)$, that is directly proportional to the time rate of change of the voltage drop, $V(t)$, across the capacitor:

$$i_C = C\frac{dV(t)}{dt}. \tag{8.18}$$

The second element we introduce is the *inductor*, which is a coil that builds up a magnetic field rate when a current flows through it. The magnetic field causes a voltage drop across the inductor that is proportional to the time rate of change of the current flowing through it:

$$\frac{di_L}{dt} = \frac{V(t)}{L}. \tag{8.19}$$

The constant, L, is the *inductance*, which is measured in henrys, named after the American physicist Joseph Henry (1797–1878). Now we integrate eq. (8.19) with respect to time,

$$i_L = \frac{1}{L}\int_{-\infty}^{t} V(t')dt', \tag{8.20}$$

where t' is a dummy variable of integration in the integral in eq. (8.20). The inductor stores energy in an amount proportional to the square of the current flowing through it:

$$E_L = \frac{1}{2} L (i_L(t))^2 = \frac{1}{2L}\left(\int_{-\infty}^{t} V(t')dt'\right)^2. \tag{8.21}$$

The third element is the *resistor*. It impedes (or resists) the flow of charge in proportion to the time rate of change of charge, or the current. The resulting voltage drop across the resistor is directly proportional to the current flowing through it:

$$i_R = \frac{V(t)}{R}, \tag{8.22}$$

where the constant, R, is the *resistance*, which is measured in ohms, named after the German physicist Georg Simon Ohm (1787–1854). The resistor, like its mechanical counterpart, the dashpot, dissipates energy by throwing it off as waste heat or power. Thus, in the context of Section 8.1.1, we can regard the resistor and the dashpot as similar dissipative elements, and the capacitor (like the mass) and the inductor (like the spring) as elements that store energy.

Why? Can we draw an analogy between the electrical elements just introduced and the spring-mass-damper system described earlier in this section? Yes. In fact, there are two well-known electrical-mechanical analogies. The choice of analogy is to some extent a matter of taste, and we describe here the one we prefer; this book's first edition presented the other.

How? We first invoke Gustav Robert *Kirchhoff's* (1824–1887) *current law* (KCL) to derive the governing equations for the parallel *RLC* circuit in Figure 8.3. The KCL states that the time rate of change of the electrical charge flowing into or out of a node or connection in a circuit must be zero. In other words, a node cannot accumulate charge. Expressed mathematically, the KCL states that

$$\frac{dq_{\text{node}}(t)}{dt} = \sum_{n=1}^{N} i_n(t) = 0, \tag{8.23}$$

where the $i_n(t)$ are the currents taken as positive flowing into the node through the N elements connected at that node. Thus, looking at the indicated currents going into and out of either of the two nodes in Figure 8.3, we see that

$$\sum_{n=1}^{N} i_n(t) = i_{\text{source}}(t) - i_C - i_L - i_R = 0, \tag{8.24}$$

where, again, $i_{\text{source}}(t)$ is the current provided by the *current source* in the circuit, and the remaining terms are the currents flowing through the capacitor, the inductor, and the resistor, respectively. Note that eq. (8.24) looks remarkably like a force balance equation [e.g., eqs. (8.3) and (8.10)]! We now replace the currents in the elements by their respective *constitutive equations* (8.18), (8.20), and (8.22), that describe how the current flows

through each relates to the voltage across each. Then eq. (8.24) becomes:

$$C\frac{dV(t)}{dt} + \frac{V(t)}{R} + \frac{1}{L}\int_{-\infty}^{t} V(t')dt' = i_{\text{source}}(t). \tag{8.25}$$

If we differentiate eq. (8.25) once with respect to time, we find:

$$C\frac{d^2V(t)}{dt^2} + \frac{1}{R}\frac{dV(t)}{dt} + \frac{1}{L}V(t) = \frac{di_{\text{source}}(t)}{dt}. \tag{8.26}$$

Equation (8.26) is a second-order, linear differential equation with constant coefficients. Its dimensions can be shown to be consistent and correct (see Problem 8.4). When solved, it yields the common voltage across the three parallel elements, from which both the currents through each and the energy stored by the capacitor and inductor can be calculated [using eqs. (8.18), (8.20), and (8.22)].

Use? Predict?

What is most noteworthy about eq. (8.26) is its uncanny resemblance to eq. (8.11), the equilibrium equation for the spring-mass-damper. It is most tempting to conclude that voltage is analogous to displacement, and that

$$C \sim m, \quad \frac{1}{R} \sim c, \quad \frac{1}{L} \sim k. \tag{8.27}$$

Some further expressions of this *electrical-mechanical analogy* are shown in Table 8.1. The analogy is interesting and useful. Consider, for example, the fact that we described the *RLC* circuit in Figure 8.3 as a parallel circuit. In the spring-mass-damper of Figure 8.2, we specifically inserted the dashpot as an element in parallel with the spring. The mass can also be said to be in parallel with the spring and the dashpot since it shares their common endpoint displacement. Further, the analogy extends into the context of system characterization: A system can be said to be very stiff if k is large or its inductance, L, is small, or as having a large effective mass or inertia if either its mass, m, or its capacitance, C, is large.

Now, to complete this introduction to the electrical-mechanical analogy, we repeat the thought that the choice of analogies is a matter of taste. The analogy presented here allows us to draw distinctions between behaviors that go *through* elements (force and current), and those measured *across* elements (displacement and voltage). The analogy also enables us to identify Newton's second law and Kirchhoff's current law as similar expressions of balance (force or current) or conservation (momentum or charge). The other analogy identifies force with voltage and displacement with charge. It, therefore, does offer some more immediately recognizable appeal because the resemblance of basic equations is even more obvious.

Table 8.1 Elements of one electrical-mechanical analogy.

Mechanical	Electrical
Momentum ($\sim$ Speed): $mv(t)$	Charge: $q(t)$
Force ($\sim d$(Momentum)$/dt$): $F = m\dfrac{dv(t)}{dt}$	Current($\sim d$: (Charge)$/dt$): $i(t) = \dfrac{dq(t)}{dt}$
Displacement: $x(t)$	Voltage: $V(t)$
Newton's 2nd @Massless Node: $\sum_{n=1}^{N} F_n(t) = \dfrac{d(mv_{\text{node}}(t))}{dt} = 0$	Kirchhoff's Current Law: $\dfrac{dq_{\text{node}}(t)}{dt} = \sum_{n=1}^{N} i_n(t) = 0$
$F_{\text{spring}} = k\int_{-\infty}^{t} v(t')dt' = kx(t)$	$i_L = \frac{1}{L}\int_{-\infty}^{t} V(t')dt'$
$F_{\text{damper}} = cv(t) = c\dot{x}(t)$	$i_R = \frac{1}{R}V(t)$
$F_{\text{net}} = m\dot{v}(t) = m\ddot{x}(t)$	$i_C = C\dot{V}(t)$
$PE = \frac{1}{2}k\,(x(t))^2$	$E_C = \frac{1}{2}C\,(V(t))^2$
$KE = \frac{1}{2}m\,(\dot{x}(t))^2$	$E_L = \frac{1}{2}L\,(i(t))^2 = \frac{1}{2}L\left(\dot{q}(t)\right)^2$

However, the preferred analogy described above is more consistent with physical principles and conforms better to our intuition of how such systems behave.

Problem 8.3. Taking as fundamental the dimensions of current, I, as charge per unit time and voltage (or *electromotive force*), V, as (force $\times$ distance) per unit charge, determine the fundamental physical dimensions of the capacitance, C, the inductance, L and the resistance, R.

Problem 8.4. Using the fundamental dimensions identified in Problem 8.3, confirm that eq. (8.26) is dimensionally consistent and correct.

Problem 8.5. Using the fundamental dimensions identified in Problem 8.3, determine whether the energy expressions for E_C and E_L given in Table 8.1 are dimensionally correct.

Problem 8.6. Determine the governing equation for the free oscillation of the voltage in a parallel LC circuit with ideal elements.

Problem 8.7. Determine the natural frequency of free vibration and the period of the ideal parallel LC circuit of Problem 8.6.

8.2 The Fundamental Period of a Tall, Slender Building

It is not surprising that buildings, especially tall and slender buildings, respond to several kinds of forces by vibrating. Buildings respond to aerodynamic forces set in motion by wind or by aircraft passing nearby. They also respond to ground-borne motion induced by traffic, earthquakes, or even explosions. These various inputs force not only the vibration of the building as a whole, but also its internal components (such as walls, floors, and windows). Further, most tall buildings have their own internal sources of vibration; for example, air conditioning systems, escalators, and elevators. What is most noteworthy is that tall buildings tend to be built lighter and with more flexibility than were earlier tall buildings (see Figure 8.4). For example, the Empire State Building is a good bit heavier and stiffer than the Sears Tower in Chicago (or were the towers of the World Trade Center in New York). As a result, building vibration, both local to a room and global to the building, has become a critical element in building design: vibration can create problems of annoyance, dysfunction, and outright danger for a building's occupants. The assessment of the

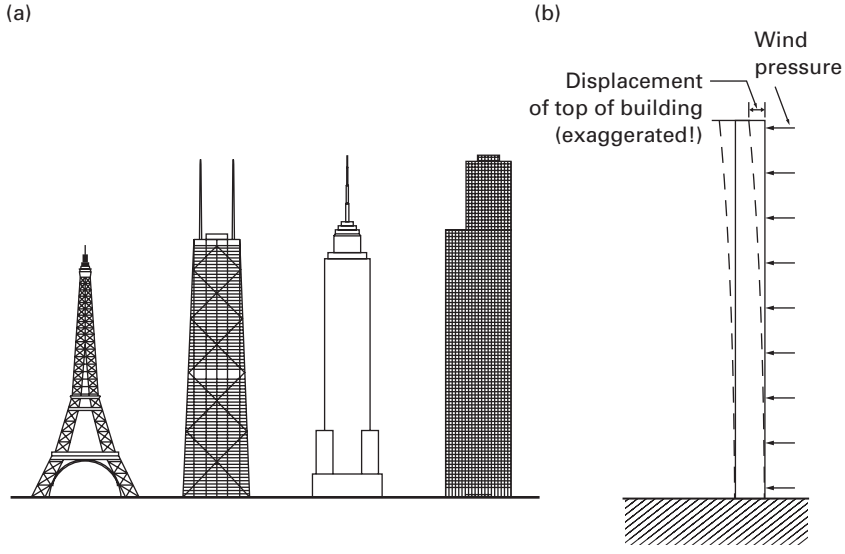


Figure 8.4 A small collection of skyscrapers, including the Eiffel Tower, the John Hancock center, the Empire State Building and the Sears Tower (after Billington, 1983). They are (mostly) tall and slender buildings that grace the skylines of modern cities. Also shown is a generic schematic of the greatly exaggerated movement of such a tall building in response to wind loading.

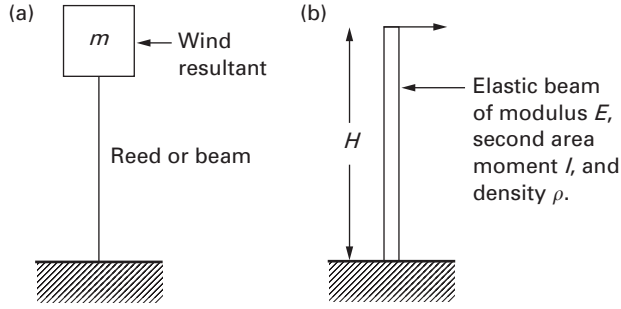


Figure 8.5 Two simple models used to estimate the fundamental period of vibration of tall, slender buildings: (a) a simple spring-mass system that is shown as a reed with a mass at its end; and (b) a cantilever beam model wherein the stiffness and the mass are distributed (uniformly) over the building height, H, but from which a simple spring-mass system can be deduced.

vibration response of a tall building, or any such structure, requires deep understanding of the building's dynamic properties, such as its own fundamental period or its natural frequency. It turns out, interestingly enough, that a "first-order" estimate of the natural frequency or period of a tall building can be obtained by making a lot of assumptions and modeling the entire building as a simple spring-mass system.

Assume? Consider the generic skyscraper shown in Figure 8.4, together with profiles of some real counterparts. We assume that the wind pressure is uniform over the building height and oriented normal to the side shown. The wind pressure produces a net force that pushes on that building face, thus making the building bend, with the largest movement at its free end at the top. Since buildings are made up of elastic structural members, which are themselves springs, we expect that the building will resist the bending motion caused by the wind and return to its original straight configuration when the wind ceases. In this sense, we can draw the building as a whole as if it were a simple elastic reed with a mass concentrated at its free end [see Figure 8.5(a)], but where this reed-and-mass system is exactly the same as the spring-mass system defined in Sections 7.3 and 8.1. We need only determine the stiffness, k, and the mass, m.

Given? One way to determine the stiffness of a building is to measure its deflection while a load or force is being applied to the building and then back-calculate the stiffness. (For a yet-to-be-built design, a similar measurement could be made on a comparable building.) Consider, for

example,a recently-built building with a *square* cross-section, $B = 30$ m (98.4 ft), on a side, and of height, $H = 300$ m (984 ft). (For a working calculation in standard American units, an experienced engineer would be likely to use $B = 100$ ft and $H = 1000$ ft.) A very strong, gale-force wind, say 100 mph (44.7 m/sec), produces a pressure of 1.23 kN/m^2 (25.7 lb/ft^2) on the building, or a total wind force of

$$\text{wind force} = \begin{cases} 1.23\,\text{kN/m}^2 \times 30\,\text{m} \times 300\,\text{m} \\ 25.7\,\text{lb/ft}^2 \times 98.4\,\text{ft} \times 984\,\text{ft} \end{cases} = \begin{cases} 11.1 \times 10^6\,\text{N} \\ 2.49 \times 10^6\,\text{lb} \end{cases} \tag{8.28}$$

We will assume that the resultant of this force acts halfway up the building. **Assume?**
The building will bend or move when it is loaded. A practical estimate is **Given?**
that the top of the building will move about 0.3% of its height, or $0.003H$.
Further, the deflection or movement of the building varies nonlinearly with **Assume?**
height, so we will assume that the movement at that height is one-third of the movement at its top. With the building top expected to move 0.9 m (2.95 ft), we can calculate its stiffness as

$$k = \begin{cases} 11.1 \times 10^6\,\text{N} \div 0.30\,\text{m} \\ 2.49 \times 10^6\,\text{lb} \div 0.98\,\text{ft} \end{cases} = \begin{cases} 37.0 \times 10^6\,\text{N/m} \\ 2.54 \times 10^6\,\text{lb/ft} \end{cases} \tag{8.29}$$

To determine the building's fundamental period or natural frequency, we
need its mass. A practical estimate of the weight of a building uses an average **Given?**
specific weight of $\gamma = 1.50\,\text{kN/m}^3$ (9.54 lb/ft^3) for a modern steel-framed tower with a 12 ft story height. In this case, the mass of the building can be calculated as:

$$m = \frac{\gamma H B^2}{g} = \begin{cases} [1.50\,\text{kN/m}^3 \times 300\,\text{m} \times (30\,\text{m})^2] \div 9.80\,\text{m/(sec)}^2 \\ [9.54\,\text{lb/ft}^3 \times 984\,\text{ft} \times (98.4\,\text{ft})^2] \div 32.2\,\text{ft/(sec)}^2 \end{cases}$$
$$= \begin{cases} 4.13 \times 10^7\,\text{kg} \\ 2.82 \times 10^6\,\text{lbm} \end{cases} \tag{8.30}$$

Thus, the fundamental period of this hypothetical generic skyscraper is

$$T_0 = 2\pi\sqrt{\frac{m}{k}} = 2\pi \begin{cases} \sqrt{4.13 \times 10^7\,\text{kg} \div 37.0 \times 10^6\,\text{N/m}} \\ \sqrt{2.82 \times 10^6\,\text{lbm} \div 2.54 \times 10^6\,\text{lb/ft}} \end{cases}$$
$$\cong \begin{cases} 6.64\,\text{sec} \\ 6.62\,\text{sec} \end{cases} \cong 6.6\ \text{sec}\,. \tag{8.31}$$

The result of eq. (8.31) is well within the range that experience suggests **Verified?**
for the period of a modern, steel-framed building,which is about 5–10 sec for buildings whose height is within the range of 214–427 m (700–1400 ft). Another estimate is that the period of a building is within the range of

0.05–0.15 times the number of stories or floors. Since our hypothetical building is likely to have something like 85 floors, our estimate of its period is once again verified.

How? Another way to estimate the period or natural frequency of a building is to model it as a simple cantilever beam where the stiffness and mass are **Given?** distributed over the length of the beam [see Figure 8.5(b)]. The theory of strength of materials says that the stiffness of a cantilever beam of length, H, measured at the top, is given by

$$k_{\text{beam}} = \frac{3EI}{H^3}, \tag{8.32}$$

and that its period of vibration is given by

$$T_{\text{beam}} \cong 1.78H^2\sqrt{\frac{\gamma A}{gEI}}. \tag{8.33}$$

Here γ is, again, the specific weight of the beam (or building), A is the beam's cross-sectional area, I its *second moment of the cross-sectional area*, and E the *modulus of elasticity* of the material of which the beam is made. **Given?** Given that the dimensions of E are force per unit area and of I are (length)^4, it is easily verified that eq. (8.33) is dimensionally correct (see Problems 8.8– **Predict?** 8.9). Note that the stiffness decreases with H^3, while the period increases **Use?** with H^2, which means that its natural frequency also drops as H^2. Thus, a short building is stiffer than a tall building. In fact, the stiffnesses of two buildings made of the same material and having the same floor plan are related to each other as the cube of the inverse ratio of their heights.

It is also clear from eqs. (8.32) and (8.33) that the beam's stiffness increases with the product EI, and the period decreases with $1/\sqrt{EI}$. What do E, I, and their product EI mean for a beam and for a building? The modulus, E, represents the stiffness of the material of which the beam is made, and, not surprisingly, $E_{\text{steel}} > E_{\text{concrete}} > E_{\text{wood}}$. So, in very loose terms, a higher modulus is more suitable for taller buildings because of their higher material stiffness. (There are other issues involved, for example, the specific weight and the failure strength of materials, but that is well beyond our current modeling scope.) The second moment of the area, I, also (erroneously) called the *moment of inertia*, reflects the distribution of the cross-sectional area about its own centerline. For a building of square cross-section, $I \sim B^4$ roughly speaking, so that both the second moment of a building and its stiffness increase with its basic plan dimension to the fourth power, B^4.

This very brief overview of building vibrations suggests why engineers have had to worry only relatively recently about the effects of wind on tall buildings. Certainly tall structures were built long ago; one can point to the amazing cathedrals built during the Middle Ages (recall the discussion

in Section 3.2.3), and even to the Eiffel Tower built in 1889. However, with the advent of both high-strength steels developed in the 20th century and new architectural styles, the flexible skyscraper came into being, bringing along both interesting problems and equally interesting opportunities. Thus, designing a building now means designing its dynamic properties and vibration response for sources of dynamic loading, including wind, earthquakes, nearby traffic, and mechanical systems within. Back-of-the-envelope estimates such as we have made play an important role in these designs because they enable engineers to make reasonable estimates of their designs long before they have to specify those designs to costly, detailed levels (see also Problems 8.41 and 8.42).

Problem 8.8. Given that the dimensions of the modulus of elasticity, E, are force per unit area, what are the dimensions of the second moment, I, that make eq. (8.32) dimensionally correct?

Problem 8.9. Using the dimensions identified in Problem 8.8, confirm that eq. (8.33) is dimensionally consistent and correct.

Problem 8.10. What is the pressure produced by a 100 mph wind expressed as a fraction of atmospheric pressure?

Problem 8.11. Show that the ordering of elastic moduli $E_{\text{steel}} > E_{\text{concrete}} > E_{\text{wood}}$ is correct in both metric and standard American units. (*Hint*: Use the library!)

Problem 8.12. For a tall cantilever of specific weight, γ, what are the physical dimensions of the parameter, $c \equiv \sqrt{E/(\gamma/g)}$? What could this parameter signify?

Problem 8.13. For the tall cantilever of Problem 8.12, with $I \sim B^4$, show that $T_0 \sim (H/c)(H/B)$. Is this result dimensionally meaningful?

8.3 The Cyclotron Frequency

Why? To show that fundamental periods and frequencies are also important in other domains, we now present a simple model of the *cyclotron*, the device that forces charged particles to move in a circular path when subjected to a magnetic field. Electrons, protons, and ions are among the charged particles spun in cyclotrons. We will determine the fundamental frequency of the **How?** cyclotron by using some basic results from electromagnetism. A charged

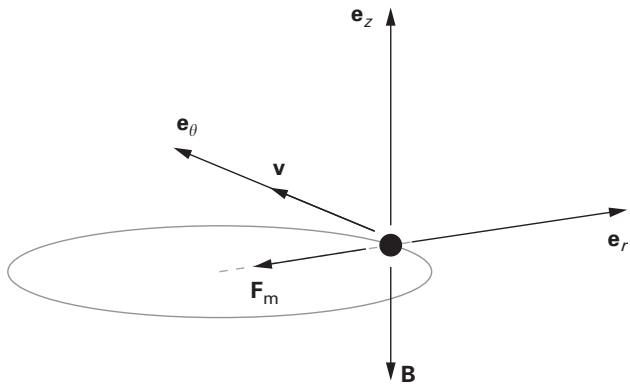


Figure 8.6 The cylindrical coordinate system and the basic vector structure needed to portray the elementary *cyclotron*. The coordinate system has the radial, tangential, and vertical unit vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$, and $\mathbf{e}_z$, respectively. The particle location is given by $\mathbf{r} = |\mathbf{r}|\mathbf{e}_r$. The magnetic field is directed in the $-z$ direction, that is, $\mathbf{B} = -|\mathbf{B}|\mathbf{e}_z$, and the magnetic force exerted on the charged particle is $\mathbf{F}_\mathrm{m}$.

particle moving through a magnetic field is subjected to a magnetic force (vector), $\mathbf{F}_\mathrm{m}$, given by:

$$\mathbf{F}_\mathrm{m} = q\mathbf{v} \times \mathbf{B}, \tag{8.34}$$

where $\mathbf{B}$ is the *magnetic induction* (vector) due to currents other than that produced by the particle charge of magnitude q, $\mathbf{v}$ is the velocity (vector) of the moving charged particle, and the symbol $\times$ denotes the *vector* or *cross product* of the $\mathbf{v}$ and $\mathbf{B}$ vectors.

The geometry underlying our cyclotron model is shown in Figure 8.6. The particle motion is described in a cylindrical set of coordinates having radial, tangential, and vertical unit vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$ and $\mathbf{e}_z$, respectively. The location of the particle is given by $\mathbf{r} = |\mathbf{r}|\mathbf{e}_r$. The magnetic field is directed in the $-z$ direction, that is, $\mathbf{B} = -|\mathbf{B}|\mathbf{e}_z$. Thus, the vector equation (8.34) can be written as

$$\mathbf{F}_\mathrm{m} = q\mathbf{v} \times (-|\mathbf{B}|\mathbf{e}_z), \tag{8.35}$$

where will soon identify the angle between the $\mathbf{v}$ and $\mathbf{B}$ vectors as θ.

Equations (8.34) and (8.35) indicate that the magnetic force, $\mathbf{F}_\mathrm{m}$, is perpendicular to the particle velocity, $\mathbf{v}$. Thus, the magnetic field, $\mathbf{B}$, imparts no power to the particle. Further, since both the force, $\mathbf{F}_\mathrm{m}$, and its consequent particle acceleration are perpendicular to the velocity, $\mathbf{v}$, the particle must

be traveling in a circle of radius, $|\mathbf{r}|$, at a (radian) frequency, ω_0, that also corresponds to simple harmonic motion. Further, that circular harmonic motion also means that the velocity vector is simply $\mathbf{v} = |\mathbf{v}|\mathbf{e}_\theta = |\mathbf{r}|\omega_0\mathbf{e}_\theta$ (see Problems 8.17–8.19). It then follows that eq. (8.35) becomes:

$$\mathbf{F}_\mathrm{m} = q(|\mathbf{v}|\mathbf{e}_\theta) \times (-|\mathbf{B}|\mathbf{e}_z) = q(|\mathbf{r}|\omega_0\mathbf{e}_\theta) \times (-|\mathbf{B}|\mathbf{e}_z) = -q|\mathbf{r}|\omega_0|\mathbf{B}|\mathbf{e}_r. \tag{8.36}$$

Equation (8.36) shows that the force, $\mathbf{F}_\mathrm{m}$, is directed radially inward, so that the acceleration is centripetal and also directed radially inward. Thus, just as with the centripetal acceleration of the pendulum (see eqs. (7.7a) and (7.8a)), the centripetal acceleration of the cyclotron particle is $-|\mathbf{r}|\omega_0^2\mathbf{e}_r$. Then, with the net force being $\mathbf{F}_\mathrm{m}$ of eq. (8.36), Newton's second law in the radial direction appears as

$$\mathbf{F}_m = -q|\mathbf{r}|\omega_0|\mathbf{B}|\mathbf{e}_r = -m|\mathbf{r}|\omega_0^2\mathbf{e}_r,$$

which finally yields the *cyclotron frequency*,

$$\omega_0 = \frac{q}{m}|\mathbf{B}|. \tag{8.37}$$

Equation (8.37) shows that the frequency depends only on the strength of the magnetic field, $\mathbf{B}$, and the charge-to-mass ratio, q/m, of the particle. It is independent of the radius of the circle and, therefore, the tangential velocity. Again, eq. (8.37) is the fundamental relationship behind cyclotron design.

Problem 8.14. If the fundamental dimension of charge is Q, determine the dimensions of the magnetic field or *magnetic flux density* $\mathbf{B}$ that ensure that eq. (8.34) is dimensionally correct.

Problem 8.15. The magnetic field $\mathbf{B}$ has units of *webers per square meter* ($\mathrm{Wb/m^2}$) in SI units. Using eq. (8.34), express these units in terms of units of charge (the *coulomb*, C) and other fundamental dimensions in SI units.

Problem 8.16. Verify that the cyclotron frequency as given in eq. (8.37) is dimensionally correct.

Problem 8.17. Calculate the velocity components of a point located in a plane by the relation $\mathbf{r} = |\mathbf{r}|\mathbf{e}^{j\omega_0 t} = x(t)\mathbf{i}_x + y(t)\mathbf{i}_y$. Express that velocity in terms of (a) the time derivatives of $x(t)$ and $y(t)$ and then (b) in terms of $|\mathbf{r}|e^{j\omega_0 t}$.

Problem 8.18. Why do the results of Problem 8.17 express simple harmonic motion?

Problem 8.19. Calculate the velocity components of a point located in a plane by the relation $\mathbf{r} = |\mathbf{r}|\mathbf{e}_r$ and express the results in plane polar coordinates. How does this result compare with that found in Problem 8.17?

8.4 The Fundamental Frequency of an Acoustic Resonator

What is an acoustic resonator? We have all blown air across the top of a bottle and heard a deep, foghorn-like response. In fact, the frequency (or *pitch*) that we hear is very much a function of the size of the air cavity in the bottle (and *not* a function of the kind of liquid in the bottle!). An *acoustic resonator* is a flask or bottle with an air cavity that is used to produce sound. Such resonators are also called *Helmholtz resonators* after the German physicist who investigated it, Herman von Helmholtz (1821–1894). By what mechanism do acoustic resonators work?

Why?

We will answer that equation by modeling the flask shown in Figure 8.7 and, in so doing, we will account for the mechanics and thermodynamics of the changes in pressure and volume of a gas as it transmits a sound signal. The flask has an "interior" cavity of volume V_0, that contains a gas of density ρ_0, at ambient pressure, p_0. The neck of the flask is of length L and has a cross-sectional area A. We will see that the gas in the flask neck moves like a mass and that the cavity exerts a spring-like response to that movement, so that our resonator model will be a mass-spring system.

How?

Predict?

We take the mass of gas in the neck as our mass, $m = \rho_0 AL$, to develop this model (or *this* analogy). The stiffness in the system comes from the gas in the cavity that resists being compressed as the neck mass moves toward it. That resistance is transmitted at the interface between neck and cavity by the pressure, p_0. The pressure, p_0, and the cavity volume of gas, V_0, that contains it are assumed to obey the *adiabatic gas law*:

Assume?

$$pV^{\gamma} = \text{constant}, \tag{8.38}$$

where in this instance, γ is the ratio of heat capacities ($\gamma = 1.4$ for air, for example), and p and V are, respectively, pressure and volume. When the mass of gas in the neck, m, moves a distance, x, to the right, the cavity volume must be reduced by

$$\delta V = -Ax. \tag{8.39}$$

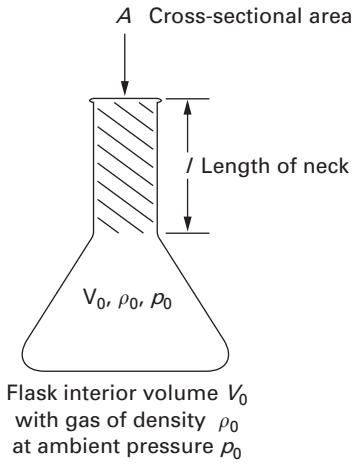


Figure 8.7 The flask used to model the acoustic or *Hemlholtz resonator*. The flask has a neck of length L, with area A, that is connected to an acoustic cavity of volume, V_0. The cavity contains a gas of density ρ_0, at pressure, p_0. When the mass in the neck moves, the cavity responds like a spring.

A small change of volume, δV, is related to a small change of pressure, δp, by the gas law (eq. (8.38)),

$$\delta(pV^\gamma) = V^\gamma(\delta p) + p(\gamma V^{\gamma-1}\delta V) = 0,$$

which, after dividing through by pV^γ, becomes

$$\frac{\delta p}{p} + \gamma\frac{\delta V}{V} = 0. \tag{8.40}$$

We now let the pressure and volume take on their ambient values, so eq. (8.40) becomes

$$\delta p = -\gamma\frac{p_0}{V_0}\delta V = 0. \tag{8.41}$$

Finally, we substitute the volume change, δV, from eq. (8.39) to find that the pressure change, δp is related to the distance moved by the neck mass, x, according to:

$$\delta p = \gamma p_0\frac{Ax}{V_0}. \tag{8.42}$$

Note that the dependence of δp on x is strongly suggestive of spring-like action, but the dimensions certainly don't look like those of a spring. On the other hand, if we recognize that the cavity-produced restoring force, F_{cavity}, acting on the neck mass is the product of pressure times area, then we see that

$$F_{\text{cavity}} = \delta pA = \gamma \frac{p_0 A^2}{V_0} x. \tag{8.43}$$

Now the resemblance to the classic spring is more evident (see Problem 8.20).

Then, if we blow across the open end of the flask with a force, $F(t)$, the mass, m, is pushed down the neck a distance, x, toward the cavity, and the cavity pushes back with a spring stiffness, k_{cavity}, the acoustic resonator behaves as a spring-mass system:

$$\rho_0 AL \frac{d^2x(t)}{dt^2} + \gamma \frac{p_0 A^2}{V_0} x(t) = F(t). \tag{8.44}$$

We can rewrite eq. (8.44) in terms of a parameter that is often used in acoustics and vibration problems, the *speed of sound* of the gas in the flask, c_0. That speed is related to the specific heat capacity, ambient pressure, and density of the gas:

$$c_0^2 = \gamma \frac{p_0}{\rho_0}. \tag{8.45}$$

Then the oscillator equation for the Helmholtz resonator is

$$\rho_0 AL \frac{d^2x(t)}{dt^2} + \frac{\rho_0 c_0^2 A^2}{V_0} x(t) = F(t). \tag{8.46}$$

The natural frequency or fundamental period follows from the homogenous version of the equation of motion (8.46) for the acoustic resonator (see Problem 8.22):

$$\omega_0 = \frac{2\pi}{T_0} = c_0 \sqrt{\frac{A}{V_0 L}}. \tag{8.47}$$

Predict? Use? Equation (8.47) could be accepted as the final result. It has the correct dimensions and shows that the frequency increases with the neck area but decreases as the neck gets longer and the cavity volume gets larger, which effects are consistent with our intuition (see Problems 8.23–8.26).

Improve? However, a bit of reflection suggests that eq. (8.47) can be further massaged by identifying the volume of the neck (which is also the volume of the moving mass m) as $V_n = AL$. Then the frequency and period become:

$$\omega_0 = \frac{2\pi}{T_0} = \frac{c_0}{L} \sqrt{\frac{V_n}{V_0}}. \tag{8.48}$$

This version of the natural frequency of the acoustic resonator is even more interesting (and satisfying) because it shows the dependencies in a more meaningful way. The frequency goes down if we elongate the neck because it takes the mass longer to move down the neck, as we see from the ratio c_0/L. Further, the effect of increasing flask volume to get deeper (lower) frequencies will not be seen unless that volume reduction is done with respect to the neck volume.

Verified?

Finally, the inhomogeneous version of the resonator model, eq. (8.46), begins to set the stage for the rest of this chapter. What does happen when there is a forcing function $F(t)$? What does $F(t)$ look like? It is easy enough to imagine that the wind blowing across the top is an acoustic signal that is, like most sounds, composed of many frequencies. Since eq. (8.46) is linear, we could obtain a complete solution by solving it for each frequency represented in $F(t)$ and then superposing or adding all of those solutions. This suggests that we seek a generic solution to

Why?

How?

$$\frac{d^2x(t)}{dt^2} + \omega_0^2 x(t) = \frac{F_0}{\rho_0 V_{\text{neck}}} \cos \omega t. \tag{8.49}$$

The radian frequency, ω, in eq. (8.49) is arbitrary and can assume any value, so the forcing function represents any oscillatory signal or input. As we will see in Sections 8.6 and 8.7, there are some very interesting effects that occur. But, first, we want to explore another way in which forcing functions occur in models of vibration (Section 8.5), and then we will talk about the mathematics (Section 8.6) and the physics (Section 8.7) that occur in governing equations like eq. (8.49).

Problem 8.20. Show that the dimensions of $\gamma(p_0 A^2/V_0)$ are such that eq. (8.43) identifies the stiffness of the flask cavity, k_{cavity}.

Problem 8.21. How does the stiffness of a cavity change if the gas is assumed to be governed by the *ideal gas law*, $pV = nRT$?

Problem 8.22. Show that the homogeneous solution of eq. (8.46) requires that the resonator's natural frequency must be given by eq. (8.47). (*Hint*: Recall Section 7.2.2.)

Problem 8.23. Estimate the natural frequency of the cavity of a standard (750 ml) wine bottle. How does that frequency compare with the note middle C, for which $f = 262$ Hz?

Problem 8.24. How long would the wine bottle flask have to be to get its cavity frequency *below* the *lowest* note produced by a piano (~ 55 Hz)?

Problem 8.25. How long would the wine bottle flask have to be to get its cavity frequency *above* the *highest* note produced by a piano (~ 8360 Hz)?

Problem 8.26. Assume that a set of acoustic resonators is built like wine bottles, each with neck radius r_n, neck length L_n, cavity radius r_0, and cavity length L_0. How would the ratio L_n/L_0 vary with the radii if every bottle were to have the same natural period?

8.5 Forcing Vibration: Modeling an Automobile Suspension

Why? We finished our discussion of the acoustic resonator by noting how it could be forced to vibrate or respond, in that case with an excitation that was external and obvious. However, excitation can show up in models in

How? other ways, as we now illustrate. Consider the damped oscillator shown in Figure 8.8(a) that is no longer connected to a fixed point or wall; rather, its free end travels over a specified contour, $y(z)$. It is a schematic for the suspension systems we are accustomed to seeing in cars, for example, and nowadays on high-end bikes. For the auto, the mass is that of the body, the power train, and the passengers and cargo. The spring is typically a coil spring that is wrapped around the shock absorber or damper. There was a time when auto springs were leaf springs, but their suspension systems would have been modeled the same way. The important feature is that both leaf and coil springs share common connection points with the shock absorber on the auto frame at one end and on the wheel at the other. Thus, spring and damper are in parallel with the auto's mass.

One way to set an auto suspension system in motion is to push rhythmically on its fender, a fairly common qualitative test of whether the shock absorbers retain much damping. This might be modeled in the same way we proposed modeling blowing over an acoustic cavity by including a forcing function, $F(t)$. In addition, however, the suspension system is excited or driven by the end connected to the wheel as it follows the road, $y(z)$. The model for the auto following the road contour is shown in Figure 8.8(b), where $a(t)$ is the amount that the wheel-end of the suspension moves with respect to a fixed wall. This means that the net extension of the spring is

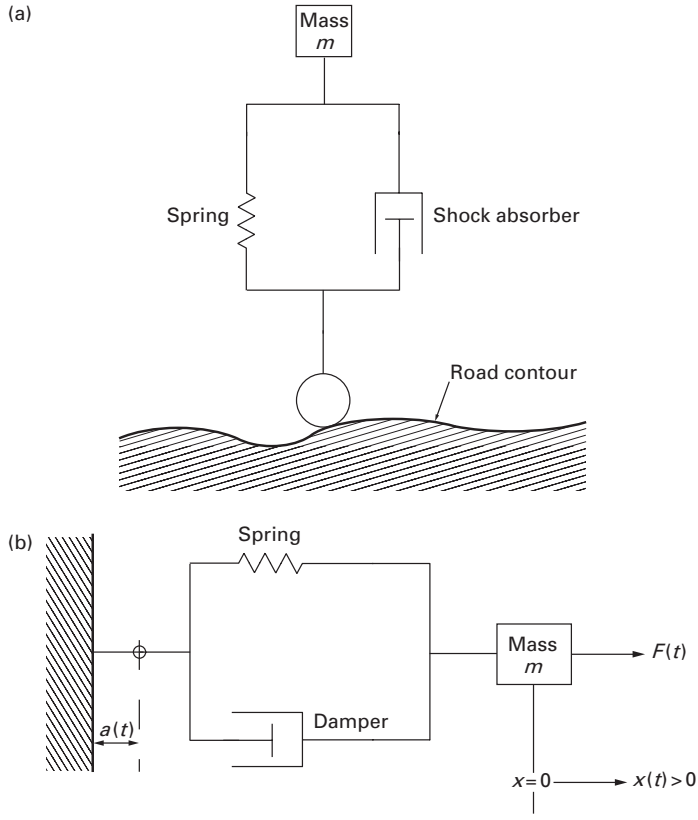


Figure 8.8 The spring-mass-damper system used to model the behavior of a *vehicle suspension system*: (a) the system's three elements (m, k, c) act in parallel and share the single coordinate, $x(t)$, while the other ends of the spring and damper share the wheel as a common connection point that follows the road contour, $y(z)$; and (b) the revision of the model to show the road-following wheel motion as a support that moves a distance, $a(t)$, with respect to the "traditional" spring-mass-damper.

$x(t) - a(t)$, and that the relative speed to which the damper responds is $d[x(t) - a(t)]/dt$. The spring force is then $k[x(t) - a(t)]$, and the damping force is $cd[x(t) - a(t)]/dt$, so that Newton's second law for this model is:

$$m\frac{d^2x(t)}{dt^2} = F(t) - k[x(t) - a(t)] - c\frac{d[x(t) - a(t)]}{dt},$$

or

$$m\ddot{x} + c\dot{x}(t) + kx(t) = F(t) + c\dot{a}(t) + ka(t). \tag{8.50}$$

Equation (8.50) shows that the terms due to the wheel motion, $a(t)$, remain on the right-hand side, because they are a known input. Thus, eq. (8.50) represents an instance of *forced* vibration even absent an explicit forcing function, that is, even when $F(t) = 0$.

Given? Consider the case of an auto without an explicit forcing function (i.e., with $F(t) = 0$) traveling in the z direction along a road whose contour $y(z)$ is given as:

$$y(z) = a_0 \sin \alpha z, \tag{8.51}$$

where α is a parameter with dimensions of (length)$^{-1}$. If the auto moves down the road at constant speed, v, it follows that $z = vt$, so that the wheel motion is

$$a(t) = y(z = vt) = a_0 \sin \alpha vt. \tag{8.52}$$

Then the governing equation for the traveling suspension system is found when eq. (8.52) is substituted into eq. (8.50):

$$m\frac{d^2x(t)}{dt^2} + c\dot{x}(t) + kx(t) = a_0(k \sin \alpha vt + c\alpha v \cos \alpha vt). \tag{8.53}$$

Thus, for this model, we once again have a non-zero right-hand side or forcing function made up of trigonometric terms. And, again, this resulted not from an explicit external forcing function, but from the fact that the system's spring and damper were not attached to an immovable point.

Problem 8.27. What are the physical dimensions of the term αv in eq. (8.53)? Explain whether or not those dimensions are correct.

Problem 8.28. Determine the values of C_1 and ϕ that allow the right-hand side of eq. (8.53) to be written in the form $a_0 C_1(\cos(\alpha vt - \phi))$.

Problem 8.29. Determine the values of C_2 and ϕ that allow the right-hand side of eq. (8.53) to be written in the form $a_0 C_2(\sin(\alpha vt + \phi))$.

8.6 The Differential Equation $md^2x/dt^2 + kx = F(t)$

How? How do we determine the solution to the inhomogeneous differential equation that describes the dynamic response of an ideal, undamped oscillator

that is driven by a harmonic forcing function (see Problems 8.28 and 8.29 above):

$$m\frac{d^2x(t)}{dt^2} + kx(t) = F_0\cos(\omega t - \phi). \tag{8.54}$$

The solutions to inhomogeneous differential equations have two parts that are superposed. The first part is the transient solution to the homogeneous equation that we had already found as eq. (7.48) or (7.49) in Section 7.2.2. The second part is the *particular* or *steady-state solution* that is crafted to solve only the differential equation without regard to the system's initial conditions.

As a trial particular solution let us assume that

$$x(t) = X_0\cos(\omega t - \phi), \tag{8.55}$$

where X_0 is a constant yet to be determined. By direct substitution of eq. (8.55) into eq. (8.54), we get:

$$(k - m\omega^2)X_0\cos(\omega t - \phi) = F_0\cos(\omega t - \phi),$$

which means that

$$X_0 = \frac{F_0}{k - m\omega^2} = \frac{F_0/k}{1 - (\omega/\omega_0)^2}, \tag{8.56}$$

where once again ω_0 is the natural frequency of the ideal oscillator defined in eq. (8.7). The final form of the steady-state solution is, then,

$$x(t) = \frac{F_0/k}{1 - (\omega/\omega_0)^2}\cos(\omega t - \phi). \tag{8.57}$$

This all seems perfectly straightforward but for one detail: If the frequency of the driving force, ω, happens to equal the natural resonance of the system, ω_0, the solution (8.57) "blows up" or becomes infinite. Now in the real world that may not literally happen because of damping, but even with the ameliorating effect of damping there is a problem when $\omega = \omega_0$. In the next section we will identify that as *resonance*, but here we want to stay focused on the formal mathematics. To complete that we note simply that a formal solution to eq. (8.54) does exist for the case $\omega = \omega_0$, and that solution can be shown to be (see Problem 8.31):

$$x(t) = \frac{F_0}{2m\omega_0}t\sin(\omega_0 t - \phi). \tag{8.58}$$

Note that $x(t)$ depends linearly on t in eq. (8.58), a result that clearly confirms the singular behavior of the ideal spring-mass system when it is excited or driven at its natural frequency. In the real world, again, damping

comes very much into play, and avoiding such resonant behavior (even with damping) is a major priority in the design of vibrating systems. We will have more to say about that in Section 8.7.

Problem 8.30. Determine the value of X_0 in eq. (8.55) by substituting eq. (8.55) into eq. (8.54) and ensuring that the equation of motion is indeed satisfied.

Problem 8.31. Confirm that the solution (8.58) does satisfy eq. (8.54) for the special case of resonance, that is, when $\omega = \omega_0$.

Problem 8.32. Determine and explain the dimensions of the coefficients $(F_0/m\omega_0)$ in eq. (8.58).

8.7 Resonance and Impedance in Forced Vibration

We now turn to the meaning and physical implications of the mathematics of simple forced oscillators. So, we again start with the equation of motion of an ideal spring-mass system that is driven by a harmonic excitation:

$$m\frac{d^2x(t)}{dt^2} + kx(t) = F_0\cos(\omega t - \phi). \tag{8.59}$$

The complete solution to eq. (8.59) is the sum of the homogeneous or transient solution (7.48) and the particular or steady-state solution (8.57):

$$x(t) = B_1\cos\omega_0 t + B_2\sin\omega_0 t + \frac{F_0/k}{1-(\omega/\omega_0)^2}\cos(\omega t - \phi). \tag{8.60}$$

where B_1 and B_2 are arbitrary constants that will be determined by the initial conditions set for the system. Having written the complete solution, it must be said that our primary interest lies in the steady-state solution because it predicts the behavior of the spring-mass system for as long as we drive it with the harmonically varying force in eq. (8.59). Further, it is independent of the initial conditions, which, as we noted in Section 7.22, affect only the transient behavior. (It should be noted that the notion of a transient solution that, implicitly, does not affect the steady state, does assume that there is at least a little bit of damping, so that solutions initiated only by the initial conditions will die out. The steady-state solution persists even in the face of damping because the excitation persists.)

Since we can always incorporate the effects of the initial conditions by suitably adjusting the two arbitrary constants in the complete solution (8.60), we take eq. (8.60) in the following form as the solution of interest:

$$x(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t - \phi). \tag{8.61}$$

We note that $x(t)$ has the same temporal behavior as the forcing function, that is, its behavior in time is the same. Thus, we say that the motion of the mass is *in phase* with the action of the driver. On the other hand, the speed of the response is given by

$$\frac{dx(t)}{dt} = -\frac{\omega F_0}{m(\omega_0^2 - \omega^2)} \sin(\omega t - \phi), \tag{8.62}$$

which shows that the speed is out of phase with the driver by 90°, that is, the speed of the mass lags behind the force by a time equal to $t = \pi/2\omega$. Now eq. (8.62) can also be written as (see Problem 8.35):

$$\frac{dx(t)}{dt} = \frac{F_0}{m\omega_0[(\omega/\omega_0) - (\omega_0/\omega)]} \sin(\omega t - \phi), \tag{8.63}$$

As we just saw in Section 8.6, the displacement and the speed become infinitely large as the forcing frequency, ω, approaches the natural frequency, ω_0. Thus, when the driving frequency equals the natural frequency, we have the condition of *resonance*. The oscillatory forcing function produces an infinite response. In fact, resonance is what we are trying to achieve when we time the pushes given to someone seated on a playground swing! In Figure 8.9 we have sketched the shape of the ideal response curve of eqs. (8.61) or (8.57) on a set of axes rendered dimensionless: $kx(t)/F_0$ against ω/ω_0. The infinite peak for the ideal case is quite obvious. We have also shown there a sketch of the damped response, which we will discuss shortly, but note that it is bounded and finite.

In acoustics and vibration research and practice, resonance and other vibratory phenomena are exhibited and measured in terms of a system's impedance, which for the system modeled here is:

$$Z(\omega) \equiv |F(t)| \Big/ \left| \frac{dx(t)}{dt} \right|,$$

which means that the *impedance for an ideal spring-mass system* is

$$Z(\omega) = m\omega_0 \left(\frac{\omega_0}{\omega} - \frac{\omega}{\omega_0} \right). \tag{8.64}$$

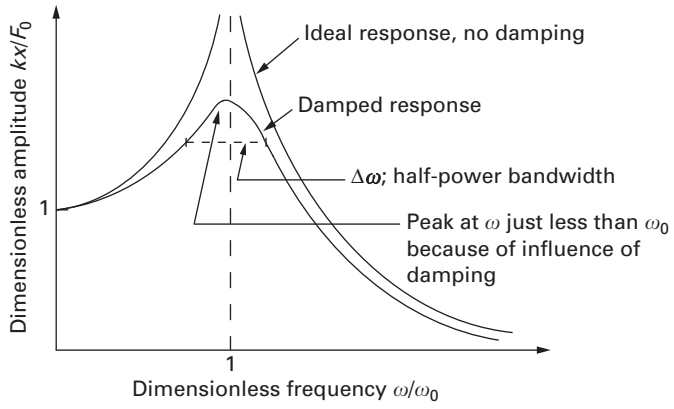


Figure 8.9 A sketch of the shape of the ideal response curve of a spring-mass system driven by a harmonic excitation. The axes are dimensionless: $kx(t)/F_0$ against ω/ω_0. The infinite peak for the ideal case is quite obvious. The damped response is bounded and finite.

We see in eq. (8.64) that the impedance vanishes at resonance, that is, $Z(\omega_0) = 0$. Thus, when the speed of the mass becomes infinite, nothing impedes its motion—even if the magnitude of the force is very small. Thus, an alternate statement of the condition of resonance is that it occurs at the frequency for which the impedance $Z(\omega) = 0$ vanishes.

The form of eq. (8.65) also suggests that the behavior of $Z(\omega)$ might be substantially different for $\omega < \omega_0$ than it would be for $\omega > \omega_0$. In fact, for frequencies below the natural frequency (i.e., for $\omega \ll \omega_0$), eq. (8.64) can be approximated as

$$Z_{\mathrm{k}}(\omega) \cong \frac{m\omega_0^2}{\omega} = \frac{k}{\omega}. \tag{8.65}$$

Thus, for low frequencies, where the excitation is applied slowly, the oscillator responds as a spring: The impedance decreases as the frequency increases toward the natural frequency. For low frequencies, of course, we are closer to the static limit of $\omega = 0$, so it should not be a surprise that stiffness dominates the response.

On the other hand, for frequencies above the natural frequency (i.e., for $\omega \gg \omega_0$), eq. (8.64) can be approximated as

$$Z_{\mathrm{m}}(\omega) \cong -m\omega. \tag{8.66}$$

At high frequencies we expect the dynamics to be more important, and so it is not unexpected that the mass dominates. It is also not surprising that

the impedance increases with frequency, meaning that it gets progressively harder to push around a mass at ever-higher frequencies.

So much for the ideal case. What happens in the "real world" where there is friction and damping and energy loss? The mathematics of modeling damped systems get more complex (see Problems 8.37 and 8.38), so we will present a few key results here. The governing equation for analyzing the dynamic response of a damped oscillator is:

$$m\frac{d^2x(t)}{dt^2} + c\frac{dx(t)}{dt} + kx(t) = F_0 e^{j\omega t}. \tag{8.67}$$

A damping element is included here, and we also have introduced complex arithmetic in the notation for the excitation: The forcing function is written in exponential form (see Sections 4.9 and 7.2.2) and, in order that eq. (8.67) remain real, the forcing amplitude must be a complex number. It can be shown that the square of the magnitude of the resulting motion of a spring-mass-damper is:

$$|x(t)|^2 = \frac{|F_0|^2}{m^2(\omega_0^2 - \omega^2)^2 + c^2\omega^2}, \tag{8.68}$$

while the magnitude of the impedance is:

$$|Z(\omega)|^2 = m^2\omega_0^2\left(\frac{\omega_0}{\omega} - \frac{\omega}{\omega_0}\right)^2 + c^2. \tag{8.69}$$

We note immediately that eqs. (8.68) and (8.69) reduce to their respective counterparts for the ideal model (eqs. (8.61) and (8.64)) when $c = 0$. Further, and still more important, note that the presence of damping eliminates both the singular response and the vanishing of the impedance at resonance. Thus, at resonance, when $\omega = \omega_0$,

$$|x(t)|^2_{\omega_0} = \frac{|F_0|^2}{c^2\omega_0^2}, \tag{8.70}$$

and

$$|Z(\omega_0)|^2 = c^2. \tag{8.71}$$

Equation (8.70) shows that the response is bounded and non-infinite as long as there is damping, and that it becomes infinite when $c = 0$. Equation (8.71) shows that the impedance vanishes altogether only if the damping vanishes altogether. In fact, eqs. (8.69) and (8.71) both also confirm our intuitive sense that damping impedes motion.

Problem 8.33. What are the fundamental physical dimensions of impedance for a mechanical oscillator?

Problem 8.34. Show that the mechanical impedance of an ideal spring-mass system can be written in the form

$$Z(\omega) = \frac{k}{\omega} - m\omega.$$

Explain why this form of impedance does not work as well as eq. (8.64) to discern the stiffness- and mass-controlled regions of response.

Problem 8.35. Write the governing equation for a parallel *LC* circuit subject to a harmonic current input $-(i_0/\omega)\cos(\omega t - \phi)$ and determine the resulting impedance.

Problem 8.36. What are the fundamental physical dimensions of impedance for an electrical oscillator? [*Hints*: Imagine eq. (8.71) and its predecessor with a resistor, R, in place of the damping coefficient, or solve Problem 8.35.]

Problem 8.37. Assume an exponential solution to the homogeneous counterpart of eq. (8.67) and determine the roots for which a solution exist.

Problem 8.38. Determine the particular solution to eq. (8.67) by assuming that $x(t) = B\exp(j\omega t)$, where B and ω may be complex.

Problem 8.39. Determine and explain the dimensions of the coefficients, $(F_0/m\omega_0)$, in eq. (8.58).

Problem 8.40. Sketch the impedance, $Z(\omega)$, of a spring-mass-damper against the dimensionless frequency and identify the regimes where stiffness, mass, or damping controls the response.

8.8 Summary

We have devoted this chapter to the simple harmonic oscillator, without and with damping, without and with a forcing function, and in several different guises. These applications have included the classical mechanical spring-mass system, inductor-capacitor oscillators, a parallel *RLC* circuit, the vibration of tall buildings, and oscillation in a cyclotron and of a vehicle

suspension system. We also developed the electrical-mechanical analogy and pointed out its usefulness for thinking about the meaning of the different terms in the various oscillator models.

In addition, we solved the differential equation and described the solution for the forced harmonic vibration of an oscillator. In so doing, we were able to bring out the very important concepts of resonance and impedance. In discussing impedance, we showed how the various elements (spring, mass, and damper) provided different response regimes, that is, frequency regimes that are controlled, respectively, by stiffness, mass, and damping.

And, finally, we pointed out the commonality of both the mathematics and the physics of such system models. Thus, to develop oscillatory behavior, systems must have elements with stiffness that store potential energy (springs and capacitors) elements with mass that store kinetic energy (masses and inductors), and elements that dissipate energy (dashpots and resistors). Stiffness may take many forms, but there must always be an element that stores potential energy in order for there to be an exchange with an element that stores kinetic energy.

8.9 References

D. P. Billington, *The Tower and the Bridge: The New Art of Structural Engineering*, Basic Books, New York, 1983.

R. E. D. Bishop, *Vibration*, Cambridge University Press, Cambridge, UK, 1965.

R. E. D. Bishop and D. C. Johnson, *The Mechanics of Vibration*, Cambridge University Press, Cambridge, UK, 1960.

M. Braun, *Differential Equations and Their Applications: Shorter Version*, Springer-Verlag, New York, 1978.

R. H. Cannon, Jr., *Dynamics of Physical Systems*, McGraw-Hill, New York, 1967.

P. D. Cha, J. J. Rosenberg, and C. L. Dym, *Fundamentals of Modeling and Analyzing Engineering Systems*, Cambridge University Press, New York, 2000.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

C. L. Dym and I. H. Shames, *Solid Mechanics: A Variational Approach*, McGraw-Hill, New York, 1973.

F. Fahy, *Sound and Structural Vibration*, Academic Press, London, 1985.

M. Farkas, *Dynamical Models in Biology*, Academic Press, San Diego, CA, 2001.

R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vols. I and II, Addison-Wesley, Reading, MA, 1963.
B. R. Gossick, *Hamilton's Principle and Physical Systems*, Academic Press, New York, 1967.
R. Haberman, *Mathematical Models*, Prentice-Hall, Englewood Cliffs, NJ, 1977.
D. Halliday and R. Resnick, *Fundamentals of Physics*, 2nd Edition, Revised Version, John Wiley & Sons, New York, 1986.
G. W. Housner and D. E. Hudson, *Applied Mechanics: Dynamics*, Von Nostrand-Reinhold, New York, 1959.
E. C. Pielou, *An Introduction to Mathematical Ecology*, Wiley Interscience, New York, 1969.
A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, McGraw-Hill, New York, 1981.
J. W. S. Rayleigh, *The Theory of Sound*, Vol. 1, 2nd Edition, Dover Publications, New York, 1945. (The 1st edition was published in 1877.)
E. Simiu and R. H. Scanlan, *Wind Effects on Structures*, John Wiley & Sons, New York, 1978.
J. M. Smith, *Mathematical Ideas in Biology*, Cambridge University Press, London and New York, 1968.
G. W. Swenson, *Principles of Modern Acoustics*, Boston Technical Publishers, Cambridge, MA, 1965.
B. S. Taranath, *Structural Analysis and Design of Tall Buildings*, McGraw-Hill, New York, 1988.
P. A. Tipler, *Physics*, Worth Publishers, New York, 1976.
J. W. Tongue, *Principles of Vibration*, 2nd Edition, Oxford University Press, New York, 2001.
M. R. Wehr and J. A. Richards, Jr., *Physics of the Atom*, Addison-Wesley, Reading, MA, 1960.
R. M. Whitmer, *Electromagnetics*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
H. B. Woolf (Editor), *Webster's New Collegiate Dictionary*, G. & C. Merriam, Springfield, MA, 1977.

8.10 Problems

8.41. The height of a World Trade Center (WTC) tower was 1370 ft (110 stories) and its fundamental period was about 11 sec. The height of the Empire State Building is 1250 ft (102 stories) and its fundamental period is about 8 sec.

(a) How do their respective average specific weights compare?

(b) If the average specific weight for the WTC is as given in Section 8.2 for slender steel-framed towers, what would be the corresponding number for the Empire State Building?

8.42. The height of the Citicorp Building is 915 ft (59 stories) and its fundamental period is about 6.5 sec. Given the data in Problem 8.41 for a WTC tower, find:

(a) how the period varies with building height; and
(b) how the period varies with number of stories.

8.43. Obtain an expression [analogous to eq. (7.28)] for the total energy in a parallel *RLC* circuit and calculate its rate of change with respect to time [analogous to eq. (7.29)].

8.44. Obtain an approximate expression [analogous to eq. (7.30)] for the total energy in a parallel *RLC* circuit that can be used with the results of Problem 8.43 to obtain a differential equation [analogous to eq. (7.29)] for the circuit's energy.

8.45. Use the results of Problem 8.44 to determine how the energy of the parallel *RLC* circuit behaves over time? What is the relevant time constant, and how would you characterize that constant? (*Hint*: Reread Section 7.1.6.)

8.46. (a) Find the impedance of an acoustic resonator as a function of ρ_0, A, L, V_0 and ω; and
(b) What are the physical dimensions of the resonator impedance?

8.47. Charged particles are accelerated in a cyclotron travel in circles of radius r that depends on their speed, v, and magnetic flux density, B, according to:

$$r = \frac{mv}{qB},$$

where m and q are the particle's mass and charge, respectively. The speed and the energy are boosted every half-cycle, so that the particles execute forced harmonic motion in circles whose radii are increasing .

(a) At what resonant frequency ω_0 must the energy be supplied?
(b) What is the impedance of this system?
(c) Show that the rate of change of the energy in the system is of the form

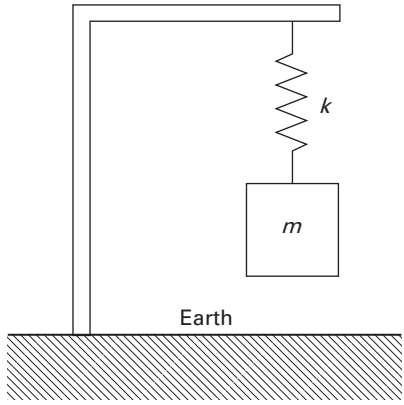
$$\frac{dE(t)}{dt} = \frac{(qB)^3(r_2^2 - r_1^2)}{2\pi m^2} > 0.$$

8.48. A simple seismograph is shown in the accompanying figure. If y denotes the displacement of m relative to the earth, and η the displacement of the earth's surface relative to the fixed stars, the

equation of motion of the mass is

$$m\frac{d^2y(t)}{dt^2} + c\frac{dy(t)}{dt} + ky(t) = -m\frac{d^2\eta(t)}{dt^2}.$$

(a) Determine the steady-state response if $\eta(t) = C\cos\omega t$.
(b) Sketch the amplitude of $y(t)$ as a function of ω.



8.49. What are the dimensions of the *damping quality factor*, $Q = \omega_0 m/c$?

8.50. A long-period seismometer has mass, $m = 0.01$ kg, period, $T_0 = 30$ sec, and damping quality factor, $Q = 3$. An earthquake triggers the earth's surface to respond with a oscillations with a period of 15 minutes and a maximum acceleration of 2×10^{-9} m/sec^2. What is the amplitude of the seismometer vibration?

8.51. Given that power equals force times velocity or speed, determine the *average power* needed to maintain the oscillations of a damped system driven by $F = F_0\cos\omega t$ and responding as $x(t) = X_0\cos(\omega t + \phi)$, where ϕ is the phase angle by which the response lags behind the force.

8.52. For the forced oscillator of Problem 8.51, let the phase angle $\phi = \pi/2$ rad, $\omega_0 = 500$ rad/sec, $Q = 4$ when

$$X_0 = \frac{F_0}{k}\frac{\omega_0/\omega}{\sqrt{\left(\frac{\omega_0}{\omega} - \frac{\omega}{\omega_0}\right)^2 + \left(\frac{1}{Q}\right)^2}},$$

with $\sin\phi = Q(kX_0/F_0)(\omega/\omega_0)$.

(a) Plot the average power input found in Problem 8.51 against the frequency, ω, of the driving force.

(b) Find the width $\Delta\omega$ of the power curve of part (a) at one-half of the maximum power, centered around the resonance frequency. This range of frequencies, the *half-power band*, is that within which resonance effectively occurs.

8.53. (a) Repeat the calculations of Problem 8.52 with a damping quality factor $Q = 6$.

(b) What does a comparison of the two half-power bands for different values of Q reveal about the effect of damping on resonance?

8.54. List resonant systems that we see in nature, over as wide a range as possible.

8.55. A weight hanging on the end of a spring causes a *static deflection* $x_{st} = W/k$. If the static deflection is measured in inches, show that the resonant frequency in cycles per second is $f = 3.13/\sqrt{x_{st}}$ (Hz).

8.56. A bridge is 100 m long and supported by steel beams whose modulus of elasticity is $E = 2\,\text{N/m}^2$ and whose second moment $I = 0.002\,\text{m}^4$. Determine the bridge's natural frequency if its mass is 10^5 kg and a weight of 1.8×10^5 N causes it to deflect 0.01 m?

8.57. A group of 200 soldiers who collectively weigh 1.8×10^5 N marches in step across the bridge of Problem 8.56. Their right feet hit the bridge at regular intervals of 0.9 sec, forcing the bridge to vibrate. Would an observer see that vibration? Explain how you know that.

9
Optimization: What Is the Best…?

This final chapter is about achieving the *best result*, obtaining the *maximum gain*, finding the *optimal outcome*. Thus, this chapter is about *optimization*—an especially interesting subject because finding an optimum result may be difficult, and at times even impossible. Our experience with finding maxima and minima in calculus suggests that we can often find a point where the derivative of a function vanishes and an extreme value exists. But in engineering design and in life generally, we often have to "satisfice," that is, in the word of Herbert A. Simon, be satisfied with an acceptable outcome, rather than an optimal one. Here, however, we will focus on modeling the ways we seek optimal solutions. In so doing, we will see that the formulation of an optimization problem depends strongly on how we express the *objective function* whose extreme values we want and the *constraints* that limit the values that our variables may assume.

Why?
How?

Much of the work on finding optimal results derives from an interest in making good decisions. Many of the ideas about formulating optimization problems emerged during and after World War II, when a compelling need to make the very best use of scarce military and economic resources translated in turn into a need to be able to formulate and make the *best decisions* about using those resources. Thus, with improved decision making as the theme, we will also present (in Section 9.4) a method of choosing the best of an available set of alternatives that can be used in a variety of settings.

We will close with a miscellany of interesting, "practical" optimization problems.

9.1 Continuous Optimization Modeling

We start with a basic minimization problem whose solution is found using
Find? elementary calculus. Suppose that we want to find the minimum values of the *objective function*

$$U(x) = \frac{x^2}{2} - x, \tag{9.1}$$

which we have drawn in Figure 9.1. That picture of the objective function $U(x)$—so called because we set our objective as finding its extreme value—is a parabolic function of x, as the algebraic form of eq. (9.1) confirms.
How? Thus, it has only a single minimum value, called the *global* minimum. The value of x at which this global minimum is found is determined by setting the first derivative of $U(x)$ to zero:

$$\frac{dU(x)}{dx} = x - 1 = 0, \tag{9.2}$$

from which it follows that the minimum value of $U(x)$ occurs when $x_{\min} = 1$ and is

$$U_{\min} = U(x_{\min}) = -\frac{1}{2}. \tag{9.3}$$

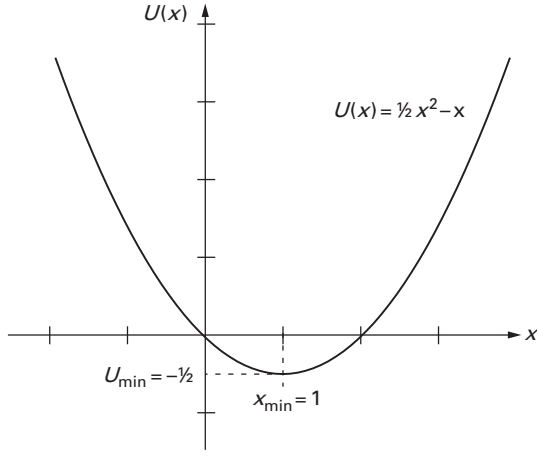


Figure 9.1 The objective function $U(x) = x^2/2 - x$ plotted over the unrestricted range of $-\infty \le x \le +\infty$. The minimum value of the objective function, $U_{\min} = -1/2$, occurs at $x_{\min} = 1$.

We also note from eq. (9.2) that the slope of $U(x)$ increases monotonically as x goes from $-\infty$ to $+\infty$, which means that $U(x)$ itself can have only one flat spot. We can confirm this by calculating the rate of change or derivative of the slope,

$$\frac{d^2U(x)}{dx^2} = 1, \tag{9.4}$$

which is always positive. Thus, there is only one minimum, and it is a global minimum. In fact, we can go a step further and identify the minimum value of eq. (9.3) as an *unconstrained minimum* because we did not constrain or limit the values that the variable x could assume.

Assume?

Suppose we did impose a constraint, say of the form $x \leq x_0$, which requires the independent variable, x, to always be less than or equal to a given constant, x_0. This means that search for the minimum of $U(x)$ is limited to the *admissible values* of x: $x \leq x_0$.We can visualize a procedure for implementing this constraint as putting a line on the same graph as the curve, $U(x)$, and then "moving" this line to different values of x_0, as shown in Figure 9.2. The constraint then shows as the set of lines, $x_{01} < x_{02} < x_{03}$, so we can now briefly consider the three problems of determining the minimum values of $U(x)$ with $x \leq x_{0i}$, $i = 1, 2, 3$. In the first case, $i = 1$, the admissible range of x is so restricted that the constrained minimum value

How?

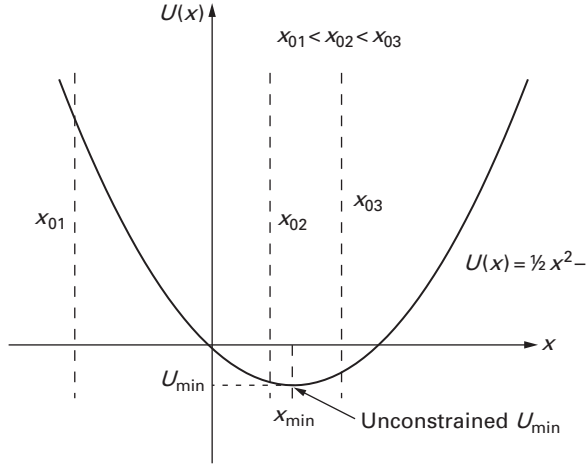


Figure 9.2 The objective function $U(x) = x^2/2 - x$ plotted together with three constraints that restrict the range of admissible values: the set of lines, $x_{01} < x_{02} < x_{03}$. These lines allow us to consider the three problems of determining the minimum values of $U(x)$ with $x \leq x_{0i}$, $i = 1, 2, 3$.

of $U(x)$ is apparently significantly greater than the unconstrained minimum of eq. (9.3). For example, if $x_{01} = -3$, the corresponding constrained minimum is $U(-3) = 7.5$. As the constraint "moves" further to the right ($i = 2, 3$), we approach and then go through the unconstrained minimum. Thus, the range of *feasible solutions* for the minimum of $U(x)$ may include the unconstrained minimum, $U_{\min}$—or it may not—depending on just where the constraint boundaries happen to be.

The constraints so far imposed are *inequality constraints*, $x \leq x_0$, that bound the range of feasible values at the upper end by the equality, $x = x_0$, and include the interior region, $x < x_0$. We might have posed only a simple *equality constraint*, $x = x_0$, in which case we would have found a (highly) constrained minimum $U(x_0)$.

If our objective function were only slightly more complicated, the search for extreme points would become significantly more complicated. Consider the objective function

$$U(x) = \sin x, \tag{9.5}$$

This elementary function could have, depending on the limits placed on the range of admissible values of x, an infinite number of maxima and of minima, or a constrained extremum somewhere between the two (see Problem 9.1). The point of this seemingly trivial example is simple. Characterizing and finding the extrema can be complicated even when the objective function is well known and its properties well understood.

Why? The objective functions (9.1) and (9.5) have only a single variable. However, *multi-dimensional optimization problems* are almost always the norm in engineering practice because engineered devices and processes rarely, if ever, depend only on a single variable. One simple example can be found at the local post office, where postal regulations typically stipulate that the rectangular package shown in Figure 9.3 can be mailed only if the sum of its girth $(2x + 2y)$ and length (z) do not exceed 84 in (2.14 m).

Find? What is the largest volume that such a rectangular package can enclose?

The objective function is the package's volume,

$$V(x, y, z) = xyz, \tag{9.6}$$

Assume? where x and y are the two smaller dimensions whose sum comprises the package's girth, and the length, z, is its longest dimension. We assume that these three dimensions are positive real numbers (i.e., $x > 0$, $y > 0$, $z > 0$).

The constraint on the package dimensions stemming from the postal regulations can be written as:

$$\underbrace{2x + 2y}_{\text{girth}} + \underbrace{z}_{\text{length}} \leq 84, \tag{9.7}$$

Since we seek the largest possible volume, this inequality constraint on the package dimensions can be expressed as an equality constraint: **Assume?**

$$2x + 2y + z = 84. \tag{9.8}$$

Thus, the volume maximization problem is expressed as the objective function (9.6) to be maximized, subject to the equality constraint (9.8). Although the problem is formulated in three dimensions, we can use the equality constraint to eliminate one variable, say the length, z, so that the objective function becomes: **How?**

$$V(x, y) = xy(84 - 2x - 2y) = 84xy - 2x^2y - 2xy^2. \tag{9.9}$$

Now we want to find the maximum value of $V(x, y)$ as a function of x and y. As we recall from calculus, the necessary condition that $V(x, y)$ takes on an extreme value is:

$$\frac{\partial V(x, y)}{\partial x} = 84y - 4xy - 2y^2 = 2y(42 - 2x - y) = 0, \tag{9.10a}$$

and

$$\frac{\partial V(x, y)}{\partial y} = 84x - 2x^2 - 4xy = 2x(42 - x - 2y) = 0. \tag{9.10b}$$

Equations (9.10a–b) can be reduced to a pair of linear algebraic equations whose non-trivial solution can be found ($x = y = 14$ in) to determine the corresponding package volume, $V = 5488$ in^3. This volume can be confirmed to be a maximum (see Problems 9.4, 9.5).

The package problem, albeit multi-dimensional, was still relatively simple because its inequality constraint could logically and appropriately be reduced to an equality constraint that could, in turn, be used to reduce the dimensionality of the problem. Then we found the maximum volume of the package by applying standard calculus tools and seemingly without any further reference to constraints (see Problem 9.6). Consider for a moment the problem of finding the minimum of the following objective function:

$$U(x, y) = x^2 + 2(x - y)^2 + 3y^2 - 11y. \tag{9.11}$$

We show a three-dimensional rendering of this parabolic surface in Figure 9.3. It has an unconstrained minimum at the point ($x = 1, y = 1.5$), where $U_{\min} = -8.25$ (see Problem 9.7). What happens if an equality constraint is imposed? That is, in the style and terminology of the field of operations research, suppose that we want to find the **Find?**

$$\begin{array}{ll} \text{minimum of} & U(x, y) = x^2 + 2(x - y)^2 + 3y^2 - 11y, \\ \text{subject to} & x + y = 3. \end{array} \tag{9.12}$$

We could again use standard calculus techniques to show that the constrained minimum occurs at the point ($x = 31/24$, $y = 41/24$), where $U_{\min} = -385/48$ (see Problem 9.8). Note that the minimum is located on the boundary plane where the constraint intersects $U(x, y)$, that is, at a point such that $x + y = 31/24 + 41/24 = 72/24 = 3$. If the equality constraint of eq. (9.12) was replaced with the (strict) inequality constraint $x + y < 3$, we would find that the minimum sought lies inside the intersecting boundary plane.

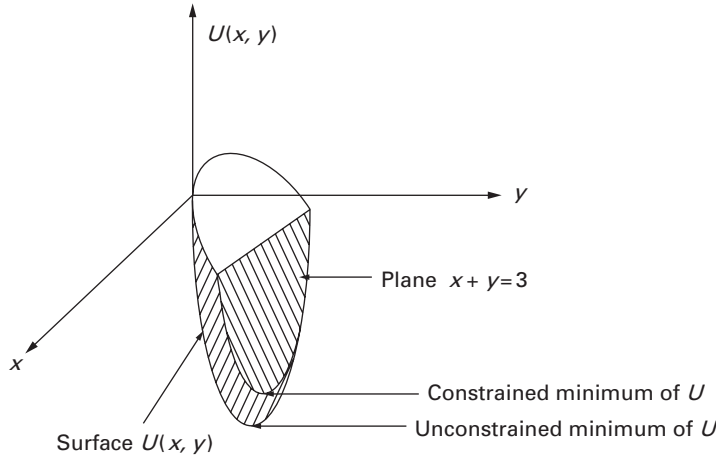


Figure 9.3 The objective function $U(x, y) = x^2 + 2(x - y)^2 + 3y^2 - 11y$ "plotted" in three dimensions, along with the plane $x + y = 3$ that could form the boundary of an equality constraint or of a corresponding inequality constraint.

Problem 9.1. Determine the maxima and minima of the elementary function, $U(x) = \sin x$, when the range of admissible values is:

(a) unconstrained;
(b) $0 < x < \pi/2$;
(c) $0 \leq x \leq \pi/2$; and
(d) $3\pi/4 \leq x \leq 9\pi/4$.

Problem 9.2. Assume an equality constraint is applied to the minimization of the objective function (9.1).

(a) Determine the corresponding value of $U_{\min}$.

(b) How would you characterize that extreme value (e.g., it is an _____ minimum)?

Problem 9.3. Solve the linear algebraic equations (9.10) and determine the three corresponding package dimensions and the package's volume, V.

Problem 9.4. If eqs. (9.10a–b) are the *necessary* conditions to find the maximum value of the function, V, of eq. (9.9), what additional requirements are needed to have *sufficient* conditions to obtain the maximum of V?

Problem 9.5. Apply the sufficient condition(s) found in Problem 9.4 to the package volume problem to confirm that the result calculated in Problem 9.3 is, in fact, a maximum.

Problem 9.6. Are there any "invisible" or implicit constraints in the package maximization problem? (*Hint*: Start with the fact that x, y, and z represent real physical quantities that can never be negative.)

Problem 9.7. Determine the location and value of the minimum of the parabolic surface given by eq. (9.11).

Problem 9.8. Use standard calculus methods to determine the constrained minimum defined by eq. (9.12). (*Hint*: Eliminate a variable.)

9.2 Optimization with Linear Programming

The section just completed showed that the search for an optimum or extreme value of a function subject to an inequality constraint requires a search over the interior of the region defined by the constraint boundary. Thus, as shown in Figure 9.2, we must search for all values of $x \leq x_{0i}$. This is true more generally because an objective function may fluctuate in value, perhaps like the sinusoid of eq. (9.5). Consider, for example, the sketch of a generic objective function in Figure 9.4. The good news is that the standard methods of calculus are usually adequate for searches where the objective functions are relatively tractable. The bad news is that, in such cases, we generally need to search the entire domain, $x_{04} \leq x \leq x_{05}$ to find global optima. However, there is a very important class of problems where a search of the interior region is not required because the optimum point must occur on one of the constraint boundaries. This class of problems is made up of objective functions that are linear functions of the independent variables, and their optimization searches are known as *linear programming* (LP).

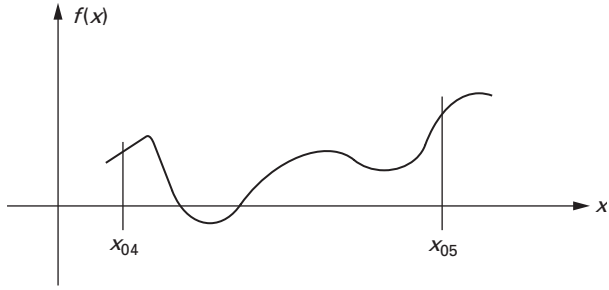


Figure 9.4 A generic sketch of an objective function that shows some variation or fluctuation, with peaks and valleys in the domain of interest. The bad news is that here we do need to search the entire domain, $x_{04} \leq x \leq x_{05}$, to find a global optimum. The good news is that the standard methods of calculus are usually adequate for searches if the objective functions are relatively straightforward.

Suppose we want to find (see Figure 9.5) the

$$\begin{array}{ll} \text{minimum of} & U(x) = mx + b, \\ \text{subject to} & x_1 \leq x \leq x_2. \end{array} \tag{9.13}$$

Now, the minimum of $U(x)$ must lie within the admissible range of values of x, defined by the two inequality constraints just given. Geometry, however, tells us that the optimal values of the linear objective function,

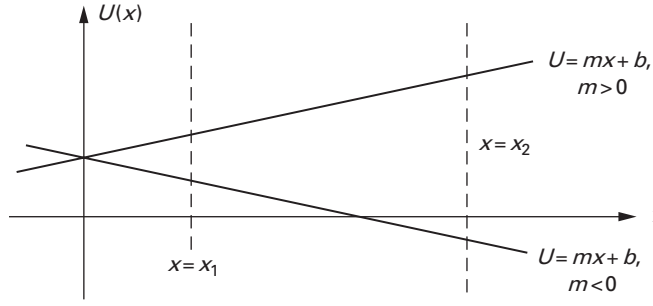


Figure 9.5 A generic *linear programming* problem which is characterized by an objective function that is a linear function of the variable x. Note that the optimal values, both maxima and minima, for $m > 0$ or $m < 0$, occur at points where the objective function intersects the constraint boundaries, that is, on the constraint boundaries themselves.

$U(x) = mx + b$, occur at points where $U(x)$ intersects one of the two constraint boundaries. For $m > 0$, $U_{\min}$ must occur at $x = x_1$ and $U_{\max}$ must occur at $x = x_2$. Thus, for this linear programming problem, we can find the optima of $U(x)$ without searching the interior region defined by the constraint boundaries: We know *a priori* that the optima must occur on the constraint boundaries. In fact, it can be shown that the optimum solutions for LP problems are found by searching only at the boundary intersections or *vertices*. The search problem is thus "reduced" to solving for a set of intersection points defined by various linear equations.

Is requiring an objective function to be linear too much of a simplification? Are LP problems useful, or a cute mathematical artifact? In fact, LP is extremely important and useful, and is one of the cornerstones of the field of *operations research*. The field of operations research (OR)—pronounced "oh r"—developed first in Britain and then in the United States during World War II when there was a compelling interest in optimizing scarce military and economic resources. Since that time, OR has been applied to both military and civil problems, including in the latter a wide variety of commercial enterprises, allocating medical resources, managing traffic, and modeling the criminal justice system. The hallmark of LP is the determination of optimal results for *single* objectives: *minimizing* transportation costs, *optimizing* the product mix, *maximizing* hospital bed availability, *minimizing* the number of highway toll attendants when traffic is slack, or *minimizing* drivers' waiting times when traffic is heavy.

9.2.1 Maximizing Profit in the Furniture Business

Suppose that we are in the furniture business and making desks and tables that are made of oak and maple. Desks and tables consume different amounts of lumber: a desk requires 6 board-feet (bft) each of oak and maple, while a table requires 3 bft of oak and 9 bft of maple. The local lumber mill will supply up to 1200 bft of oak at \$6.00/bft and up to 1800 bft of maple at \$4.00/bft. The market for desks and tables is such that they can be sold for, respectively, \$90.00 and \$84.00. How many desks and how many tables should we make to maximize our profit?

Why?
Given?
Assume?
Find?

We will soon find out (see eq. (9.16)) that under the conditions assigned here, the profits earned by selling a desk are the same as the profits earned by selling a table, namely, \$30.00 each. Suppose that this was not the case and that the profit in selling a table was only \$18.00. Then it might seem reasonable to first make only desks to maximize profit—except that we will run out of oak after only 200 desks are made and have an excess, unusable supply of maple left over. It will also turn out that the profit earned in this case, \$6000, is not the maximum profit possible. This problem is

interesting because the constraints supply *limits* on the available materials, which means in turn that we must make *trade-offs* between desks and tables to maximize our overall profit.

How? We formulate this profit optimization problem as an LP problem, meaning that we build an objective function—the difference between sales income and cost of manufacture—and the relevant operating constraints. If x_1 is the number of desks made, and x_2 the number of tables, the income derived by selling desks and tables is:

$$\begin{aligned}\text{income} &= (\$/\text{desk})x_1 + (\$/\text{table})x_2 \\ &= \$(90x_1 + 84x_2). \end{aligned} \tag{9.14}$$

The cost of manufacture is reckoned in terms of the quantity of lumber required for each product and the unit costs of that lumber:

$$\begin{aligned}\text{cost} =& (\$/\text{oak})[(\text{oak}/\text{desk})x_1 + (\text{oak}/\text{table})x_2] \\ &+ (\$/\text{maple})[(\text{maple}/\text{desk})x_1 + (\text{maple}/\text{table})x_2] \\ =& (\$6.00)(6x_1 + 3x_2) + (\$4.00)(6x_1 + 9x_2) \\ =& \$(60x_1 + 54x_2). \end{aligned} \tag{9.15}$$

The profit to be maximized, the objective function, is the difference between the income (eq. (9.14)) and the cost (eq. (9.15)):

$$\begin{aligned}\text{profit} &= \$(90x_1 + 84x_2) - \$(60x_1 + 54x_2) \\ &= \$(30x_1 + 30x_2). \end{aligned} \tag{9.16}$$

Note that the income, cost, and profit are all expressed in a common unit of currency, in this case U.S. dollars ($).

There are three constraints for this linear programming problem, deriving from the limitations of the wood's availability from the lumber mill and the fact that wood is a real physical object. On availability, the manufacturer can't use any more wood than the lumber mills can make available, so that

$$\begin{aligned}\text{amount of oak used} &= (\text{oak}/\text{desk})x_1 + (\text{oak}/\text{table})x_2 \\ &= 6x_1 + 3x_2 \leq 1200 \text{ (bft)}, \end{aligned} \tag{9.17a}$$

and

$$\begin{aligned}\text{amount of maple used} &= (\text{maple}/\text{desk})x_1 + (\text{maple}/\text{table})x_2 \\ &= 6x_1 + 9x_2 \leq 1800 \text{ (bft)}. \end{aligned} \tag{9.17b}$$

Note that both constraints are expressed in terms of a common dimension, board-feet (see Problems 9.9 and 9.10).

The third constraint follows from the simple fact that the numbers of tables and desk must be real, positive numbers. Thus, there is a *non-negativity constraint*:

$$x_1, x_2 \geq 0. \tag{9.18}$$

So, to sum up our furniture LP problem, then, we want to find the

$$\begin{array}{ll} \text{maximum of} & \$(30x_1 + 30x_2), \\ \text{subject to} & \begin{cases} 6x_1 + 3x_2 \leq 1200(\text{bft}), \\ 6x_1 + 9x_2 \leq 1800(\text{bft}), \\ x_1, x_2 \geq 0. \end{cases} \end{array} \tag{9.19}$$

Note that the non-negativity constraint is almost always a part of LP formulations, largely because the variables involved in LP or optimization problems are real physical variables that by their very nature are greater than (or sometimes equal to) zero. In addition, it is fairly easy to convert "negative" variables to "positive" variables by suitable sign changes in the objective function and in the constraints. Finally, and likely most importantly, there is a significant computational advantage when all of the variables are positive because the non-negativity constraint limits the admissible space of the variables substantially. **How?**

The solution to the LP problem posed in eq. (9.19) will be found graphically. In Figure 9.6 we show the admissible space as the first quadrant in the (x_1, x_2) plane; the objective function (9.16) as a series of dotted lines; and the two inequality constraints (9.17) as (labeled) solid lines. The feasible region, wherein *all* (in this case *both*) constraints are satisfied, is shaded. The objective function can be thought of as a series of parallel dotted lines that can take on different, *increasing* values as the variables x_1 and x_2 increase (see Problem 9.11). We observe that the objective function will reach its largest value when it reaches the point (150, 100) because it is the last point on the constraint boundary that is within the feasible space—here at the intersection of the two inequality constraints (9.17). The value attained by the objective function at that point is \$6300. **How?**

We might have spotted this solution immediately on the basis of our previous observation that LP optima must lie on the constraint boundary. In the present case, the objective function, $30(x_1 + x_2)$, must reach its maximum at one of the vertices formed by the intersection of the constraint boundary with the feasible region ((200, 0), (0, 200)) or with each other (150, 100). Clearly, the point among these three that produces the largest objective function is that at the intersection of the two inequality constraints. So, again, the furniture maker will reap the most profit by making (and selling) 150 desks and 100 tables, which will yield a profit of \$6300. **Valid?** **Use?**

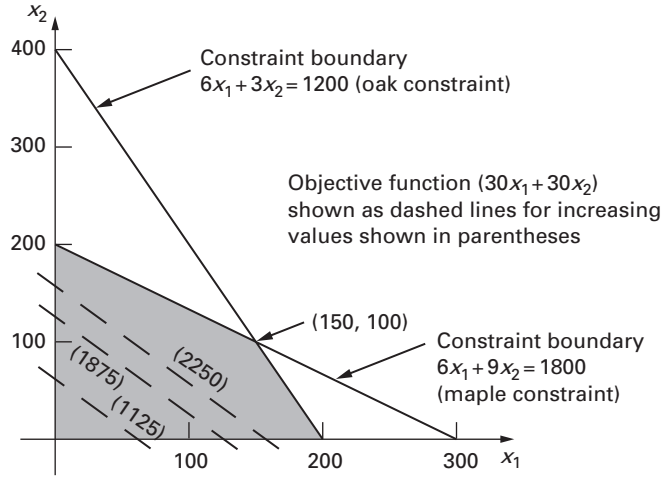


Figure 9.6 The graphical solution to the LP problem posed in eq. (9.19). Note that the admissible space is the first quadrant in the (x_1, x_2) plane; the objective function (9.16) is portrayed in a series of dotted lines; and the two inequality constraints (9.17) are the (labeled) solid lines. The region of feasible solutions, wherein *all* (in this case *both*) constraints are satisfied, is shaded.

9.2.2 On Linear Programming and Extensions

The example presented just above is rather simple because it has only two variables, which made it amenable to graphical solution. LP problems often have hundreds or even thousands of variables and so cannot be handled graphically, but they can be solved straightforwardly with a variety of standard computational approaches. All of these approaches to LP problems, the most notable of which is the *simplex* method, work by identifying the boundary vertices at which optima must lie in ways that are analogous to our graphical solution. There are also many other classical OR/LP problems, including the *feed-mix* and *product-mix* problems that occur repeatedly in industry, and the *transportation* problem that we will discuss in the next section.

There are other so-called programming problems and methodologies that are used to solve more complicated optimization problems. For example, *nonlinear programming* (NLP) refers to the set of techniques used when the objective functions are nonlinear. *Dynamic programming* refers to the class of problems that require sequential, hierarchical decisions, that is, problems where the solution to one problem serves as input to or a starting

point for another problem. For example, in an extension of the furniture problem just solved, the lumber prices might vary over time because of external supply factors, in which case different production decisions might be made. Similarly, *integer programming* is designed to deal with those problems in which variables must be treated as integers, rather than as continuous variables. For example, we treated the numbers of desks and tables as continuous variables in the furniture optimization problems, but it is hard to imagine that we would make 150.7 desks or 99.6 tables. The results could, of course, be rounded off, but then we lose our guarantee of optimality. More importantly, however, integer programming is used for problems that have binary variables (for example, zero or one). For example, *scheduling* problems, such as when an election agency locates polling places within particular zip codes, are the kind of "go or no go" situations where integer programming is of most use.

9.2.3 On Defining and Assessing Optima

The optimal or best solution may well depend on the perspective of the person conducting or sponsoring the study. Someone has to say what he or she means by "the best." Consider the three cities that are spaced as in Figure 9.7. New connecting highways are to be built between the three cities, and the taxpayers clearly want the best result. The question is: What is the best result? If the best result is defined as the shortest travel time between two adjacent cities, then the configuration shown in Figure 9.7(a) will be the best. If the best result is defined as the least amount of road construction, then the configuration shown in Figure 9.7(b) will be the best. Thus, as the old saying goes, "Where you stand depends on where you sit." Optimizing travel time (the commuter's perspective) may be the best, but minimizing road construction (the taxpayer's perspective) may also be the best.

Further, the choices are rarely as simple as that. The reduction of commuter travel time may lead to tangible economic benefits, perhaps from more rapid delivery of goods, perhaps from a reduced pollution burden, or it may prompt more travel that increases noise and air pollution. A careful calculation of such benefits might be used to decrease the net cost of the first highway configuration, which might change the assessment of which highway pattern is truly the best. Clearly, doing such calculations requires the assignment of economic values to waiting time, delivery time, inventory costs, pollution burdens, and to other aspects of "reality" that may be relevant. Even the choice between scenic and direct routes can be modeled as an economic choice because it reflects a value judgment about whether

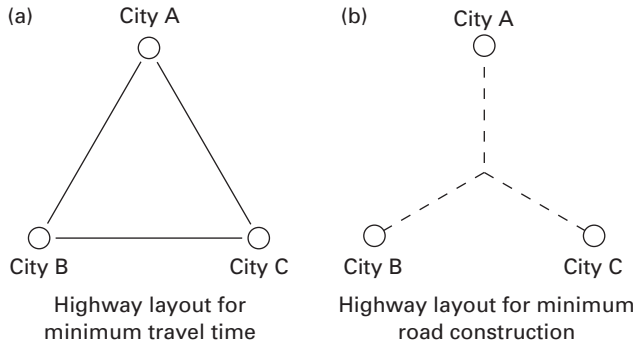


Figure 9.7 Alternate highway configurations for connecting three cities. (a) The first configuration minimizes travel time between adjacent cities. (b) The second configuration minimizes the amount of road construction needed.

it is more important to enjoy the landscape or to get to the destination as quickly as possible. Thus, three important points are:

- With LP and other OR techniques, we are modeling *decisions*, rather than physical behavior or the like.
- When making such decisions, we are making *trade-offs* between *costs* and *benefits*.
- When formulating and modeling such decisions, we are using cost-benefit analysis to make explicit our values and preferences.

Problem 9.9. Verify the dimensions of eqs. (9.14–9.16).

Problem 9.10. Verify the dimensions of eqs. (9.17a–b).

Problem 9.11. Determine the slope of the dotted lines in Figure 9.6 that represents the objective function (9.16) of the furniture LP problem. How does it compare with the slopes of the two constraint boundaries?

9.3 The Transportation Problem

Why? Having decided in Section 9.2 how many desks and tables we must make in order to maximize our profit from making furniture, we now turn to selling our desks through a *distribution network* of furniture outlets. The stores are at varying distances from the furniture maker's two plants—we have

been doing well and have expanded our operations!—and each store has its own demand, based on its own marketing analyses. Thus, we have the logistical problem of deciding how to allocate the desks among the stores. This class of OR problems is called the *transportation problem*.

Three furniture stores have ordered desks: Mary's Furniture Emporium wants 30, Lori's Custom Furniture wants 50, and Jenn's Furniture Bazaar wants 45. We have made 70 desks at Plant 1 and another 80 at Plant 2. The distances between the two plants and the three stores are given in Table 9.1, and the shipping cost is $1.50 per mile per desk. We want to minimize the shipping costs of filling the three orders. Since the cost of shipping a desk is easily calculated (see Problem 9.12), we have to calculate how many desks go from a specified plant (of two) to a particular store (of three).

Given?

What?

How?

Table 9.1 The distances (in miles) between Plants 1 and 2, where the desks are made, to Mary's Furniture Emporium, Lori's Custom Furniture, and Jenn's Furniture Bazaar, where the desks will be sold.

	Product		
Plant	**(1) Mary's**	**(2) Lori's**	**(3) Jenn's**
1	10	5	30
2	7	20	5

Is an optimal solution available easily? Doesn't it seem reasonable to ship first along the shortest (and cheapest) routes? Here we would send 50 desks from Plant 1 to Lori's, 45 desks from Plant 2 to Jenn's, and so on; and this approach might yield the optimal solution in this case (see Problem 9.13). However, this is a simple problem that has only two plants and three stores, and thus only six plant-store combinations to consider. In addition, the supply of desks is distributed so that the demand for all three stores can be met with only the shortest routes (see Problem 9.14). Thus, this simple problem does not need the kind of trade-off among alternatives that was needed to maximize profit for making furniture. However, it is a useful template for more complex problems, so let us solve it in a formal way.

This transportation problem of shipping desks from two plants to three stores can be represented as an elementary *network problem*, as depicted in Figure 9.8. The circles represent *nodes* at which desks are either supplied (the plants) or consumed (the stores). In more elaborate problems the nodes may be points to which material is supplied and from which material is distributed. The directed line segments (the arrows) are *links*

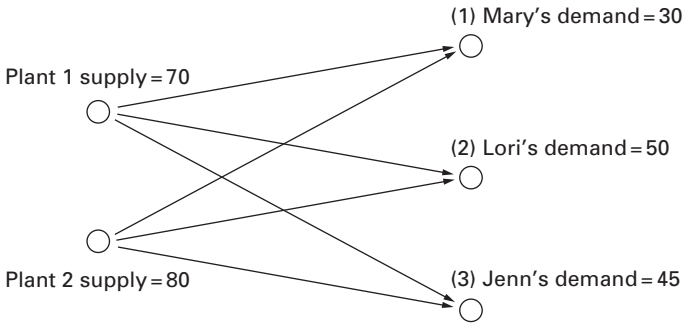


Figure 9.8 A *network representation* of the elementary transportation problem of shipping desks from two plants to three stores. The circles represent *nodes* at which desks are either supplied (the plants) or consumed (the stores). The directed line segments (the arrows) are *links* that represent the routes along which the desks could be shipped.

that represent the routes along which the desks could be shipped, and in more elaborate problems, these directed line segments may thus signify two-way or bi-directional links.

One possible—although *sub-optimal*—solution to this transportation problem is shown in Figure 9.9. We can calculate the shipping cost for this solution as \$1387.50 (see Problem 9.15), which is substantially higher than the optimal solution (see Problem 9.13 again).

The transportation problem can be formulated as an LP problem, for which some additional notation will be useful. Thus, we now identify x_{ij} as the number of desks shipped from Plant $i = 1, 2$ to Store $j = 1, 2, 3$. (Note

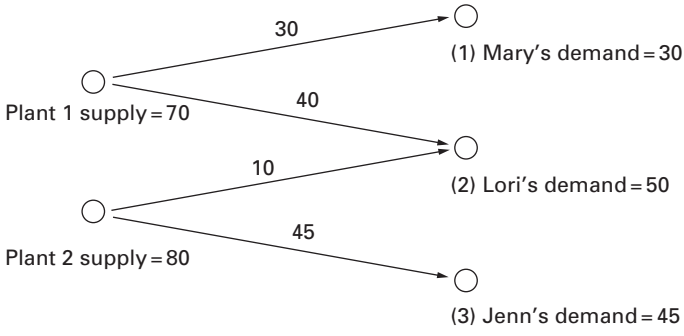


Figure 9.9 One possible solution to the elementary transportation problem of shipping desks from Plants 1 and 2 to Mary's Furniture Emporium, Lori's Custom Furniture, and Jenn's Furniture Bazaar.

that the two plants were numbered from the beginning, and the stores were assigned numbers in Table 9.1.) Then we have

x_{11} = number of desks from Plant 1 to Store 1 (Mary's),
x_{12} = number of desks from Plant 1 to Store 2 (Lori's),
x_{13} = number of desks from Plant 1 to Store 3 (Jenn's).
x_{21} = number of desks from Plant 2 to Store 1 (Mary's),
x_{22} = number of desks from Plant 2 to Store 2 (Lori's), and
x_{23} = number of desks from Plant 2 to Store 3 (Jenn's).

Since the unit cost of shipping is a constant ($1.50 per desk per mile), we can use the data in Table 1 to establish an objective function that is equal to the shipping cost:

$$\text{shipping cost} = (\$1.50)(10x_{11} + 5x_{12} + 30x_{13} + 7x_{21} + 20x_{22} + 5x_{23}). \tag{9.20}$$

The constraints for this problem arise from the supply of desks produced by the two plants and the demand for the desks by the three stores. The two plants cannot exceed their capacities for producing desks:

$$\begin{aligned} x_{11} + x_{12} + x_{13} &\leq 70, \\ x_{21} + x_{22} + x_{23} &\leq 80. \end{aligned} \tag{9.21}$$

The stores, in turn, must have enough desks shipped to them to meet their demands:

$$\begin{aligned} x_{11} + x_{21} &\geq 30, \\ x_{12} + x_{22} &\geq 50, \\ x_{13} + x_{23} &\geq 45. \end{aligned} \tag{9.22}$$

Finally, the numbers of desks must satisfy a non-negativity constraint because the desks are real, so that:

$$x_{ij} \geq 0. \tag{9.23}$$

Thus, to sum up the formulation of our shipping problem as an LP problem, we want to find the

$$\text{minimum of} \quad (\$1.50)(10x_{11} + 5x_{12} + 30x_{13} + 7x_{21} + 20x_{22} + 5x_{23}),$$

$$\text{subject to} \quad \begin{cases} x_{11} + x_{12} + x_{13} \leq 70, \\ x_{21} + x_{22} + x_{23} \leq 80, \\ x_{11} + x_{21} \geq 30, \\ x_{12} + x_{22} \geq 50, \\ x_{13} + x_{23} \geq 45, \\ x_{ij} \geq 0. \end{cases} \tag{9.24}$$

The shipping problem posed thus far has supply exceeding demand. A more restricted version, the *classical transportation problem*, sets the total supply equal to the total demand. The five inequality constraints (9.21) and (9.22) become simple equality constraints, which reduces the number of independent constraints by one. Suppose that Plant 1 produces only 45 desks (instead of 70). This reduces the (total) supply to a level that equals the demand level of 125 desks. Thus, the supply constraints (9.21) become

$$\begin{aligned} x_{11} + x_{12} + x_{13} &= 45, \\ x_{21} + x_{22} + x_{23} &= 80, \end{aligned} \tag{9.25}$$

whose sum, a constraint on the total supply, can then be found to be:

$$x_{11} + x_{12} + x_{13} + x_{21} + x_{22} + x_{23} = 125. \tag{9.26}$$

Similarly, the demand constraints (9.22) become

$$\begin{aligned} x_{11} + x_{21} &= 30, \\ x_{12} + x_{22} &= 50, \\ x_{13} + x_{23} &= 45, \end{aligned} \tag{9.27}$$

and their sum, a constraint on the total demand, adds up to the same result as for the total supply [eq. (9.26)]. Since the total demand and the total supply equations are the same, the set of constraints (9.25) and (9.27) represent only four—not five—independent equations.

This reduction in the number of independent constraints produces some real computational benefits in solving classical transportation problems. One of the benefits is that all of the variables turn out to be integers when the constraints are expressed as integers. Further, the "extra" constraint produced by equating supply to demand results in comparatively straightforward and efficient computations of the optimum.

If the demand exceeds the supply, the LP model cannot even get started because it is impossible to get into the feasible region—from which the solutions derive. This is clear from summing the supply constraint inequalities (9.21),

$$\sum_{i,j} x_{ij} \leq \text{supply}, \tag{9.28}$$

and comparing it to the sum of the demand inequalities (9.22),

$$\sum_{i,j} x_{ij} \geq \text{demand}. \tag{9.29}$$

Of course, if supply exceeds demand, so that there is a net, positive surplus, an LP solution can proceed in a straightforward fashion.

Problem 9.12. Formulate and verify the dimensions of the equation needed to calculate the cost of shipping a desk from either plant to any of the three stores.

Problem 9.13. Complete the easily available optimal solution to the desk shipping problem and determine the minimum shipping cost.

Problem 9.14. Which of the six available plant-store routes were used in achieving the optimal solution of Problem 9.13?

Problem 9.15. Find the actual shipping cost of the transportation solution shown in Figure 9.9.

Problem 9.16. Confirm the objective function (9.20). (*Hint*: Use the result of Problem 9.12.)

Problem 9.17. Verify the constraints (9.21) and (9.22).

Problem 9.18. Verify that the sums of the supply (9.25) and demand (9.28) equality constraints add to the same sum as eq. (9.26).

9.4 Choosing the Best Alternative

Why?

People choose among alternatives all of the time: voters rank candidates; designers rank objectives; and students rank colleges. In each of these circumstances, the *voter* or *decision maker* is charged with choosing among the alternatives.

9.4.1 Rankings and Pairwise Comparisons

In recent years, questions have been raised about *how* voters establish rankings of alternatives. Further, since people seem to compare objects in a list on a pairwise basis before rank ordering the entire list, there is a special focus on how pairwise comparisons are performed as a means of assembling information for doing rank orderings. In *pairwise comparisons*, the elements in a set (i.e., the candidates, design objectives, or colleges) are ranked two at a time, on a pair-by-pair basis, until all of the permutations have been exhausted. Points are awarded to the winner of each comparison. Then the points awarded to each element in the set are summed, and the rankings are obtained by ordering the elements according to points accumulated. However, it is worth noting that as both described here and

practiced, the number of points awarded in such pairwise comparisons is often non-uniform and arbitrarily weighted. But, as we will note below, it is quite important that the points awarded be measured in fixed increments.

The pairwise comparison methodology has been criticized particularly because it violates the famous *Arrow impossibility theorem* for which Kenneth J. Arrow was awarded the 1972 Nobel Prize in Economics. In that theorem, Arrow proved that a perfect or fair voting procedure cannot be developed whenever there are more than two candidates or alternatives that are to be chosen. He started by analyzing the properties that would typify a *fair* election system, and stated (mathematically) that a voting procedure can be characterized as fair *only* if four axioms are obeyed:

1. *Unrestricted*: All conceivable rankings registered by individual voters are actually possible.
2. *No Dictator*: The system does not allow one voter to impose his/her ranking as the group's aggregate ranking.
3. *Pareto Condition*: If every individual ranks A over B, the societal ranking has A ranked above B.
4. *Independence of Irrelevant Alternatives* (IIA): If the aggregate ranking would choose A over B when C is not considered, then it will not choose B over A when C is considered.

Arrow proved that at *least one of these properties must be violated* for problems of reasonable size (at least three voters expressing only ordinal preferences among more than two alternatives). It is worth noting that a consistent social choice (voting) procedure can be achieved by violating any one of the four conditions. Further, some *voting procedures* based on pairwise comparisons are faulty in that they can produce ranking results that offend our intuitive sense of a reasonable outcome—and quite often a desired final ranking can be arrived at by specifying a voting procedure.

Among pairwise comparison procedures, the *Borda count* (which we describe below, in Section 9.4.2) most "respects the data" in that it avoids the counter-intuitive results that can arise with other methods. As D. G. Saari notes, the Borda count "*never elects the candidate which loses all pairwise elections … always ranks a candidate who wins all pairwise comparisons above the candidate who loses all such comparisons.*"

The Borda count does violate Arrow's final axiom, the *independence of irrelevant alternatives* (IIA). What does it mean that IIA is violated? And, is that important? The meaning depends to some extent on the domain and whether or not there are meaningful alternatives or options that are being excluded. In an election with a finite number of candidates, the IIA axiom is likely not an issue. In conceptual design, where the possible space of design choice is large or even infinite, IIA could be a problem. However, rational designers must find a way to limit their set of design alternatives to a finite,

relatively small set of options. Thus, options that don't meet some criteria or are otherwise seen as poor designs may be eliminated. It is unlikely that IIA matters much if it is violated for one of these two reasons—and there is some evidence to support this—unless it is shown that promising designs were wrongly removed early in the process.

The violation of IIA leads to the possibility of *rank reversals*, that is, changes in order among n alternatives that may occur when one alternative is dropped from a once-ranked set before a second ranking of the remaining n–1 alternatives (as we will soon see below). The elimination of designs or candidates *can* change the tabulated rankings of those designs or candidates that remain under consideration. The determination of which design is "best" or which candidate is "preferred most" may well be sensitive to the set of designs considered.

Rank reversals occur when there are *Condorcet cycles* in the voting patterns: $[A \succ B \succ C, B \succ C \succ A, C \succ A \succ B]$. When aggregated over all voters and alternatives, these cycles cancel each other out because each option has the same Borda count. When one of the alternatives is removed, this cycle no longer cancels. Thus, removing C from the above cycle unbalances the Borda count between A and B, resulting in a unit gain for A that is propagated to the final ranking results. Thus, the rank reversals symbolize a loss of information that occurs when an alternative is dropped or removed from the once-ranked set.

We now describe a way to use pairwise comparisons in a structured approach that parallels the role of the Borda count in voting procedures and, in fact, produces results that are identical to the accepted vote-counting standard, the Borda count. The method is a structured extension of pairwise comparisons to a *pairwise comparison chart* (PCC) or matrix. The PCC produces consistent results quickly and efficiently, and these results are identical with results produced by a Borda count.

9.4.2 Borda Counts and Pairwise Comparisons

We begin with an example that highlights some of the problems of (non-Borda count) pairwise comparison procedures. It also suggests the equivalence of the Borda count with a structured pairwise comparison chart (PCC).

Twelve (12) voters are asked to rank order three candidates: A, B, and C. In doing so, the twelve voters have, collectively, produced the following sets of orderings:

$$\begin{array}{ll} 1 \text{ preferred } A \succ B \succ C, & 4 \text{ preferred } B \succ C \succ A, \\ 4 \text{ preferred } A \succ C \succ B, & 3 \text{ preferred } C \succ B \succ A. \end{array} \tag{9.30}$$

Pairwise comparisons other than the Borda count can lead to inconsistent results for this case. For example, in a widely used plurality voting process called *the best of the best*, A gets 5 first-place votes, while B and C each get 4 and 3, respectively. Thus, A is a clear winner. On the other hand, in an "antiplurality" procedure characterized as *avoid the worst of the worst*, C gets only 1 last-place vote, while A and B get 7 and 4, respectively. Thus, under these rules, C could be regarded as the winner. In an iterative process based on *the best of the best*, if C were eliminated for coming in last, then a comparison of the remaining pair A and B quickly shows that B is the winner:

$$\begin{array}{ll} 1 \text{ preferred } A \succ B, & 4 \text{ preferred } B \succ A, \\ 4 \text{ preferred } A \succ B, & 3 \text{ preferred} B \succ A. \end{array} \tag{9.31}$$

On the other hand, a Borda count produces a clear result. The Borda count procedure assigns numerical ratings separated by a common constant to each element in the list. Thus, sets such as (3, 2, 1), (2, 1, 0) and (10, 5, 0) could be used to rank a three-element list. If we use (2, 1, 0) for the rankings presented in eq. (9.30), we find total vote counts of $(A\colon 2+8+0+0 = 10)$, $(B\colon 1+0+8+3 = 12)$ and $(C\colon 0+4+4+6 = 14)$, which clearly shows that C is the winner. Furthermore, if A is eliminated and C is compared only to B in a second Borda count,

$$\begin{array}{ll} 1 \text{ preferred } B \succ C, & 4 \text{ preferred } B \succ C, \\ 4 \text{ preferred } C \succ B, & 3 \text{ preferred } C \succ B. \end{array} \tag{9.32}$$

C remains the winner, as it also would here by a simple vote count. It must be remarked that this consistency cannot be guaranteed, as the Borda count violates the IIA axiom.

We now make the same comparisons in a PCC matrix, as illustrated in Table 9.2. As noted above, a point is awarded to the winner in each pairwise comparison, and then the points earned by each alternative are summed. In the PCC of Table 9.2, points are awarded row-by-row, proceeding along

Table 9.2 A pairwise comparison chart (PCC) for the ballots cast by twelve (12) voters choosing among the candidates *A*, *B* and *C* (see eq. (9.30)).

Win/Lose	A	B	C	Sum/Win
A	— —	$1+4+0+0$	$1+4+0+0$	10
B	$0+0+4+3$	— —	$1+0+4+0$	12
C	$0+0+4+3$	$0+4+0+3$	— —	14
Sum/Lose	14	12	10	— —

each row while comparing the row element to each column alternative in an individual pairwise comparison. This PCC result shows that the rank ordering of preferred candidates is entirely consistent with the Borda results just obtained:

$$C \succ B \succ A. \tag{9.33}$$

Note that the PCC matrix exhibits a special kind of symmetry, as does the ordering in the "Win" column (largest number of points) and the "Lose" row (smallest number of points): the sum of corresponding off-diagonal elements, $X_{ij} + X_{ji}$, is a constant equal to the number of comparison sets.

We have noted that a principal complaint about some pairwise comparisons is that they lead to rank reversals when the field of candidate elements is reduced by removing the lowest-ranked element between orderings. (Strictly speaking, rank reversal can occur when any alternative is removed. In fact, and as we note further in Section 9.4.3, examples can be constructed to achieve a specific rank reversal outcome. Such examples usually include a dominated option that is not the worst. Also, rank reversals are possible if new alternatives are *added*.) Practical experience suggests that the PCC generally preserves the original rankings if one alternative is dropped. If element A is removed above and a two-element runoff is conducted for B and C, we find the results given in Table 9.3. Hence, once again we find

$$C \succ B. \tag{9.34}$$

The results in inequality (9.34) clearly preserve the ordering of inequality (9.33), that is, no rank reversal is obtained as a result of applying the PCC approach. In those instances where some rank reversal does occur, it is often among lower-ranked elements where the information is strongly influenced by the removed element (see Section 9.4.3).

Table 9.3 A reduced pairwise comparison chart (PCC) for the problem in Table 9.2 wherein the "loser" A in the first ranking is removed from consideration.

Win/Lose	B	C	**Sum/Win**
B	— —	$1+0+4+0$	5
C	$0+4+0+3$	— —	7
Sum/Lose	7	5	— —

9.4.3 Pairwise Comparisons and Rank Reversals

Rank reversals do sometimes occur when alternatives are dropped and the PCC procedure is repeated. We now show how such an example can be constructed.

Thirty (30) designers (or consumers) are asked to rank order five designs, A, B, C, D, and E, as a result of which they produce the following sets of orderings:

$$\begin{aligned} &10 \text{ preferred} \quad A \succ B \succ C \succ D \succ E, \\ &10 \text{ preferred} \quad B \succ C \succ D \succ E \succ A, \\ &10 \text{ preferred} \quad C \succ D \succ E \succ A \succ B. \end{aligned} \tag{9.35}$$

Here too, the procedure chosen to rank order these five designs can decidedly influence or alter the results. For example, all of the designers ranked C and D ahead of E in the above tally. Nonetheless, if the following sequence of pairwise comparisons is undertaken, an inconsistent result obtains:

$$C \text{ vs } D \Rightarrow C; \;\; C \text{ vs } B \Rightarrow B; \;\; B \text{ vs } A \Rightarrow A; \;\; A \text{ vs } E \Rightarrow E. \tag{9.36}$$

Table 9.4 shows the PCC matrix for this five-design example, and the results clearly indicate the order of preferred designs to be:

$$C \succ B \succ D \succ A \succ E. \tag{9.37}$$

If the same data are subjected to a Borda count, using the weights (4, 3, 2, 1, 0) for the place rankings, we then find the results displayed in Table 9.5. When we compare these results to the PCC results shown in Table 9.4, we see that the PCC has achieved the same Borda count results, albeit in a slightly different fashion.

Table 9.4 A collective pairwise comparison chart (PCC) for a case before alternatives are dropped and the PCC is repeated.

Win/Lose	A	B	C	D	E	Sum/Win
A	— —	$10+0+10$	$10+0+0$	$10+0+0$	$10+0+0$	50
B	$0+10+0$	— —	$10+10+0$	$10+10+0$	$10+10+0$	70
C	$0+10+10$	$0+0+10$	— —	$10+10+10$	$10+10+10$	90
D	$0+10+10$	$0+0+10$	$0+0+0$	— —	$10+10+10$	60
E	$0+10+10$	$0+0+10$	$0+0+0$	$0+0+0$	— —	30
Sum/Lose	70	50	30	60	90	— —

What happens if we drop the lowest-ranked design and redo our assessment of alternatives? Here design E is least preferred, and we find the results shown in Table 9.5 if it is dropped. These results show a rank ordering of

$$C \succ B \succ A \succ D. \tag{9.38}$$

Rank order is preserved here for the two top designs, C and B, while the last two change places. Why does this happen? Quite simply, because of the relative narrowness of the gap between A and D when compared to the gap between A and E, the two lowest ranked in the first application of the PCC in this example.

Table 9.5 The Borda count with weights (4, 3, 2, 1) for the case where alternative E is dropped and the PCC is repeated. Compare these results with those in Table 9.4.

Element	Points
A	$40 + 10 + 20 = 70$
B	$30 + 40 + 10 = 80$
C	$20 + 30 + 40 = 90$
D	$10 + 20 + 30 = 60$

It is also useful to "reverse engineer" this example. Evidently, it was constructed by taking a Condorcet cycle [$A \succ B \succ C$, $B \succ C \succ A$, $C \succ A \succ B$] and replacing C with an ordered set ($C \succ D \succ E$) that introduces two dominated (by C) options that are irrelevant by inspection. Removing only E produces a minor rank reversal of the last two alternatives, A and D. Removing only D, the third best option, produces the same result among A, B, and C as removing E, although without creating a rank reversal. Removing both D and E produces a tie among A, B, and C.

9.4.4 Pay Attention to All of the Data

We now present an example that shows how pairwise ranking that does not consider other alternatives can lead to a result exactly opposite to a Borda count, which does consider other alternatives. It also indicates that attempting to select a single best alternative may be the wrong approach.

One hundred (100) customers are "surveyed on their preferences" with respect to five mutually exclusive design alternatives, A, B, C, D, and E. The survey reports that "45 customers prefer A, 25 prefer B, 17 prefer C,

13 prefer D, and no one prefers E." These data suggest that A is the preferred choice, and that E is entirely "off the table."

However, as reported, these results assume either that the customers are asked to list only one choice or, if asked to rank order all five designs, that only their first choices are abstracted from their rank orderings. Suppose that the 100 customers were asked for rankings and that those rankings are:

$$\begin{aligned} &45 \text{ preferred } A \succ E \succ D \succ C \succ B, \\ &25 \text{ preferred } B \succ E \succ D \succ C \succ A, \\ &17 \text{ preferred } C \succ E \succ D \succ B \succ A, \\ &13 \text{ preferred } D \succ E \succ C \succ B \succ A. \end{aligned} \tag{9.39}$$

Again, the procedure used to choose among the rank orderings of these five designs can decidedly influence or alter the results. For example, if A and B are compared as a (single) pair, B beats A by a margin of 55 to 45. And, continuing a sequence of pairwise comparisons, we can find that (see Problem 9.23):

$$A \text{ vs } B \Rightarrow B; \;\; B \text{ vs } C \Rightarrow C; \;\; C \text{ vs } D \Rightarrow D; \;\; D \text{ vs } E \Rightarrow E. \tag{9.40}$$

Proposition (9.40) provides an entirely different outcome, one that is not at all apparent from the vote count originally reported. How do we sort out this apparent conflict?

We resolve this dilemma by constructing a PCC matrix for this five-product example, and the results clearly indicate the order of preferred designs to be (see Problem 9.24):

$$E(300) \succ D(226) \succ A(180) \succ C(164) \succ B(130). \tag{9.41}$$

A Borda count of the same data (of eqs. (9.39)), using the weights (4, 3, 2, 1, 0) for the place rankings, confirms the PCC results, with the Borda count numbers being identical to those in eq. (9.41) (see Problem 9.25). In this case, removing B and re-voting generates a relatively unimportant rank reversal between A and C, thus demonstrating the meaning of IIA and showing that dropping information can have consequences.

This example is one where the "best option" as revealed by the PCC/Borda count is not the one most preferred by anyone. Is the PCC lying to us? In a real market situation, where all five options are available, none of the surveyed customers would buy E. Perhaps this data was collected across too broad a spectrum of customers in a very segmented market in which design E provided a "common denominator," while the other four designs responded better to their separate market "niches." There is really no "best design" under these circumstances. It is also possible that these

designs were extremely close to each other in performance, so that small variations in performance have translated into large differences in the PCC. Both of the above explanations point to the need to treat PCC results with caution because there are cases where more detailed selection procedures might be more appropriate.

9.4.5 On Pairwise Comparisons and Making Decisions

The structured PCC—an implementation of the Borda count—can support consistent decision making and choice, notwithstanding concerns raised about pairwise comparisons and violations of Arrow's theorem. Rank reversals and other infelicities do result when "losing" alternatives are dropped from further consideration. But simulation suggests that such reversals are limited to alternatives that are nearly indistinguishable. Pairwise comparisons that are properly aggregated in a pairwise comparison chart (PCC) produce results that are identical to the Borda count, which in Saari's words is a "unique positional procedure which should be trusted."

Practicing designers use the PCC and similar methods very early in the design process where rough ordinal rankings are used to bound the scope of further design work. The PCC is more of a discussion tool than a device intended to aggregate individual orderings of design team members into a "group" decision. Indeed, design students are routinely cautioned against over-interpreting or relying too heavily on small numerical differences. In political voting, we usually end up with only one winner, and any winner *must* be one of the entrants in the contest. In early design, it is perfectly fine to keep two or more winners around, and the ultimate winner often does not appear on the initial ballot. Indeed, it is often suggested that designers look at *all* of the design alternatives and try to incorporate the good points of each to create an improved, composite design. In this framework, the PCC is a useful aid for understanding the strengths and weaknesses of individual design alternatives. Still, pairwise comparison charts should be applied carefully and with restraint. As noted above, it is important to cluster similar choices and to perform the evaluations at comparable levels of detail.

In addition, given the subjective nature of these rankings, when we use such a ranking tool, we should ask *whose* values are being assessed. Marketing values are easily included in different rankings, as in product design, for example, where a design team might need to know whether it's "better" for a product to be cheaper or lighter. On the other hand, there might be deeper issues involved that, in some cases, may touch upon the fundamental values of both clients and designers. For example, suppose

two competing companies, GRAFT and BJIC, are trying to rank order design objectives for a new beverage container. We show the PCCs for the GRAFT- and BJIC-based design teams in Tables 9.6(a) and (b), respectively. It is clear from these two charts and the scores in their right-hand columns that the GRAFT designers were far more interested in a container that would generate a strong brand identity and be easy to distribute than in it being environmentally benign or having appeal for parents. At BJIC, on the other hand, the environment and taste preservation ranked more highly, thus demonstrating that subjective values show up in PCCs and, eventually, in the marketplace!

Table 9.6 Using PCCs to rank order design objectives at two different companies designing new beverage containers (Dym and Little, 2003).

Goals	Environ. Benign	Easy to Distribute	Preserve Taste	Appeals to Parents	Market Flexibility	Brand ID	Score
(a) GRAFT's weighted objectives							
Environ. Benign	● ● ●	0	0	0	0	0	0
Easy to Distribute	1	● ● ●	1	1	1	0	4
Preserve Taste	1	0	● ● ●	0	0	0	1
Appeals to Parents	1	0	1	● ● ●	0	0	2
Market Flexibility	1	0	1	1	● ● ●	0	3
Brand ID	1	1	1	1	1	● ● ●	5
(b) BJIC's weighted objectives							
Environ. Benign	● ● ●	1	1	1	1	1	5
Easy to Distribute	0	● ● ●	0	0	1	0	1
Preserve Taste	0	1	● ● ●	1	1	1	4
Appeals to Parents	0	1	0	● ● ●	1	1	3
Market Flexibility	0	0	0	0	● ● ●	0	0
Brand ID	0	1	0	0	1	● ● ●	2

It is also tempting to take our *ranked* or ordered objectives and put them on a *scale* so that we can manipulate the rankings in order to attach relative weights to goals or to do some other calculation. It would be nice to be able to answer questions such as: *How much more* important is portability than cost in a ladder? Or, in the case of a beverage container, *How much more* important is environmental friendliness than durability? A little more? A lot more? Ten times more? We can easily think of cases where one of the objectives is substantially more important than any of the others, such as safety compared to attractiveness or to cost in an air traffic control system, and other cases where the objectives are essentially very close to

one another. However, and sadly, there is *no mathematical foundation* for normalizing the rankings obtained with tools such as the PCC. The numbers obtained with a PCC are *approximate*, *subjective* views or judgments about relative importance. We must *not* inflate their importance by doing further calculations with them or by giving them unwarranted precision.

Valid?

Use?

Problem 9.19. Would you find election procedures that violated Arrow's third axiom offensive? Explain your answer.

Problem 9.20. Would you find election procedures that violated the Pareto condition, Arrow's fourth axiom, offensive? Explain your answer.

Problem 9.21. Engineering designers often use quantified performance rankings to compare alternatives on the basis of measurable criteria. If this comparison were done on a pairwise basis, would it violate Arrow's fourth axiom? Explain your answer.

Problem 9.22. Defend or refute the proposition that ranking criteria that are of the less-is-better, more-is-better, or nominal-is-best varieties will violate Arrow's first axiom. (*Hint*: Are all theoretically possible orders admissible in practice?)

Problem 9.23. Verify the ordering of the five alternatives displayed in eq. (9.40) by performing the appropriate individual pair-by-pair comparisons.

Problem 9.24. Construct a PCC of the data presented in eq. (9.39) and confirm the Borda count results given in eq. (9.41).

Problem 9.25. Using the weights (4, 3, 2, 1, 0), perform a Borda count of the preferences expressed in eq. (9.39) and confirm the results obtained in eq. (9.41) and in the previous problem.

9.5 A Miscellany of Optimization Problems

In this section we present some simple yet interesting optimization and "Can we do better?" problems. Their interest derives more from their subject matter than from the optimization technique applied. As a result, some elementary models are introduced and described in just enough detail to make the search for optimum behavior meaningful. These optimization

problems include forming nuclei in solids, maximizing the range of planes or birds, and reducing the weight of a cantilever beam. Along the way we will introduce some further wrinkles in the modeling of searches for a good—if not globally optimum—method of searching for an optimum result.

9.5.1 Is There Enough Energy to Create a Sphere?

Nucleation refers to the formation of tiny, even submicroscopic, particles. Such particles or *nuclei* initiate the phase transformations in which the microstructures of materials are changed during various materials processes. For example, steel alloys come in various forms (e.g., cementite and ferrite) that have substantially different properties (e.g., ferrite is softer than cementite, but it is also less brittle). How do such nuclei form?

Why?

Given?

The nucleation process occurs in a solution that has, for example, a small number of β atoms relative to a much larger number of α atoms. The β atoms diffuse together, form a small volume and then re-arrange into a crystal structure that is enclosed in a volume, V, with an interfacial (with the surrounding α atoms) area, A. This process can occur only if an *activation energy barrier* is overcome. In the simplest formulation, wherein the distribution of the interfacial energy is *isotropic* (or independent of direction within the solution), the total *free energy exchange* ΔG needed to bring about this change is given by

$$\Delta G = -(\Delta G_V - \Delta G_S)V + \gamma A, \tag{9.42}$$

where $(\Delta G_V - \Delta G_S)$ is the (positive) difference between the volume free energy and the misfit strain energy, and γ is the surface free energy per unit area. Further, the misfit strain energy reduces the free energy exchange because it is subtracted from the volume free energy. Note that all of the energy terms are *specific*, expressed as they are in terms of energy per unit volume of β.

Notwithstanding this meager, skeletal introduction to some of the language of thermodynamics, the important point is that the free energy exchange needed to allow creation of a volume, V, is the sum of a term that decreases with V but increases with its area, A. Is there a point below which the free energy exchange cannot happen? If so, what is the value of that free energy exchange barrier?

Predict?

Assume?

To answer these questions,we assume that the volume will form, at least initially, a sphere of radius r. With appropriate substitutions for the sphere's

surface area and volume, eq. (9.42) becomes

$$\Delta G = -(\Delta G_V - \Delta G_S)\left(\frac{4}{3}\pi r^3\right) + \gamma\left(\pi r^2\right). \tag{9.43}$$

We can now employ the standard techniques of calculus to show that there is a minimum radius, r^*, below which there is not enough free energy to overcome the free energy exchange barrier, ΔG^* (see Problems 9.25 and 9.26):

$$r^* = \frac{2\gamma}{(\Delta G_V - \Delta G_S)}, \tag{9.44}$$

and

$$\Delta G^* = \frac{16\pi\gamma^3}{3(\Delta G_V - \Delta G_S)^2}. \tag{9.45}$$

The free energy exchange is plotted on Figure 9.10 and it shows the barrier that needs to be overcome quite clearly. It also shows how the behavior of the free energy exchange depends differently on r, depending on whether the sphere is smaller or larger than that with the minimum radius, r^* (see Problems 9.27 and 9.28).

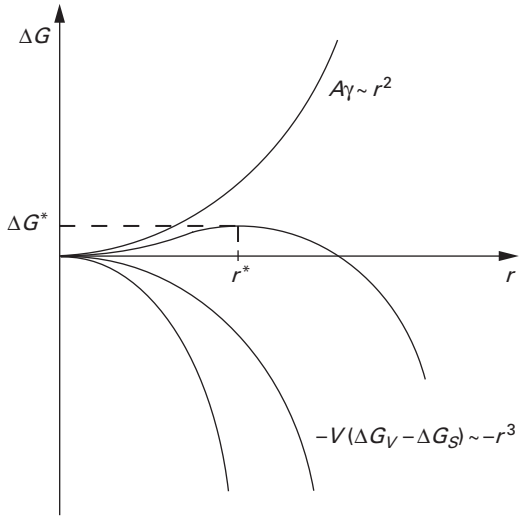


Figure 9.10 The variation of the free energy exchange, ΔG, with the radius, r, of a nucleating sphere (Porter and Easterling, 1992). We see that there is an activation energy barrier, ΔG^*, that must be overcome, and that the free energy exchange decreases for values of $r < r^*$, while it increases for $r > r^*$.

Problem 9.25. Verify eqs. (9.44) and (9.45) by performing the appropriate calculus.

Problem 9.26. Demonstrate that eqs. (9.44) and (9.45) have the correct dimensions.

Problem 9.27. Write eq. (9.43) in terms of a dimensionless coordinate, $\rho = r/r^*$, and expand it in a power series valid for *small* values of ρ. What part of the curves in Figure 9.10 does that result portray?

Problem 9.28. Expand eq. (9.43) in a power series valid for *large* values of the dimensionless coordinate $\rho = r/r^*$. What part of the curves in Figure 9.10 does that result portray?

Problem 9.29. Is the surface area-to-volume ratio of a cylinder of radius, R, and length, L, smaller or larger than that of a sphere of radius R? (Hint: Write a ratio of the ratios as a function of R/L.)

9.5.2 Maximizing the Range of Planes and Birds

Why? Airplane pilots share a challenge with flying birds: How far can they go—what is their *range*—for a fixed amount of fuel? Still better, can they maximize their range? It turns out that for a given amount of fuel, the speed that maximizes the range is the one that maximizes the aerodynamic quantity, called the *lift-to-drag ratio,* or, conversely, minimizes its inverse, the *drag-to-lift ratio.*

Given? We show a typical jet in Figure 9.11 with a free-body diagram (FBD) superposed. The plane is climbing at an angle, α, at a speed, V, relative to the ground. The climb or flight direction angle, α, is zero for level flight, and positive for ascending flight and negative for descending flight. The FBD shows the forces that act to support the plane and move it forward, as described in the aerodynamic literature. The plane's weight, W, is supported by a *lift* (force), L, that is perpendicular to the flight path. The engines provide a thrust, T, that moves the plane along the flight path by overcoming the *drag* (force), D, that also acts along the flight path, albeit it in a direction that retards flight. Due largely to preceding experimental work and subsequent confirming analysis, aerodynamicists have known since the end of the 17th century that the lift and drag forces on a flying body can be expressed in terms of the density of the surrounding air, ρ, the wing or *lifting surface* area, S, and the body's speed, V, as,

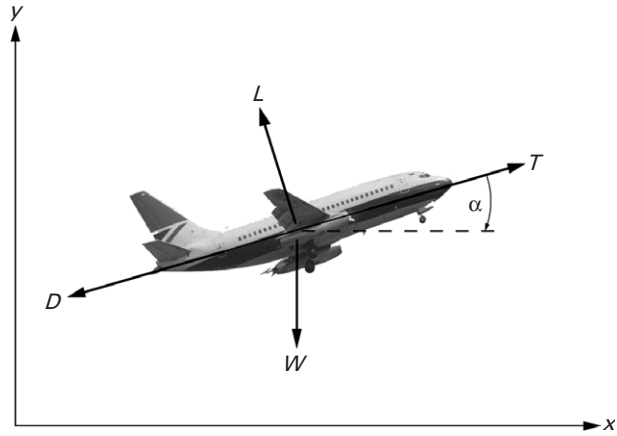


Figure 9.11 A typical jet with a superposed free-body diagram showing the aerodynamic forces acting. The plane is climbing at an angle, α, at a speed, V, relative to the earth below. The plane's weight, W, is supported by a *lift* (force), L, that is perpendicular to the flight path. The jet engines provide a thrust, T, that moves the plane along the flight path by overcoming the *drag* (force), D, that also acts along the flight path, although in a direction that opposes flight. The plane's wing has a surface area, S, and span, b.

respectively,

$$L = \frac{1}{2}\rho S V^2 C_{\mathrm{L}}, \tag{9.46a}$$

and

$$D = \frac{1}{2}\rho S V^2 C_{\mathrm{D}}, \tag{9.46b}$$

where C_{L} and C_{D} are the corresponding *lift* and *drag coefficients*. (We should note that the drag-velocity relation is more complicated when planes fly closer to the speed of sound, due to drag produced by compressibility effects either on rapidly rotating propellers or on the wings of jet aircraft.) The makeup of the C_{L} and C_{D} coefficients and their relationship provide the complexity we will see in our search for an optimum flight speed. But first we need to do a little equilibrium analysis because taken superficially,

eqs. (9.46a–b) suggest that the drag-to-lift ratio L/D is independent of the speed V, so how could it be minimized with respect to V?

We sum the forces superposed on the plane in Figure 9.11 in the x and y directions:

$$\sum F_x = -T\cos\alpha + L\sin\alpha + D\cos\alpha = 0, \tag{9.47a}$$

and

$$\sum F_y = T\sin\alpha + L\cos\alpha - W - D\sin\alpha = 0. \tag{9.47b}$$

If the climb angle, α, is assumed to be small, along the lines of the approximations introduced in Section 4.1.2, eqs. (9.47a–b) can be simplified and solved to show that the lift L is, in fact, a constant (see Problem 9.30),

$$L \cong \frac{W}{1+\alpha^2} \cong W, \tag{9.48a}$$

which means that the *drag-to-lift ratio* is simply

$$D/L \cong D/W. \tag{9.48b}$$

Equation (9.48a) clearly shows that the lift force supports the plane's weight, while eq. (9.48b) provides a speed-dependent ratio of the drag force to the weight.

Now we return to the drag coefficients because that is the logical step for casting the D/L ratio in terms of the plane's speed, V. It turns out that the drag coefficient is expressed as a sum of two terms,

$$C_D = C_{D_0} + \frac{kSC_L^2}{\pi b^2}. \tag{9.49}$$

The first term represents the *parasite* or *friction drag* caused by shear stresses resulting from the air speeding over and separating from the wing. The second term is the *induced drag*: it is independent of the air viscosity and is created by wings of finite span (i.e., real wings!) because of momentum changes needed to produce lift, according to Newton's second law. Note that the induced drag is proportional to the square of the lift coefficient, C_L^2.

Now we can combine eqs. (9.46b) and (9.48b) to write the drag-to-lift ratio as

$$\frac{D}{L} = \frac{\rho S V^2 C_D}{2W}, \tag{9.50}$$

after which we can further combine eqs. (9.46a), (9.48a), and (9.49) to rewrite eq. (9.50) as (see Problem 9.31):

$$\frac{D}{L} = C_{01}V^2 + C_{02}V^{-2}, \tag{9.51}$$

with the *constants* C_{01} and C_{02} defined as:

$$C_{01} = \frac{\rho S C_{D_0}}{2W}, \quad C_{02} = \frac{2kW}{\pi\rho b^2}. \tag{9.52}$$

Thus, the objective function or *cost* for this optimization problem is defined in eq. (9.51), and its coefficients as presented in eq. (9.52) are simply constants reflecting the values of the problem's physical parameters: ρ, S, W, the wing span, b, the parasite drag coefficient, C_{D_0}, and a dimensionless shape constant, k (see Problem 9.32).

The extreme value of this unconstrained optimization problem is then found by the standard calculus approach, that is,

$$\frac{d}{dV}\left(\frac{D}{L}\right) = 2C_{01}V - 2C_{02}V^{-3} = 0, \tag{9.53}$$

which has the following extreme value:

$$\left(\frac{D}{L}\right)_{\min} = 2\sqrt{C_{01}C_{02}} \quad \text{at} \quad V_{\min} = \left(\frac{C_{02}}{C_{01}}\right)^{1/4}. \tag{9.54}$$

With the aid of eq. (9.52), the minimum drag-to-lift ratio can then be written in its final form (see Problem 9.33):

$$\left(\frac{D}{L}\right)_{\min} = 2\sqrt{\frac{kSC_{D_0}}{\pi b^2}}. \tag{9.55}$$

This is a classical result in aerodynamics. Further, it is also easily demonstrated that this minimum D/L ratio occurs only when the parasite drag and the induced drag are equal and, consequently, independent of the plane weight W (see Problem 9.34). In the next section we will obtain this result again by introducing still another method of searching for optimal results.

Problem 9.30. Solve eqs. (9.47a–b) for $(T-D)$ as a function of L and W and confirm that eqs. (9.48a–b) are correct while identifying any additional needed approximations.

Problem 9.31. Combine eqs. (9.46a), (9.48a), and (9.49) and develop eq. (9.51).

Problem 9.32. Show that the constants C_{01} and C_{02} have the correct physical dimensions. (*Hints*: What are their physical dimensions according to eq. (9.51)? Do they have those dimensions?)

Problem 9.33. Use the standard calculus test to confirm that the value of D/L given in eqs. (9.54) and (9.55) is a minimum.

Problem 9.34. Show that the induced drag equals the parasite drag at the minimum D/L ratio, and that both are independent of the plane weight, W.

Problem 9.35. The minimum of eq. (9.51) can also be seen "by inspection." Inspect eq. (9.51) and explain why that minimum can be so determined.

9.5.3 Geometric Programming for a Plane's Optimum Speed

The objective function (9.51) for the plane range problem considered just above is a member of the class of functions called *posynomials*, polynomials whose coefficients are always positive. Clarence Zener, inventor of the Zener diode, noted that if the objective functions whose minima were being sought were posynomials, then each *term* in such an objective function could be considered an independent variable whose contribution to the overall minimum sum could be established. Zener proposed doing that by constructing a dual function that would be maximized. The mathematician Richard J. Duffin recognized that a posynomial cost function could be viewed as a *weighted arithmetic mean*, and Zener's dual function as its *weighted geometric mean*. Cauchy's inequality—the arithmetic mean is always greater than or equal to its geometric mean—could be brought to bear, and thus Zener's optimization invention became known as *geometric programming*.

How?

Consider a rectangle bounded by lines of length a and b. Geometrically, then, the rectangle's perimeter is $P = 2(a + b)$ and its area is $A = ab$. The Greeks asked, What is the smallest perimeter of a rectangle of given area? Well, the answer to that equation is not hard to find. The perimeter can be written as

$$P = 2(a + b) = 4\left[\left(\frac{a}{2} + \frac{b}{2}\right)^2\right]^{1/2}, \tag{9.56}$$

from which it follows that (see Problem 9.36):

$$P = 4\left[ab + \left(\frac{a-b}{2}\right)^2\right]^{1/2} \geq 4\sqrt{ab}. \tag{9.57}$$

Thus, in terms of the rectangle's perimeter and area, eqs. (9.56) and (9.57) tell us that (see Problem 9.37):

$$P \geq 4\sqrt{A}. \tag{9.58}$$

Equation (9.58) also tells us something else. For any two numbers a and b, we can define their *arithmetic mean*, $\overline{a}_{\text{arith}} = \frac{1}{2}(a+b)$, and their *geometric mean*, $\overline{a}_{\text{geom}} = \sqrt{ab}$. Then eq. (9.58) turns out to be a very simple expression of the *Cauchy inequality*:

$$\overline{a}_{\text{arith}} \geq \overline{a}_{\text{geom}}. \tag{9.59}$$

For a collection of numbers or functions, U_i, the Cauchy inequality becomes

$$\overline{U}_{\text{arith}} = \frac{1}{N}\sum_{i=1}^{N} U_i \geq \prod_{i=1}^{N} U_i^{1/N} = \overline{U}_{\text{geom}}. \tag{9.60}$$

Equation (9.60) can be generalized still further. Consider that each object in the sum that is the arithmetic mean is weighted by a positive constant, w_i. Then the extended Cauchy inequality is:

$$\overline{U}_{\text{arith}} = \sum_{i=1}^{N} w_i U_i \geq \prod_{i=1}^{N} U_i^{w_i} = \overline{U}_{\text{geom}}. \tag{9.61}$$

Finally, if we define a set of modified numbers or functions, $V_i = w_i U_i$, we can write the central inequality of eq. (9.61) as

$$\sum_{i=1}^{N} V_i \geq \prod_{i=1}^{N} \left(\frac{V_i}{w_i}\right)^{w_i}, \tag{9.62a}$$

or, written *in extenso*,

$$V_1 + V_2 + \cdots V_N \geq \left(\frac{V_1}{w_1}\right)^{w_1}\left(\frac{V_2}{w_2}\right)^{w_2}\cdots\left(\frac{V_N}{w_N}\right)^{w_N}. \tag{9.62b}$$

The weights, w_i, are restricted in two ways that reflect their roots in geometry. The first is that they must satisfy a *normality condition*, that is, their values must sum to one:

$$w_1 + w_2 + \cdots w_N = 1. \tag{9.63}$$

The second restriction is an *orthogonality condition*, which requires that the geometric mean terms on the right-hand sides of eqs. (9.62a–b) must be free of—or dimensionless in—the independent variables that make up the functions, V_i. The weights that satisfy both the normality and orthogonality conditions then *maximize the geometric mean* and, consequently, *minimize the arithmetic mean.* This lovely piece of geometry brings us back to our optimization problem.

We start with an objective or cost function that is written as a sum of posynomials, $V_i(x)$, each of which is a function of some or all of a set of independent design variables, $x = (x_1, x_2, \ldots x_k)$:

$$V(x) = V(x)_1 + V_2(x) + \cdots V_N(x) = \sum_{i=1}^{N} V_i(x). \tag{9.64}$$

We then define the following weighted product as the *dual* to the cost function, $U(x)$:

$$d(w) = \left(\frac{V_1}{w_1}\right)^{w_1} \left(\frac{V_2}{w_2}\right)^{w_2} \cdots \left(\frac{V_N}{w_N}\right)^{w_N} = \prod_{i=1}^{N} \left(\frac{V_j}{w_j}\right)^{w_j}, \tag{9.65}$$

where $w = (w_1, w_2, \ldots w_k)$ is the set of weights that satisfy the appropriate normality and orthogonality conditions. By the geometric analysis culminating in eqs. (9.62a–b), we can then say that:

$$\min V(x) = \max d(w). \tag{9.66}$$

To illustrate the application of geometric programming (GP), consider once again the determination of the optimum D/L ratio for maximizing a plane's range. The objective function is the D/L ratio given in eq. (9.51), and it is clearly a sum of two posynomials: $V_1 = C_{01} V^2$ and $V_2 = C_{01} V^{-2}$. The corresponding dual function can then be constructed as defined by eq. (9.65):

$$d_\gamma(w) = \left(\frac{C_{01} V^2}{w_1}\right)^{w_1} \left(\frac{C_{02} V^{-2}}{w_2}\right)^{w_2} = \left(\frac{C_{01}}{w_1}\right)^{w_1} \left(\frac{C_{02}}{w_2}\right)^{w_2} \left(V^{2(w_1 - w_2)}\right) \tag{9.67}$$

The orthogonality condition that renders eq. (9.67) dimensionless with respect to the independent variable V is:

$$2w_1 - 2w_2 = 0. \tag{9.68}$$

In conjunction with the appropriate ($N = 2$) version of the normality condition (9.63), eq. (9.68) produces the weights $w_1 = w_2 = 1/2$, which

can then immediately be substituted into the dual function (9.67) to yield the minimum D/L ratio:

$$d_\gamma(w) = \left(\frac{C_{01}}{1/2}\right)^{1/2} \left(\frac{C_{02}}{1/2}\right)^{1/2} = 2\sqrt{C_{01}C_{02}} \tag{9.69}$$

This result is, of course, exactly the same as the one we obtained before (see eq. (9.54)).

Two features of this solution are worth special note. The first is that the solution proceeded quite directly to the sought minimum D/L ratio. This contrasts with the calculus solution, which yielded first the velocity at which the minimum ratio occurs, with the ratio itself being determined after its additional calculation from the critical value of the velocity. The second point is that this solution was remarkably simple. In principle, and by extension to more complicated cases (see Problem 9.40), all we had to do was solve a set of linear equations—the normality and orthogonality conditions—to immediately obtain the optimum we were after. The typical calculus approach, for such evidently nonlinear cost functions, clearly requires much more work.

Finally on GP, we note that we have presented GP in its simplest form. We have not dealt with any constraints, whether equality or inequality. The principles applied to more complicated, more "real" problems are similar—but they will require more work. Given the role of computers in our lives, techniques such as GP are not invoked much any more. However, GP still offers a neat and direct approach to an interesting class of (posynomial) problems.

Problem 9.36. Construct the steps that get one from eq. (9.56) to eq. (9.57).

Problem 9.37. When does the soft inequality in eq. (9.58) become a simple equality?

Problem 9.38. Use the principle of induction to prove the general statement (9.60) of Cauchy's inequality.

Problem 9.39. Can eq. (9.43)—in the discussion of nucleation in Section 9.5.1—be cast into a form suitable for solution by geometric programming? Explain your answer.

Problem 9.40. Minimize the objective function $2x_1^2x_2^{-1} + 4x_2^3x_3^{1/2} + x_1^{-2}x_2x_3^{-1/2} + 2x_2^{-3}$ using geometric programming.

Problem 9.41. Confirm the results of Problem 9.40 using the standard calculus approach of determining extrema.

9.5.4 The Lightest Diving Board (or Cantilever Beam)

Cantilever beams are ubiquitous in life, appearing as diving boards, trees, tall slender buildings, freeway signs, and the arms of grandparents picking up their grandchildren. Their optimal design will vary with their situational circumstance. We present here an analysis that leads toward significantly improved designs that may or may not be optimal. The question answered by such analysis is, therefore, much less like "What is the best … ?" and much more like "Can we do better than … ?".

Find? Our "doing better" problem is simple. We want to determine the profile or shape of a tip-loaded cantilever beam that weighs significantly less while **Given?** yielding the same tip deflection, δ, for a given tip load, P. We will use the classic elementary model of beam bending, the origin of which can be traced to Galileo Galilei (1564–1642), to model the cantilever beam shown in Figure 9.12. This widely applicable model assumes that the beam is long and slender, meaning that its thickness, h, and width, b, are both small compared to its length, L, and that its response to an applied load is almost entirely due to bending, meaning that the stress through the thickness is distributed as

$$\sigma_{xx}(x, z) = \frac{M(x)z}{I}, \tag{9.70}$$

where $\sigma_{xx}(x, z)$ is the axial stress that occurs when the beam is bent, $M(x)$ is the moment that forces the beam to bend, z the vertical coordinate measured positive downward from the centerline of the beam's cross-sectional area, and I is the second moment of that area (see Figure 9.12).

Assume? We will assume that the cross-sectional area is rectangular, with constant width but thickness that may vary with the axial coordinate, x. In this case the second moment, I, is

$$I(x) = \frac{bh^3(x)}{12}. \tag{9.71}$$

Finally, the bending theory of beams states that the moment produced by a force, P, at the cantilever tip ($x = 0$) is $M(x) = -Px$, and that the resulting deflection at the tip is:

$$\delta(x = 0) = \int_0^I \frac{Px^2 dx}{EI(x)}. \tag{9.72}$$

Our base model for comparison is the case of a uniform cantilever of constant thickness, h_0, and length, L_0. For this case, the second moment of the area is the constant

$$I_0 = \frac{bh_0^3}{12}, \tag{9.73}$$

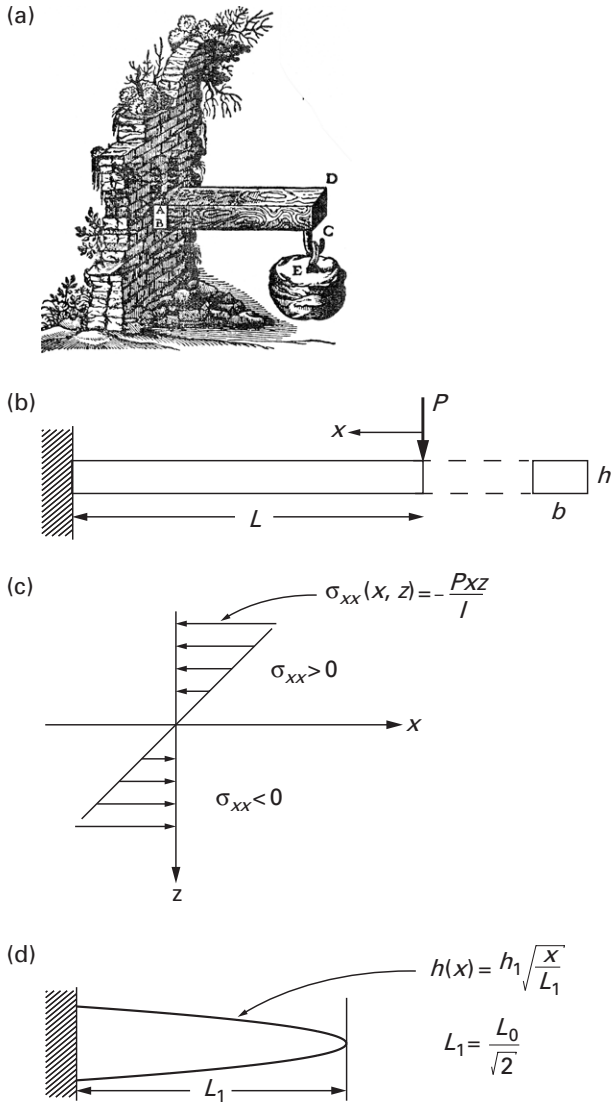


Figure 9.12 The classic cantilever beam: (a) Galileo Galilei's famous picture; (b) a modern incarnation; (c) the distribution of stress through the thickness, as indicated by eq. (9.70); and (d) the optimal shape, including a significantly shorter length, that produces a 67% reduction in the total volume (and thus the beam's weight) (after Bejan, 2000).

and its volume is

$$V_0 = bh_0L_0, \tag{9.74}$$

while its tip deflection is found by integrating eq. (9.72):

$$\delta_0 = \frac{PL_0^3}{3EI_0} = \frac{4PL_0^3}{Ebh_0^3}. \tag{9.75}$$

The maximum stress in the beam will occur at the root or support, $x = L_0$, at the beam's outer edges, $z = h_0/2$, and is determined by eq. (9.70) to be:

$$\sigma_{\max} = \sigma_{xx}(L_0, h_0/2) = \frac{6PL_0}{bh_0^2}. \tag{9.76}$$

Therefore, if a maximum stress that must not be exceeded is given as a design constraint, it follows that the minimum thickness required for a uniform beam to support a load, P, is:

$$h_0 = \sqrt{\frac{6PL_0}{b\sigma_{\max}}}. \tag{9.77}$$

Finally, for the base case, and in view of eqs. (9.75) and (9.76), we can rewrite the volume (9.74) in a form that is independent of the beam's geometrical parameters:

$$V_0 = 9\frac{EP\delta_0}{\sigma_{\max}^2}. \tag{9.78}$$

Consider now a second case where the cantilever has variable thickness, $h(x)$, and length, L_1. The variation of the thickness is determined by the requirement that, again, a given maximum stress not be exceeded. In this instance, eq. (9.70) states that (see Problem 9.42):

$$h(x) = \sqrt{\frac{6PL_1}{b\sigma_{\max}}}\sqrt{\frac{x}{L_1}} \equiv h_1\sqrt{\frac{x}{L_1}}, \tag{9.79}$$

where h_1 is the maximum thickness of a parabolic profile that begins at the tip ($x = 0$) and reaches its maximum at the support ($x = L_1$).

The volume for this beam of varying thickness is

$$V_1 = \int_0^{L_1} bh(x)\,dx = \frac{2}{3}bh_1L_1. \tag{9.80}$$

The corresponding tip deflection is found by substituting eq. (9.79) into eq. (9.72) and then performing the indicated integration:

$$\delta_1 = \frac{8PL_1^3}{Ebh_1^3}. \tag{9.81}$$

Once again we can cast the volume (9.80) of the beam with parabolically-varying thickness in a form independent of the beam's geometrical parameters, now in view of eqs. (9.79) and (9.81):

$$V_1 = 3\frac{EP\delta_1}{\sigma_{\max}^2}. \tag{9.82}$$

How are we to compare these two cases? Which is the better beam? What do we mean by the "better beam"? There are (at least) two bases for comparison. In the first, we ask: How do the volumes compare if we require each beam to have the same tip deflection while supporting the same load, P? This question is easily answered by comparing the volumes given by eqs. (9.78) and (9.82):

$$\frac{V_1}{V_0} = \frac{1}{3}\frac{\delta_1}{\delta_0} = \frac{1}{3}. \tag{9.83}$$

Thus, we have the astounding result that we can reduce the volume by 2/3 or 67%! An amazing improvement. By equating the formulas (9.77) and (9.81) for the respective deflections, and by rewriting the volumes in terms of their geometries, we can find that (see Problem 9.43):

$$\frac{h_1}{h_0} = \frac{L_1}{L_0} = \frac{1}{\sqrt{2}}. \tag{9.84}$$

Our volume savings come at a price of a beam that is not only thinner, but almost 30% shorter. This length shortening may or may not matter; it depends on the context in which this beam will be used.

Suppose we looked for a different "better beam." Suppose we require that the beams carry the same load, P, have the same maximum thickness, $h_0 = h_1$, and the same length, $L_0 = L_1 = L$. It is then easy enough to show that (see Problems 9.44 and 9.45):

$$\frac{V_1}{V_0} = \frac{2}{3} \tag{9.85}$$

and

$$\frac{\delta_1}{\delta_0} = 2. \tag{9.86}$$

In this case we still have a substantial volume reduction of 33% , but at the price of doubling the deflection. We have maintained the original length (and maximum thickness), but we now pay a different price for a different saving.

There are other ways to improve the behavior of a beam. While we have focused on the more visible external structure of the beam (i.e., its thickness and its length), we could also change the inner structure. For example, we might consider the volume (and material) saving that results

from taking a beam of a solid circular cross-section, and then hollowing it out to make it a tube (see Problems 9.57 and 9.58). Further, we might also combine changes in the internal and external structure to see what costs are reduced (see Problems 9.59 and 9.60). In fact, we can see such examples in nature any time we choose to look. Tree trunks are thicker at their bases, and branches thicker at their initial branching points, thus exemplifying external structuring. And the internal structuring of tubes shows up in bamboo and various reeds. Thus, nature seems to be paying attention to the search for optimal behavior.

Problem 9.42. Show that a beam with varying thickness, $h(x)$, that is required to have a maximum stress $\leq \sigma_{\max}$, will have the thickness distribution given in eq. (9.79). (*Hint*: Where does the maximum stress occur for varying $h(x)$?)

Problem 9.43. Show that eq. (9.84) is correct for a comparison of beams required to have the same tip deflection, δ, when under the same load, P.

Problem 9.44. Confirm that the volume ratio (9.85) is correct when beams whose lengths and maximum thickness are required to support the same load, P.

Problem 9.45. Confirm that the tip deflection ratio (9.86) is correct when beams whose lengths and maximum thickness are required to support the same load, P.

9.6 Summary

This final chapter has been devoted to optimization, the search for the optimum or best outcome to a problem. We have briefly reviewed several well-founded techniques, including calculus, linear programming (LP), and geometric programming (GP). We also talked about making the best decision when voting for candidates and choosing among alternatives. Our emphasis throughout has been less on the intricacies of the particular techniques, and more on framing the question. In this context, it is particularly important to recognize that any search for an optimum solution is to some extent "biased" or influenced by the way the question is framed. This was most evident in the discussion of voting and the expression of preferences, but it is also the case in the more "rigorous" calculus- and programming-based approaches. When we ask which is the cheapest design or product,

we are choosing money as our metric, not the design's esthetics or the product's effects on the environment. To be sure, such externals can be taken into account, but the means for so doing are neither rigorous nor entirely objective.

Having said that, we also note that we have only scratched the surface of tools that support the making of optimal decisions. For example, while linear programming is a very valuable tool and an important part of operations research, there are many other optimization techniques, including nonlinear, integer, dynamic, and geometric programming. Operations research also includes queueing theory, game theory, and simulation (particularly Monte Carlo simulation). These approaches are concerned with such issues as assessing the costs of having too few or too many service lines at a service facility, rationalizing economic and strategic decisions in the face of uncertainty, and performing simulations of problems that are analytically intractable or experimentally too expensive. There is also a vast body of literature on and experience with what might be called "continuous optimization" techniques, and their digital implementations are often used with finite element methods (FEM) and other numerical programs to seek the best designs of large complex designs, such as aircraft. All in all, the foregoing discussion is only an appetizer; more than a few full meals remain.

9.7 References

J. D. Anderson, Jr., *A History of Aerodynamics and Its Impact on Flying Machines*, Cambridge University Press, New York, 1997.

J. S. Arora, *Introduction to Optimum Design*, McGraw-Hill, New York, 1989.

K. J. Arrow, *Social Choice and Individual Values.* 1st Edition, John Wiley, New York, 1951.

T. Au and E. T. Stelson, *Introduction to Systems Engineering*, Addison-Wesley, Reading, MA, 1969.

A. Bejan, *Shape and Structure: From Engineering to Nature*, Cambridge University Press, New York, 2000.

A. S. Belegundu and T. R. Chandrupatla, *Optimization Concepts and Applications in Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1999.

W. D. Callister, Jr., *Materials Science and Engineering: An Introduction*, 4th Edition, John Wiley & Sons, New York, 1997.

A.-M. Chung, *Linear Programming*, Merrill, New York, 1963.

R. J. Duffin, E. L. Peterson, and C. Zener, *Geometric Programming*, John Wiley & Sons, New York, 1967.

C. L. Dym and E. S. Ivey, *Principles of Mathematical Modeling*, 1st Edition, Academic Press, New York, 1980.

C. L. Dym and P. Little, *Engineering Design: A Project-Based Introduction*, 2nd Edition, John Wiley & Sons, New York, 2003.

C. L. Dym, W. H. Wood, and M. J. Scott, "Rank ordering engineering designs: Pairwise comparison charts and Borda counts," *Research in Engineering Design*, **13**, 236–242, 2003.

R. L. Fox, *Optimization Methods for Engineering Design*, Addison-Wesley, Reading, MA, 1971.

R. J. Giglio and R. Wrightington, "Methods for Apportioning Costs Among Participants in Regional Systems," *Water Resources Research*, **8**(5), 1133–1144, 1972.

F. R. Giordano, M. D. Weir, and W. P. Fox, *A First Course in Mathematical Modeling*, 2nd Edition, Brooks/Cole Publishing, Pacific Grove, CA, 1997.

A. M. Lee, *Applied Queueing Theory*, St. Martin's Press, New York, 1966.

J. C. Liebman, J. W. Male, and M. Wathne, "Minimum Cost in Residential Refuse Vehicle Routes," *Journal of the Environmental Engineering Division*, Proceedings of The American Society of Civil Engineers, **101**(EE3), 399–412, June 1975.

D. P. Maki and M. Thompson, *Mathematical Models and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

J. M. McMasters, "Geometric Programming and Its Applications to Aerodynamic Optimization Problems," unpublished manuscript, courtesy of the author, undated.

H. J. Miser, "Introducing Operational Research," *Operations Research Quarterly*, **27** (3), 655–670, 1976.

P. Y. Papalambros and D. J. Wilde, *Principles of Optimal Design: Modeling and Computation*, Cambridge University Press, Cambridge, UK, 1988.

D. A. Porter and K. E. Easterling, *Phase Transformations in Metals and Alloys*, 2nd Edition, Chapman & Hall, London, 1992.

D. G. Saari, "Bad Decisions: Experimental Error or Faulty Decision Procedures," unpublished manuscript, courtesy of the author, 2001.

D. G. Saari, *Chaotic Elections!: A Mathematician Looks at Voting*, American Mathematical Society, Providence, RI, 2001.

D. G. Saari, *Decisions and Elections: Explaining the Unexpected*, Cambridge University Press, New York, 2001.

D. G. Saari, *Basic Geometry of Voting*, Springer-Verlag, New York, 1995.

M. J. Scott, "On Rank Reversals in the Borda Count," *Proceedings of DETC 2003*, American Society of Mechanical Engineering, Chicago, IL, September 2003.

M. J. Scott and E. K. Antonsson, "Arrow's theorem and engineering decision making," *Research in Engineering Design*, **11**, 218–228, 1999.

H. A. Simon, *The Sciences of the Artificial*, 3rd Edition, MIT Press, Cambridge, MA, 1996.
J. Singh, *Great Ideas of Operations Research*, Dover Publications, New York, 1968.
R. M. Stark and R. L. Nicholls, *Mathematical Foundations for Design*, McGraw-Hill, New York, 1972.
H. Theil, J. C. G. Boot, and T. Kloek, *Operations Research and Quantitative Economics*, McGraw-Hill, New York, 1965.
C. Toregas and C. ReVelle, "Binary Logic Solutions to a Class of Location Problems," *Geographical Analysis*, **5**(7), 145–155, 1973.
H. M. Wagner, *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
C. Zener, *Engineering Design by Geometric Programming*, John Wiley & Sons, New York, 1971.

9.8 Problems

9.46. (a) Find the extreme values of the function $y = \sin x$ for $0 \leq x \leq \pi$. Are the extreme values maxima or minima?
(b) What are maxima and minima of y=sin x in the interval $0 \leq x \leq 2\pi$. Are the extreme values maxima or minima?

9.47. (a) What are the extreme values of the function $y = x$ in the interval $0 \leq x \leq 2\pi$?
(b) What are the extreme values of the function $y = x - x^3/3!$ in the interval $0 \leq x \leq 2\pi$?
(c) How do the answers to parts (a) and (b) of this question relate to the answers to Problem 9.46?

9.48. A string of length, l, can be used to outline many simple geometrical figures, such as an equilateral triangle with sides $l/3$, a square with sides $l/4$, a pentagon with sides $l/5$, and a circle of circumference, l. For the figures mentioned:
(a) calculate their areas and show how they vary with the number of sides; and
(b) guess (and explain!) the maximum area that can be enclosed by a string of given length l.

9.49. Determine the maximum area of a triangle that can be inscribed in the shown semicircle of diameter, d. (*Hint*: Show that the area of the triangle is $bc/2$ and that the height, c, can be expressed [and eliminated] through a relationship between the triangle's sides and the semicircle's diameter.)

9.50. Graphically solve the following linear programming problem cast in terms of two nonnegative variables, x_1 and x_2.

$$\begin{aligned} &\text{Maximize} && z = 5x_1 + 3x_2 \\ &\text{subject to} && \begin{cases} 3x_1 + 5x_2 \leq 15 \\ 5x_1 + 2x_2 \leq 10 \end{cases} \end{aligned}$$

9.51. Graphically solve the following linear programming problem cast in terms of two nonnegative variables, x_1 and x_2.

$$\begin{aligned} &\text{Maximize} && z = 2x_1 + x_2 \\ &\text{subject to} && \begin{cases} 4x_1 + 3x_2 \leq 24 \\ 3x_1 + 5x_2 \leq 15 \end{cases} \end{aligned}$$

9.52. Graphically solve the following linear programming problem cast in terms of two nonnegative variables, x_1 and x_2.

$$\begin{aligned} &\text{Maximize} && z = 2x_1 + x_2 \\ &\text{subject to} && \begin{cases} x_1 + x_2 \leq 4 \\ 3x_1 + x_2 \leq 10 \end{cases} \end{aligned}$$

9.53. A manufacturing company regularly produces three products that are sold at unit prices of, respectively, \$6, \$11, and \$22. These prices seem to be independent of the firm's output, that is, the market seems able to absorb any amount of product without any adverse effect on their price. Four input factors are needed to make these three products, with the specific amounts, costs, and available supplies shown in the table below. Assuming that no other restrictions are placed on the company's manufacturing decisions, how much of each product should be made to maximize the company's profits? (*Hint*: Formulate the problem as a linear programming problem.)

		Product			
Input	Unit cost (\$)	1	2	3	Supply of input
1	2.0	0	1	2	150
2	1.0	1	2	1	200
3	0.5	4	6	10	400
4	2.0	0	0	2	100

9.54. A vendor customarily produces three washers whose materials and *unit* (per washer) costs are, respectively, brass at \$0.60, steel at \$1.20, and aluminum at \$1.00. The washers are sold in two collections of mixed types, as shown in the table below, as is the supply of raw materials. Mixture *A* sells at \$1.50/lb and mixture *B* sells at \$1.80/lb. The vendor would like to know, how much of each mixture should she make? (*Hint*: Formulate the problem as a linear programming problem.)

Mixture	Brass	Steel	Aluminum
A	0.25	0.50	0.25
B	0.00	0.50	0.50
Supply (lb)	1,000	400	400

9.55. A trucking firm has received an order to move 3000 tons of miscellaneous goods. The firm has fleets of 150 15-ton trucks and 100 10-ton trucks whose operating costs per ton are, respectively, \$30.00 and \$40.00. The firm also has a policy of retaining in reserve at least one 150-ton truck with every two 10-ton trucks. How many of each fleet should be dispatched to move the goods at minimal operating cost? (*Hint*: Formulate the problem as a linear programming problem.)

9.56. Ingredients *A* and *B* are mixed in varying proportions to make massage oil and machine oil, each of which is sold at the (same) wholesale price of \$3.00 per quart. The cost of massage oil is \$1.50 per quart, while machine oil costs \$2.00 per quart. While there is no fixed formula or algorithm for mixing *A* and *B* to obtain a specific type of oil, two rules are generally followed: (1) Massage oil may contain no less than 25% of *A* and no less than 50% of *B*; and (2) machine oil may contain no more than 75% of *A*. If 30 quarts of *A* and 20 quarts of *B* are available for mixing, how much of each oil should be made to maximize profit? (*Hint*: Formulate the problem as a linear programming problem.)

9.57. Determine the volume and tip deflection of a tip-loaded cantilever beam of length, *L*, and circular cross-section of constant radius, *R*.

9.58. What savings of volume (or material) could be made for the beam of Problem 9.57 if the beam cross-section were a hollow tube of constant mean radius, *R*, and tube wall thickness, *t*?

9.59. What savings of volume (or material) if a beam of constant rectangular cross-section $b \times h$ were replaced with a beam of the same length

whose cross-section is an idealized I-beam that has two, symmetrically placed small rectangles of thickness $t < h$ and area $A = b \times t$ that are separated by the beam's height, h? (*Hint*: Remember that I is the second moment of the area, $I = \int_A z^2 dA$.)

9.60. What savings would be made if the radius of the solid circular beam varied along the axis, $R = R(x)$, and was restricted to have the same deflection under the same load?

9.61. What savings would be made if (only) the radius of the tubular circular beam varied along the axis, $R = R(x)$, and was restricted to have the same deflection under the same load?

9.62. Develop a model for the *glide angle* γ of a glider. (*Hints*: Reconsider the small plane model developed in Section 9.5.2 in the absence of thrust. How does the climb angle, α, relate to the glide angle, γ?)

9.63. In the light of Problem 9.62, what is the optimum glide angle for a glider?

9.64. Show that the power $P = D \times V$ of a small, propeller-driven plane for equilibrium flight, during which the plane's acceleration is zero, can be modeled as:

$$P = C'_{01} V^3 + C'_{02} V^{-1}.$$

How do the constants, C'_{01} and C'_{02}, relate to those given for the small plane model of Section 9.5.2?

9.65. Determine the optimum speed that minimizes the power consumption for the plane model developed in Problem 9.64 using the standard calculus approach.

9.66. Determine the optimum speed that minimizes the power consumption for the plane model developed in Problem 9.64 using geometric programming (GP). Does this answer agree with that obtained in Problem 9.65?

Index