Tenth Edition

ERWIN KREYSZIG

ADVANCED ENGINEERING MATHEMATICS

Systems of Units. Some Important Conversion Factors

The most important systems of units are shown in the table below. The mks system is also known as the *International System of Units* (abbreviated *SI*), and the abbreviations sec (instead of s), gm (instead of g), and nt (instead of N) are also used.

System of units	Length	Mass	Time	Force
cgs system	centimeter (cm)	gram (g)	second (s)	dyne
mks system	meter (m)	kilogram (kg)	second (s)	newton (nt)
Engineering system	foot (ft)	slug	second (s)	pound (lb)

1 inch (in.) = 2.540000 cm

1 foot (ft) = 12 in. = 30.480000 cm

1 yard (yd) = 3 ft = 91.440000 cm

1 statute mile (mi) = 5280 ft = 1.609344 km

1 nautical mile = 6080 ft = 1.853184 km

1 acre = 4840 yd^2 = 4046.8564 m^2

1 mi^2 = 640 acres = 2.5899881 km^2

1 fluid ounce = 1/128 U.S. gallon = 231/128 in.^3 = 29.573730 cm^3

1 U.S. gallon = 4 quarts (liq) = 8 pints (liq) = 128 fl oz = 3785.4118 cm^3

1 British Imperial and Canadian gallon = 1.200949 U.S. gallons = 4546.087 cm^3

1 slug = 14.59390 kg

1 pound (lb) = 4.448444 nt

1 newton (nt) = 10^5 dynes

1 British thermal unit (Btu) = 1054.35 joules

1 joule = 10^7 ergs

1 calorie (cal) = 4.1840 joules

1 kilowatt-hour (kWh) = 3414.4 Btu = $3.6 \cdot 10^6$ joules

1 horsepower (hp) = 2542.48 Btu/h = 178.298 cal/sec = 0.74570 kW

1 kilowatt (kW) = 1000 watts = 3414.43 Btu/h = 238.662 cal/s

$°\text{F} = °\text{C} \cdot 1.8 + 32$

$1° = 60' = 3600'' = 0.017453293$ radian

For further details see, for example, D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics.* 9th ed., Hoboken, N. J: Wiley, 2011. See also AN American National Standard, ASTM/IEEE Standard Metric Practice, Institute of Electrical and Electronics Engineers, Inc. (IEEE), 445 Hoes Lane, Piscataway, N. J. 08854, website at www.ieee.org.

Differentiation

$(cu)' = cu' \qquad (c \text{ constant})$

$(u + v)' = u' + v'$

$(uv)' = u'v + uv'$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$$

$$\frac{du}{dx} = \frac{du}{dy} \cdot \frac{dy}{dx} \qquad \text{(Chain rule)}$$

$(x^n)' = nx^{n-1}$

$(e^x)' = e^x$

$(e^{ax})' = ae^{ax}$

$(a^x)' = a^x \ln a$

$(\sin x)' = \cos x$

$(\cos x)' = -\sin x$

$(\tan x)' = \sec^2 x$

$(\cot x)' = -\csc^2 x$

$(\sinh x)' = \cosh x$

$(\cosh x)' = \sinh x$

$$(\ln x)' = \frac{1}{x}$$

$$(\log_a x)' = \frac{\log_a e}{x}$$

$$(\arcsin x)' = \frac{1}{\sqrt{1 - x^2}}$$

$$(\arccos x)' = -\frac{1}{\sqrt{1 - x^2}}$$

$$(\arctan x)' = \frac{1}{1 + x^2}$$

$$(\text{arccot } x)' = -\frac{1}{1 + x^2}$$

Integration

$$\int uv'\,dx = uv - \int u'v\,dx \text{ (by parts)}$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + c \qquad (n \neq -1)$$

$$\int \frac{1}{x}\,dx = \ln|x| + c$$

$$\int e^{ax}\,dx = \frac{1}{a}e^{ax} + c$$

$$\int \sin x\,dx = -\cos x + c$$

$$\int \cos x\,dx = \sin x + c$$

$$\int \tan x\,dx = -\ln|\cos x| + c$$

$$\int \cot x\,dx = \ln|\sin x| + c$$

$$\int \sec x\,dx = \ln|\sec x + \tan x| + c$$

$$\int \csc x\,dx = \ln|\csc x - \cot x| + c$$

$$\int \frac{dx}{x^2 + a^2} = \frac{1}{a}\arctan\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \arcsin\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{x^2 + a^2}} = \text{arcsinh}\,\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{x^2 - a^2}} = \text{arccosh}\,\frac{x}{a} + c$$

$$\int \sin^2 x\,dx = \tfrac{1}{2}x - \tfrac{1}{4}\sin 2x + c$$

$$\int \cos^2 x\,dx = \tfrac{1}{2}x + \tfrac{1}{4}\sin 2x + c$$

$$\int \tan^2 x\,dx = \tan x - x + c$$

$$\int \cot^2 x\,dx = -\cot x - x + c$$

$$\int \ln x\,dx = x\ln x - x + c$$

$$\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\sin bx - b\cos bx) + c$$

$$\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\cos bx + b\sin bx) + c$$

PART E

Numeric Analysis

Numeric analysis or briefly **numerics** continues to be one of the fastest growing areas of engineering mathematics. This is a natural trend with the ever greater availability of computing power and global Internet use. Indeed, good software implementation of numerical methods are readily available. Take a look at the ***updated*** list of ***Software*** starting on p. 788. It contains software for purchase (commercial software) and software for free download (public-domain software). For convenience, we provide Internet addresses and phone numbers. The software list includes computer algebra systems (CASs), such as ***Maple*** and ***Mathematica***, along with the *Maple Computer Guide*, 10th ed., and *Mathematica Computer Guide*, 10th ed., by E. Kreyszig and E. J. Norminton related to this text that teach you stepwise how to use these computer algebra systems and with complete engineering examples drawn from the text. Furthermore, there is scientific software, such as ***IMSL***, ***LAPACK*** (free download), and scientific calculators with graphic capabilities such as ***TI-Nspire***. Note that, although we have listed frequently used quality software, this list is by no means complete.

In your career as an engineer, appplied mathematician, or scientist you are likely to use commercially available software or proprietary software, owned by the company you work for, that uses numeric methods to solve engineering problems, such as modeling chemical or biological processes, planning ecologically sound heating systems, or computing trajectories of spacecraft or satellites. For example, one of the collaborators of this book (Herbert Kreyszig) used proprietary software to determine the value of bonds, which amounted to solving higher degree polynomial equations, using numeric methods discussed in Sec. 19.2.

However, the availability of quality software does not alleviate your effort and responsibility to first ***understand*** *these numerical methods.* Your effort will pay off because, with your mathematical expertise in numerics, you will be able to plan your solution approach, judiciously select and use the appropriate software, judge the quality of software, and, perhaps, even write your own numerics software.

Numerics extends your ability to solve problems that are either difficult or impossible to solve analytically. For example, certain integrals such as error function [see App. 3, formula (35)] or large eigenvalue problems that generate high-degree characteristic polynomials cannot be solved analytically. Numerics is also used to construct approximating polynomials through data points that were obtained from some experiments.

Part E is designed to give you a solid background in numerics. We present many numeric methods as **algorithms**, which give these methods in detailed steps suitable for software implementation on your computer, CAS, or programmable calculator. The first chapter, Chap. 19, covers three main areas. These are general numerics (floating point, rounding errors, etc.), solving equations of the form $f(x) = 0$ (using Newton's method and other methods), interpolation along with methods of numeric integration that make use of it, and differentiation.

Chapter 20 covers the essentials of numeric linear algebra. The chapter breaks into two parts: solving linear systems of equations by methods of Gauss, Doolittle, Cholesky, etc. and solving eigenvalue problems numerically. Chapter 21 again has two themes: solving ordinary differential equations and systems of ordinary differential equations as well as solving partial differential equations.

Numerics is a very active area of research as new methods are invented, existing methods improved and adapted, and old methods—impractical in precomputer times—are rediscovered. A main goal in these activities is the development of well-structured software. And in large-scale work—millions of equations or steps of iterations—even small algorithmic improvements may have a large significant effect on computing time, storage demand, accuracy, and stability.

Remark on Software Use. Part E is designed in such a way as to *allow compelete flexibility on the use of CASs, software, or graphing calculators.* The computational requirements range from very little use to heavy use. The choice of computer use is at the discretion of the professor. The material and problem sets (except where clearly indicated such as in CAS Projects, CAS Problems, or CAS Experiments, which can be omitted without loss of continuity) do not require the use of a CAS or software. A scientific calculator perhaps with graphing capabilities is all that is required.

Software

See also http://www.wiley.com/college/kreyszig/

The following list will help you if you wish to find software. You may also obtain information on known and new software from websites such as Dr. Dobb's Portal, from articles published by the *American Mathematical Society* (see also its website at www.ams.org), the *Society for Industrial and Applied Mathematics* (SIAM, at www.siam.org), the *Association for Computing Machinery* (ACM, at www.acm.org), or the *Institute of Electrical and Electronics Engineers* (IEEE, at www.ieee.org). Consult also your library, computer science department, or mathematics department.

TI-Nspire. Includes TI-Nspire CAS and programmable graphic calculators. Texas Instruments, Inc., Dallas, TX. Telephone: 1-800-842-2737 or (972) 917-8324; website at www.education.ti.com.

EISPACK. See LAPACK.

GAMS (Guide to Available Mathematical Software). Website at http://gams.nist.gov. Online cross-index of software development by NIST.

IMSL (International Mathematical and Statistical Library). Visual Numerics, Inc., Houston, TX. Telephone: 1-800-222-4675 or (713) 784-3131; website at www.vni.com. Mathematical and statistical FORTRAN routines with graphics.

LAPACK. FORTRAN 77 routines for linear algebra. This software package supersedes LINPACK and EISPACK. You can download the routines from www.netlib.org/lapack. The LAPACK User's Guide is available at www.netlib.org.

LINPACK see LAPACK

Maple. Waterloo Maple, Inc., Waterloo, ON, Canada. Telephone: 1-800-267-6583 or (519) 747-2373; website at www.maplesoft.com.

Maple Computer Guide. For Advanced Engineering Mathematics, 10th edition. By E. Kreyszig and E. J. Norminton. John Wiley and Sons, Inc., Hoboken, NJ. Telephone: 1-800-225-5945 or (201) 748-6000.

Mathcad. Parametric Technology Corp. (PTC), Needham, MA. Website at www.ptc.com.

Mathematica. Wolfram Research, Inc., Champaign, IL. Telephone: 1-800-965-3726 or (217) 398-0700; website at www.wolfram.com.

Mathematica Computer Guide. For Advanced Engineering Mathematics, 10th edition. By E. Kreyszig and E. J. Norminton. John Wiley and Sons, Inc., Hoboken, NJ. Telephone: 1-800-225-5945 or (201) 748-6000.

Matlab. The MathWorks, Inc., Natick, MA. Telephone: (508) 647-7000; website at www.mathworks.com.

NAG. Numerical Algorithms Group, Inc., Lisle, IL. Telephone: (630) 971-2337; website at www.nag.com. Numeric routines in FORTRAN 77, FORTRAN 90, and C.

NETLIB. Extensive library of public-domain software. See at www.netlib.org.

NIST. National Institute of Standards and Technology, Gaithersburg, MD. Telephone: (301) 975-6478; website at www.nist.gov. For Mathematical and Computational Science Division telephone: (301) 975-3800. See also http://math.nist.gov.

Numerical Recipes. Cambridge University Press, New York, NY. Telephone: 1-800-221-4512 or (212) 924-3900; website at www.cambridge.org/us. Book, 3rd ed. (in C++) see App. 1, Ref. [E25]; source code on CD ROM in C++, which also contains old source code (but not text) for (out of print) 2nd ed. C, FORTRAN 77, FORTRAN 90 as well as source code for (out of print) 1st ed. To order, call office at West Nyack, NY, at 1-800-872-7423 or (845) 353-7500 or online at www.nr.com.

FURTHER SOFTWARE IN STATISTICS. See Part G.

CHAPTER 19

Numerics in General

Numeric analysis or briefly numerics has a distinct flavor that is different from basic calculus, from solving ODEs algebraically, or from other (nonnumeric) areas. Whereas in calculus and in ODEs there were very few choices on how to solve the problem and your answer was an algebraic answer, in numerics you have many more choices and your answers are given as tables of values (numbers) or graphs. You have to make judicous choices as to what numeric method or algorithm you want to use, how accurate you need your result to be, with what value (starting value) do you want to begin your computation, and others. This chapter is designed to provide a good transition from the algebraic type of mathematics to the numeric type of mathematics.

We begin with the general concepts such as floating point, roundoff errors, and general numeric errors and their propagation. This is followed in Sec. 19.2 by the important topic of solving equations of the type $f(x) = 0$ by various numeric methods, including the famous Newton method. Section 19.3 introduces interpolation methods. These are methods that construct new (unknown) function values from known function values. The knowledge gained in Sec. 19.3 is applied to spline interpolation (Sec. 19.4) and is useful for understanding numeric integration and differentiation covered in the last section.

Numerics provides an invaluable extension to the knowledge base of the problem-solving engineer. Many problems have no solution formula (think of a complicated integral or a polynomial of high degree or the interpolation of values obtained by measurements). In other cases a complicated solution formula may exist but may be practically useless. It is for these kinds of problems that a numerical method may generate a good answer. Thus, it is very important that the applied mathematician, engineer, physicist, or scientist becomes familiar with the essentials of numerics and its ideas, such as estimation of errors, order of convergence, numerical methods expressed in algorithms, and is also informed about the important numeric methods.

Prerequisite: Elementary calculus.
References and Answers to Problems: App. 1 Part E, App. 2.

19.1 Introduction

As an engineer or physicist you may deal with problems in elasticity and need to solve an equation such as $x \cosh x = 1$ or a more difficult problem of finding the roots of a higher order polynomial. Or you encounter an integral such as

$$\int_0^1 \exp(-x^2)\, dx$$

[see App. 3, formula (35)] that you cannot solve by elementary calculus. Such problems, which are difficult or impossible to solve algebraically, arise frequently in applications. They call for **numeric methods**, that is, systematic methods that are suitable for solving, numerically, the problems on computers or calculators. Such solutions result in tables of numbers, graphical representation (figures), or both. Typical numeric methods are iterative in nature and, for a well-choosen problem and a good starting value, will frequently converge to a desired answer. The evolution from a given problem that you observed in an experimental lab or in an industrial setting (in engineering, physics, biology, chemistry, economics, etc.) to an approximation suitable for numerics to a final answer usually requires the following steps.

1. **Modeling.** We set up a mathematical model of our problem, such as an integral, a system of equations, or a differential equation.
2. **Choosing a numeric method** and parameters (e.g., step size), perhaps with a preliminary error estimation.
3. **Programming.** We use the algorithm to write a corresponding program in a CAS, such as Maple, Mathematica, Matlab, or Mathcad, or, say, in Java, C or C^{++}, or FORTRAN, selecting suitable routines from a software system as needed.
4. **Doing the computation.**
5. **Interpreting the results** in physical or other terms, also deciding to rerun if further results are needed.

Steps 1 and 2 are related. A slight change of the model may often admit of a more efficient method. To choose methods, we must first get to know them. Chapters 19–21 contain efficient algorithms for the most important classes of problems occurring frequently in practice.

In Step 3 the program consists of the given data and a sequence of instructions to be executed by the computer in a certain order for producing the answer in numeric or graphic form.

To create a good understanding of the nature of numeric work, we continue in this section with some simple general remarks.

Floating-Point Form of Numbers

We know that in decimal notation, every real number is represented by a finite or an infinite sequence of decimal digits. Now most computers have two ways of representing numbers, called *fixed point* and *floating point*. In a **fixed-point** system all numbers are given with a fixed number of decimals after the decimal point; for example, numbers given with 3 decimals are 62.358, 0.014, 1.000. In a text we would write, say, 3 decimals as 3D. Fixed-point representations are impractical in most scientific computations because of their limited range (explain!) and will not concern us.

In a **floating-point** system we write, for instance,

$$0.6247 \cdot 10^3, \qquad 0.1735 \cdot 10^{-13}, \qquad -0.2000 \cdot 10^{-1}$$

or sometimes also

$$6.247 \cdot 10^2, \qquad 1.735 \cdot 10^{-14}, \qquad -2.000 \cdot 10^{-2}.$$

We see that in this system the number of significant digits is kept fixed, whereas the decimal point is "floating." Here, a **significant digit** of a number c is any given digit of c, except possibly for zeros to the left of the first nonzero digit; these zeros serve only to fix the position of the decimal point. (Thus any other zero is a significant digit of c.) For instance,

$$13600, \quad 1.3600, \quad 0.0013600$$

all have 5 significant digits. In a text we indicate, say, 5 significant digits, by 5S.

The use of exponents permits us to represent very large and very small numbers. Indeed, theoretically any nonzero number a can be written as

$$a = \pm m \cdot 10^n, \qquad 0.1 \leqq |m| < 1, \qquad n \text{ integer.} \tag{1}$$

On modern computers, which use binary (base 2) numbers, m is limited to k binary digits (e.g., $k = 8$) and n is limited (see below), giving representations (for finitely many numbers only!)

$$\bar{a} = \pm \bar{m} \cdot 2^n, \qquad \bar{m} = 0.d_1 d_2 \cdots d_k, \qquad d_1 > 0. \tag{2}$$

These numbers $\bar{a}$ are called *k-digit binary* **machine numbers**. Their fractional part m (or $\bar{m}$) is called the *mantissa.* This is not identical with "mantissa" as used for logarithms. n is called the *exponent* of $\bar{a}$.

It is important to realize that there are only finitely many machine numbers and that they become less and less "dense" with increasing a. For instance, there are as many numbers between 2 and 4 as there are between 1024 and 2048. Why?

The smallest positive machine number *eps* with $1 + \text{eps} > 1$ is called the *machine accuracy*. It is important to realize that there are no numbers in the intervals $[1, 1 + \text{eps}]$, $[2, 2 + 2 \cdot \text{eps}], \cdots, [1024, 1024 + 1024 \cdot \text{eps}], \cdots$. This means that, if the mathematical answer to a computation would be $1024 + 1024 \cdot \text{eps}/2$, the computer result will be *either* 1024 or $1024 \cdot \text{eps}$ so it is impossible to achieve greater accuracy.

Underflow and Overflow. The range of exponents that a typical computer can handle is very large. The IEEE (Institute of Electrical and Electronic Engineers) floating-point standard for **single precision** is from 2^{-126} to 2^{128} (1.175×10^{-38} to 3.403×10^{38}) and for **double precision** it is from 2^{-1022} to 2^{1024} (2.225×10^{-308} to 1.798×10^{308}).

As a minor technicality, to avoid storing a minus in the exponent, the ranges are shifted from $[-126, 128]$ by adding 126 (for double precision 1022). Note that shifted exponents of 255 and 1047 are used for some special cases such as representing infinity.

If, in a computation a number outside that range occurs, this is called **underflow** when the number is smaller and **overflow** when it is larger. In the case of underflow, the result is usually set to zero and computation continues. Overflow might cause the computer to halt. Standard codes (by IMSL, NAG, etc.) are written to avoid overflow. Error messages on overflow may then indicate programming errors (incorrect input data, etc.). From here on, we will be discussing the decimal results that we obtain from our computations.

Roundoff

An error is caused by **chopping** (= discarding all digits from some decimal on) or **rounding**. This error is called **roundoff error**, regardless of whether we chop or round. The rule for rounding off a number to k decimals is as follows. (The rule for rounding off to k significant digits is the same, with "decimal" replaced by "significant digit.")

Roundoff Rule. To round a number x to k decimals, and $5 \cdot 10^{-(k+1)}$ to x and chop the digits after the $(k + 1)$st digit.

EXAMPLE 1 **Roundoff Rule**

Round the number 1.23454621 to **(a)** 2 decimals, **(b)** 3 decimals, **(c)** 4 decimals, **(d)** 5 decimals, and **(e)** 6 decimals.

Solution. **(a)** For 2 decimals we add $5 \cdot 10^{-(k+1)} = 5 \cdot 10^{-3} = 0.005$ to the given number, that is, $1.2345621 + 0.005 = 1.23\,954621$. Then we chop off the digits "954621" after the space or equivalently $1.23954621 - 0.00954621 = 1.23$.
(b) $1.23454621 + 0.0005 = 1.235\,04621$, so that for 3 decimals we get 1.234.
(c) 1.23459621 after chopping give us 1.2345 (4 decimals).
(d) 1.23455121 yields 1.23455 (5 decimals).
(e) 1.23454671 yields 1.234546 (6 decimals).
Can you round the number to 7 decimals? ■

Chopping is not recommended because the corresponding error can be larger than that in rounding. (Nevertheless, some computers use it because it is simpler and faster. On the other hand, some computers and calculators improve accuracy of results by doing intermediate calculations using one or more extra digits, called *guarding digits.*)

Error in Rounding. Let $\bar{a} = \mathrm{fl}(a)$ in (2) be the floating-point computer approximation of a in (1) obtained by rounding, where fl suggests **floating**. Then the roundoff rule gives (by dropping exponents) $|m - \bar{m}| \leqq \frac{1}{2} \cdot 10^{-k}$. Since $|m| \geqq 0.1$, this implies (when $a \neq 0$)

$$\left|\frac{a-\bar{a}}{a}\right| \approx \left|\frac{m-\bar{m}}{m}\right| \leqq \frac{1}{2} \cdot 10^{1-k}. \tag{3}$$

The right side $u = \frac{1}{2} \cdot 10^{1-k}$ is called the **rounding unit**. If we write $\bar{a} = a(1 + \delta)$, we have by algebra $(\bar{a} - a)/a = \delta$, hence $|\delta| \leqq u$ by (3). *This shows that the rounding unit u is an error bound in rounding.*

Rounding errors may ruin a computation completely, even a small computation. In general, these errors become the more dangerous the more arithmetic operations (perhaps several millions!) we have to perform. It is therefore important to analyze computational programs for expected rounding errors and to find an arrangement of the computations such that the effect of rounding errors is as small as possible.

As mentioned, the arithmetic in a computer is not exact and causes further errors; however, these will not be relevant to our discussion.

Accuracy in Tables. Although available software has rendered various tables of function values superfluous, some tables (of higher functions, of coefficients of integration formulas, etc.) will still remain in occasional use. If a table shows k significant digits, it is conventionally assumed that any value $\tilde{a}$ in the table deviates from the exact value a by at most $\pm\frac{1}{2}$ unit of the kth digit.

Loss of Significant Digits

This means that a result of a calculation has fewer correct digits than the numbers from which it was obtained. This happens if we subtract two numbers of about the same size, for example, $0.1439 - 0.1426$ ("subtractive cancellation"). It may occur in simple problems, but it can be avoided in most cases by simple changes of the algorithm—if one is aware of it! Let us illustrate this with the following basic problem.

EXAMPLE 2 **Quadratic Equation. Loss of Significant Digits**

Find the roots of the equation

$$x^2 + 40x + 2 = 0,$$

using 4 significant digits (abbreviated 4S) in the computation.

Solution. A formula for the roots x_1, x_2 of a quadratic equation $ax^2 + bx + c = 0$ is

$$x_1 = \frac{1}{2a}\,(-b + \sqrt{b^2 - 4ac}), \qquad x_2 = \frac{1}{2a}\,(-b - \sqrt{b^2 - 4ac}). \tag{4}$$

Furthermore, since $x_1 x_2 = c/a$, another formula for those roots

$$x_1 = \frac{c}{ax_2}, \qquad x_2 \text{ as in (4)}. \tag{5}$$

We see that this avoids cancellation in x_1 for positive b.

If $b < 0$, calculate x_1 from (4) and then $x_2 = c/(ax_1)$.

For $x^2 + 40x + 2 = 0$ we obtain from (4) $x = -20 \pm \sqrt{398} = -20 \pm 19.95$, hence $x_2 = -20.00 - 19.95$, involving no difficulty, and $x_1 = -20.00 + 19.95 = -0.05$, a poor value involving loss of digits by subtractive cancellation.

In contrast, (5) gives $x_1 = 2.000/(-39.95) = -0.05006$, the absolute value of the error being less than one unit of the last digit, as a computation with more digits shows. The 10S-value is -0.05006265674. ■

Errors of Numeric Results

Final results of computations of unknown quantities generally are **approximations**; that is, they are not exact but involve errors. Such an error may result from a combination of the following effects. **Roundoff errors** result from rounding, as discussed above. **Experimental errors** are errors of given data (probably arising from measurements). **Truncating errors** result from truncating (prematurely breaking off), for instance, if we replace a Taylor series with the sum of its first few terms. These errors depend on the computational method used and must be dealt with individually for each method. [“Truncating” is sometimes used as a term for chopping off (see before), a terminology that is not recommended.]

Formulas for Errors. If $\tilde{a}$ is an approximate value of a quantity whose exact value is a, we call the difference

$$\epsilon = a - \tilde{a} \tag{6}$$

the **error** of $\tilde{a}$. Hence

$$a = \tilde{a} + \epsilon, \qquad \text{True value} = \text{Approximation} + \text{Error}. \tag{6*}$$

For instance, if $\tilde{a} = 10.5$ is an approximation of $a = 10.2$, its error is $\epsilon = -0.3$. The error of an approximation $\tilde{a} = 1.60$ of $a = 1.82$ is $\epsilon = 0.22$.

CAUTION! In the literature $|a - \tilde{a}|$ (“absolute error”) or $\tilde{a} - a$ are sometimes also used as definitions of error.

The **relative error** ϵ_r of $\tilde{a}$ is defined by

$$\epsilon_r = \frac{\epsilon}{a} = \frac{a - \tilde{a}}{a} = \frac{\text{Error}}{\text{True value}} \qquad (a \neq 0). \tag{7}$$

This looks useless because a is unknown. But if $|\epsilon|$ is much less than $|\tilde{a}|$, then we can use $\tilde{a}$ instead of a and get

$$\epsilon_r \approx \frac{\epsilon}{\tilde{a}}. \tag{7'}$$

This still looks problematic because ϵ is unknown—if it were known, we could get $a = \tilde{a} + \epsilon$ from (6) and we would be done. But what one often can obtain in practice is an **error bound** for $\tilde{a}$, that is, a number β such that

$$|\epsilon| \leqq \beta, \qquad \text{hence} \qquad |a - \tilde{a}| \leqq \beta.$$

This tells us how far away from our computed $\tilde{a}$ the unknown a can at most lie. Similarly, for the relative error, an error bound is a number β_r such that

$$|\epsilon_r| \leqq \beta_r, \qquad \text{hence} \qquad \left|\frac{a - \tilde{a}}{a}\right| \leqq \beta_r.$$

Error Propagation

This is an important matter. It refers to how errors at the beginning and in later steps (roundoff, for example) propagate into the computation and affect accuracy, sometimes very drastically. We state here what happens to error bounds. Namely, bounds for the *error* add under addition and subtraction, whereas bounds for the *relative error* add under multiplication and division. You do well to keep this in mind.

THEOREM 1

Error Propagation

(a) *In addition and subtraction, a bound for the* ***error*** *of the results is given by the sum of the error bounds for the terms.*

(b) *In multiplication and division, an error bound for the* ***relative error*** *of the results is given (approximately) by the sum of the bounds for the relative errors of the given numbers.*

PROOF **(a)** We use the notations $x = \tilde{x} + \epsilon_x$, $y = \tilde{y} + \epsilon_y$, $|\epsilon_x| \leqq \beta_x$, $|\epsilon_y| \leqq \beta_y$. Then for the error ϵ of the *difference* we obtain

$$\begin{aligned} |\epsilon| &= |x - y - (\tilde{x} - \tilde{y})| \\ &= |x - \tilde{x} - (y - \tilde{y})| \\ &= |\epsilon_x - \epsilon_y| \leqq |\epsilon_x| + |\epsilon_y| \leqq \beta_x + \beta_y. \end{aligned}$$

The proof for the *sum* is similar and is left to the student.

(b) For the relative error ϵ_r of $\tilde{x}\tilde{y}$ we get from the relative errors ϵ_{rx} and ϵ_{ry} of $\tilde{x}$, $\tilde{y}$ and bounds β_{rx}, β_{ry}

$$\begin{aligned} |\epsilon_r| &= \left|\frac{xy - \tilde{x}\tilde{y}}{xy}\right| = \left|\frac{xy - (x - \epsilon_x)(y - \epsilon_y)}{xy}\right| = \left|\frac{\epsilon_x y + \epsilon_y x - \epsilon_x \epsilon_y}{xy}\right| \\ &\approx \left|\frac{\epsilon_x y + \epsilon_y x}{xy}\right| \leqq \left|\frac{\epsilon_x}{x}\right| + \left|\frac{\epsilon_y}{y}\right| = |\epsilon_{rx}| + |\epsilon_{ry}| \leqq \beta_{rx} + \beta_{ry}. \end{aligned}$$

This proof shows what "approximately" means: we neglected $\epsilon_x\epsilon_y$ as small in absolute value compared to $|\epsilon_x|$ and $|\epsilon_y|$. The proof for the quotient is similar but slightly more tricky (see Prob. 13). ■

Basic Error Principle

Every numeric method should be accompanied by an error estimate. If such a formula is lacking, is extremely complicated, or is impractical because it involves information (for instance, on derivatives) that is not available, the following may help.

Error Estimation by Comparison. *Do a calculation twice with different accuracy. Regard the difference* $\tilde{a}_2 - \tilde{a}_1$ *of the results* $\tilde{a}_1, \tilde{a}_2$ *as a (perhaps crude) estimate of the error* ϵ_1 *of the inferior result* $\tilde{a}_1$. Indeed, $\tilde{a}_1 + \epsilon_1 = \tilde{a}_2 + \epsilon_2$ by formula (4*). This implies $\tilde{a}_2 - \tilde{a}_1 = \epsilon_1 - \epsilon_2 \approx \epsilon_1$ because $\tilde{a}_2$ is generally more accurate than $\tilde{a}_1$, so that $|\epsilon_2|$ is small compared to $|\epsilon_1|$.

Algorithm. Stability

Numeric methods can be formulated as algorithms. An **algorithm** is a step-by-step procedure that states a numeric method in a form (a "**pseudocode**") understandable to humans. (See Table 19.1 to see what an algorithm looks like.) The algorithm is then used to write a program in a programming language that the computer can understand so that it can execute the numeric method. Important algorithms follow in the next sections. For routine tasks your CAS or some other software system may contain programs that you can use or include as parts of larger programs of your own.

Stability. To be useful, an algorithm should be **stable**; that is, small changes in the initial data should cause only small changes in the final results. However, if small changes in the initial data can produce large changes in the final results, we call the algorithm **unstable**.

This "*numeric instability*," which in most cases can be avoided by choosing a better algorithm, must be distinguished from "*mathematical instability*" of a problem, which is called "*ill-conditioning*," a concept we discuss in the next section.

Some algorithms are stable only for certain initial data, so that one must be careful in such a case.

PROBLEM SET 19.1

1. **Floating point.** Write 84.175, −528.685, 0.000924138, and −362005 in floating-point form, rounded to 5S (5 significant digits).
2. Write −76.437125, 60100, and −0.00001 in floating-point form, rounded to 4S.
3. **Small differences of large numbers** may be particularly strongly affected by rounding errors. Illustrate this by computing $0.81534/(35 \cdot 724 - 35.596)$ as given with 5S, then rounding stepwise to 4S, 3S, and 2S, where "stepwise" means round the rounded numbers, not the given ones.
4. **Order of terms**, in adding with a fixed number of digits, will generally affect the sum. Give an example. Find empirically a rule for the best order.
5. **Rounding and adding.** Let $a_1, \cdots, a_n$ be numbers with a_j correctly rounded to S_j digits. In calculating the sum $a_1 + \cdots + a_n$, retaining $S = \min S_j$ significant digits, is it essential that we first add and then round the result or that we first round each number to S significant digits and then add?
6. **Nested form.** Evaluate

$$f(x) = x^3 - 7.5x^2 + 11.2x + 2.8$$
$$= ((x - 7.5)x + 11.2)x + 2.8$$

at $x = 3.94$ using 3S arithmetic and rounding, in both of the given forms. The latter, called the *nested form*, is usually preferable since it minimizes the number of operations and thus the effect of rounding.

7. **Quadratic equation.** Solve $x^2 - 30x + 1 = 0$ by (4) and by (5), using 6S in the computation. Compare and comment.

8. Solve $x^2 - 40x + 2 = 0$, using 4S-computation.

9. Do the computations in Prob. 7 with 4S and 2S.

10. **Instability.** For small $|a|$ the equation $(x - k)^2 = a$ has nearly a double root. Why do these roots show instability?

11. **Theorems on errors.** Prove Theorem 1(a) for addition.

12. **Overflow and underflow** can sometimes be avoided by simple changes in a formula. Explain this in terms of $\sqrt{x^2 + y^2} = x\sqrt{1 + (y/x)^2}$ with $x^2 \geqq y^2$ and x so large that x^2 would cause overflow. Invent examples of your own.

13. **Division.** Prove Theorem 1(b) for division.

14. **Loss of digits. Square root.** Compute $\sqrt{x^2 + 4} - 2$ with 6S arithmetic for $x = 0.001$ **(a)** as given and **(b)** from $x^2/(\sqrt{x^2 + 4} + 2)$ (derive!).

15. **Logarithm.** Compute $\ln a - \ln b$ with 6S arithmetic for $a = 4.00000$ and $b = 3.99900$ **(a)** as given and **(b)** from $\ln(a/b)$.

16. **Cosine.** Compute $1 - \cos x$ with 6S arithmetic for $x = 0.02$ **(a)** as given and **(b)** by $2 \sin^2 \frac{1}{2}x$ (derive!).

17. Discuss the numeric use of (12) in App. A3.1 for $\cos v - \cos u$ when $u \approx v$.

18. **Quotient near 0/0. (a)** Compute $(1 - \cos x)/\sin x$ with 6S arithmetic for $x = 0.005$. **(b)** Looking at Prob. 16, find a much better formula.

19. **Exponential function.** Calculate $1/e = 0.367879$ (6S) from the partial sums of 5–10 terms of the Maclaurin series **(a)** of e^{-x} with $x = 1$, **(b)** of e^x with $x = 1$ and then taking the reciprocal. Which is more accurate?

20. Compute e^{-10} with 6S arithmetic in two ways (as in Prob. 19).

21. **Binary conversion.** Show that

$$\begin{aligned}23 &= 20 \cdot 10^1 + 3 \cdot 10^0 = 16 + 4 + 2 + 1 \\ &= 2^4 + 2^2 + 2^1 + 2^0 = (1 \quad 0 \quad 1 \quad 1 \quad 1.)_2\end{aligned}$$

can be obtained by the division algorithm

$$\begin{array}{lll} 2\underline{|23} & \text{Remainder} & 1 = c_0 \\ 2\underline{|11} & & 1 = c_1 \\ 2\underline{|5\ } & & 1 = c_2 \\ 2\underline{|2\ } & & 0 = c_3 \\ \ \ 0 & & 1 = c_4 \end{array}$$

22. Convert $(0.59375)_{10}$ to $(0.10011)_2$ by successive multiplication by 2 and dropping (removing) the integer parts, which give the binary digits $c_1, c_2, \cdots$:

$$\begin{array}{rl} & 0\ .59375 \cdot 2 \\ c_1 = & \boxed{1}\ .1875 \cdot 2 \\ c_2 = & \boxed{0}\ .375 \cdot 2 \\ c_3 = & \boxed{0}\ .75 \cdot 2 \\ c_4 = & \boxed{1}\ .5 \cdot 2 \\ c_5 = & \boxed{1}\ .0 \end{array}$$

23. Show that 0.1 is not a binary machine number.

24. Prove that any binary machine number has a finite decimal representation. Is the converse true?

25. **CAS EXPERIMENT. Approximations.** Obtain $x = 0.1 = \dfrac{3}{2} \displaystyle\sum_{m=1}^{\infty} 2^{-4m}$ from Prob. 23. Which machine number (partial sum) S_n will first have the value 0.1 to 30 decimal digits?

26. **CAS EXPERIMENT. Integration from Calculus.** Integrating by parts, show that $I_n = \int_0^1 e^x x^n \, dx = e - nI_{n-1}, I_0 = e - 1$. **(a)** Compute $I_n, n = 0, \cdots,$ using 4S arithmetic, obtaining $I_8 = -3.906$. Why is this nonsense? Why is the error so large?

    (b) Experiment in (a) with the number of digits $k > 4$. As you increase k, will the first negative value $n = N$ occur earlier or later? Find an empirical formula for $N = N(k)$.

27. **Backward Recursion. In Prob. 26.** Using $e^x < e$ $(0 < x < 1)$, conclude that $|I_n| \leqq e/(n + 1) \rightarrow 0$ as $n \rightarrow \infty$. Solve the iteration formula for $I_{n-1} = (e - I_n)/n$, start from $I_{15} \approx 0$ and compute 4S values of $I_{14}, I_{13}, \cdots, I_1$.

28. **Harmonic series.** $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ diverges. Is the same true for the corresponding series of computer numbers?

29. **Approximations of π = 3.14159265358979** $\cdots$ are 22/7 and 355/113. Determine the corresponding errors and relative errors to 3 significant digits.

30. **Compute π by Machin's approximation** $16 \arctan(\frac{1}{5}) - 4 \arctan(\frac{1}{239})$ to 10S (which are correct). [In 1986, D. H. Bailey (NASA Ames Research Center, Moffett Field, CA 94035) computed almost 30 million decimals of π on a CRAY-2 in less than 30 hrs. The race for more and more decimals is continuing. See the Internet under pi.]

19.2 Solution of Equations by Iteration

For each of the remaining sections of this chapter, we select basic kinds of problems and discuss numeric methods on how to solve them. The reader will learn about a variety of important problems and become familiar with ways of thinking in numerical analysis.

Perhaps the easiest conceptual problem is to find solutions of a single equation

$$f(x) = 0, \tag{1}$$

where f is a given function. A **solution** of (1) is a number $x = s$ such that $f(s) = 0$. Here, s suggests "solution," but we shall also use other letters.

It is interesting to note that the task of solving (1) is a question made for numeric algorithms, as in general there are no direct formulas, except in a few simple cases.

Examples of single equations are $x^3 + x = 1$, $\sin x = 0.5x$, $\tan x = x$, $\cosh x = \sec x$, $\cosh x \cos x = -1$, which can all be written in the form of (1). The first of the five equations is an **algebraic equation** because the corresponding f is a polynomial. In this case the solutions are called **roots** of the equation and the solution process is called *finding roots.* The other equations are **transcendental equations** because they involve transcendental functions.

There are a very large number of applications in engineering, where we have to solve a single equation (1). You have seen such applications when solving characteristic equations in Chaps. 2, 4, and 8; partial fractions in Chap. 6; residue integration in Chap. 16, finding eigenvalues in Chap. 12, and finding zeros of Bessel functions, also in Chap. 12. Moreover, methods of finding roots are very important in areas outside of classical engineering. For example, in finance, the problem of determining how much a bond is worth amounts to solving an algebraic equation.

To solve (1) when there is no formula for the exact solution available, we can use an approximation method, such as an **iteration method**. This is a method in which we start from an initial guess x_0 (which may be poor) and compute step by step (in general better and better) approximations $x_1, x_2, \cdots$ of an unknown solution of (1). We discuss three such methods that are of particular practical importance and mention two others in the problem set.

It is very important that the reader understand these methods and their underlying ideas. The reader will then be able to select judiciously the appropriate software from among different software packages that employ variations of such methods and not just treat the software programs as "black boxes."

In general, iteration methods are easy to program because the computational operations are the same in each step—just the data change from step to step—and, more importantly, if in a concrete case a method converges, it is stable in general (see Sec. 19.1).

Fixed-Point Iteration for Solving Equations $f(x) = 0$

Note: Our present use of the word "fixed point" has absolutely nothing to do with that in the last section.

By some *algebraic steps* we transform (1) into the form

$$x = g(x). \tag{2}$$

Then we choose an x_0 and compute $x_1 = g(x_0)$, $x_2 = g(x_1)$, and in general

$$x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots). \tag{3}$$

A solution of (2) is called a **fixed point** of g, motivating the name of the method. This is a solution of (1), since from $x = g(x)$ we can return to the original form $f(x) = 0$. From (1) we may get several different forms of (2). The behavior of corresponding iterative sequences $x_0, x_1, \cdots$ may differ, in particular, with respect to their speed of convergence. Indeed, some of them may not converge at all. Let us illustrate these facts with a simple example.

EXAMPLE 1 An Iteration Process (Fixed-Point Iteration)

Set up an iteration process for the equation $f(x) = x^2 - 3x + 1 = 0$. Since we know the solutions

$$x = 1.5 \pm \sqrt{1.25}, \qquad \text{thus} \qquad 2.618034 \qquad \text{and} \qquad 0.381966,$$

we can watch the behavior of the error as the iteration proceeds.

Solution. The equation may be written

$$x = g_1(x) = \tfrac{1}{3}(x^2 + 1), \qquad \text{thus} \qquad x_{n+1} = \tfrac{1}{3}(x_n^2 + 1). \tag{4a}$$

If we choose $x_0 = 1$, we obtain the sequence (Fig. 426a; computed with 6S and then rounded)

$$x_0 = 1.000, \quad x_1 = 0.667, \quad x_2 = 0.481, \quad x_3 = 0.411, \quad x_4 = 0.390, \cdots$$

which seems to approach the smaller solution. If we choose $x_0 = 2$, the situation is similar. If we choose $x_0 = 3$, we obtain the sequence (Fig. 426a, upper part)

$$x_0 = 3.000, \quad x_1 = 3.333, \quad x_2 = 4.037, \quad x_3 = 5.766, \quad x_4 = 11.415, \cdots$$

which diverges.

Our equation may also be written (divide by x)

$$x = g_2(x) = 3 - \frac{1}{x}, \qquad \text{thus} \qquad x_{n+1} = 3 - \frac{1}{x_n}, \tag{4b}$$

and if we choose $x_0 = 1$, we obtain the sequence (Fig. 426b)

$$x_0 = 1.000, \quad x_1 = 2.000, \quad x_2 = 2.500, \quad x_3 = 2.600, \quad x_4 = 2.615, \cdots$$

which seems to approach the larger solution. Similarly, if we choose $x_0 = 3$, we obtain the sequence (Fig. 426b)

$$x_0 = 3.000, \quad x_1 = 2.667, \quad x_2 = 2.625, \quad x_3 = 2.619, \quad x_4 = 2.618, \cdots.$$

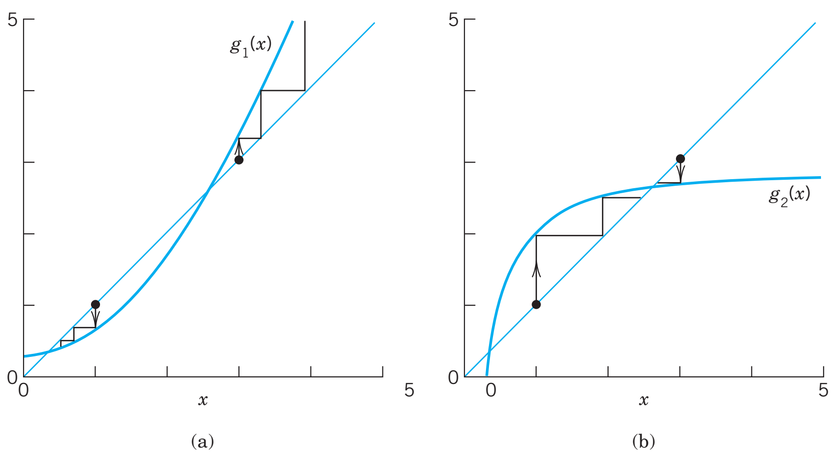


Fig. 426. Example 1, iterations (4a) and (4b)

Our figures show the following. In the lower part of Fig. 426a the slope of $g_1(x)$ is less than the slope of $y = x$, which is 1, thus $|g_1'(x)| < 1$, and we seem to have convergence. In the upper part, $g_1(x)$ is steeper ($g_1'(x) > 1$) and we have divergence. In Fig. 426b the slope of $g_2(x)$ is less near the intersection point ($x = 2.618$, fixed point of g_2, solution of $f(x) = 0$), and both sequences seem to converge. From all this we conclude that convergence seems to depend on the fact that, in a neighborhood of a solution, the curve of $g(x)$ is less steep than the straight line $y = x$, and we shall now see that this condition $|g'(x)| < 1$ (= slope of $y = x$) is sufficient for convergence. ■

An iteration process defined by (3) is called **convergent** for an x_0 if the corresponding sequence $x_0, x_1, \cdots$ is convergent.

A sufficient condition for convergence is given in the following theorem, which has various practical applications.

THEOREM 1

Convergence of Fixed-Point Iteration

Let $x = s$ be a solution of $x = g(x)$ and suppose that g has a continuous derivative in some interval J containing s. Then, if $|g'(x)| \leqq K < 1$ in J, the iteration process defined by (3) *converges for any x_0 in J. The limit of the sequence $\{x_n\}$ is s.*

PROOF By the mean value theorem of differential calculus there is a t between x and s such that

$$g(x) - g(s) = g'(t)(x - s) \qquad (x \text{ in } J).$$

Since $g(s) = s$ and $x_1 = g(x_0)$, $x_2 = g(x_1), \cdots$, we obtain from this and the condition on $|g'(x)|$ in the theorem

$$|x_n - s| = |g(x_{n-1}) - g(s)| = |g'(t)||x_{n-1} - s| \leqq K|x_{n-1} - s|.$$

Applying this inequality n times, for $n, n-1, \cdots, 1$ gives

$$|x_n - s| \leqq K|x_{n-1} - s| \leqq K^2|x_{n-2} - s| \leqq \cdots \leqq K^n|x_0 - s|.$$

Since $K < 1$, we have $K^n \to 0$; hence $|x_n - s| \to 0$ as $n \to \infty$. ■

We mention that a function g satisfying the condition in Theorem 1 is called a **contraction** because $|g(x) - g(v)| \leqq K|x - v|$, where $K < 1$. Furthermore, K gives information on the speed of convergence. For instance, if $K = 0.5$, then the accuracy increases by at least 2 digits in only 7 steps because $0.5^7 < 0.01$.

EXAMPLE 2 **An Iteration Process. Illustration of Theorem 1**

Find a solution of $f(x) = x^3 + x - 1 = 0$ by iteration.

Solution. A sketch shows that a solution lies near $x = 1$. **(a)** We may write the equation as $(x^2 + 1)x = 1$ or

$$x = g_1(x) = \frac{1}{1 + x^2}, \qquad \text{so that} \qquad x_{n+1} = \frac{1}{1 + x_n^2}. \qquad \text{Also} \qquad |g_1'(x)| = \frac{2|x|}{(1 + x^2)^2} < 1$$

for any x because $4x^2/(1 + x^2)^4 = 4x^2/(1 + 4x^2 + \cdots) < 1$, so that by Theorem 1 we have convergence for any x_0. Choosing $x_0 = 1$, we obtain (Fig. 427)

$$x_1 = 0.500, \quad x_2 = 0.800, \quad x_3 = 0.610, \quad x_4 = 0.729, \quad x_5 = 0.653, \quad x_6 = 0.701, \cdots.$$

The solution exact to 6D is $s = 0.682328$.

(b) The given equation may also be written

$$x = g_2(x) = 1 - x^3. \qquad \text{Then} \qquad |g_2'(x)| = 3x^2$$

and this is greater than 1 near the solution, so that we cannot apply Theorem 1 and assert convergence. Try $x_0 = 1, x_0 = 0.5, x_0 = 2$ and see what happens.

The example shows that the transformation of a given $f(x) = 0$ into the form $x = g(x)$ with g satisfying $|g'(x) \leqq K < 1$ may need some experimentation. ■

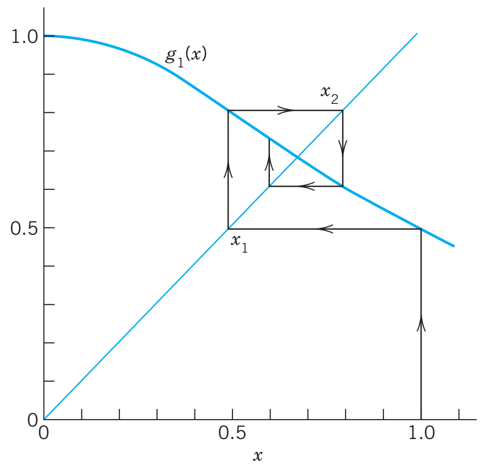


Fig. 427. Iteration in Example 2

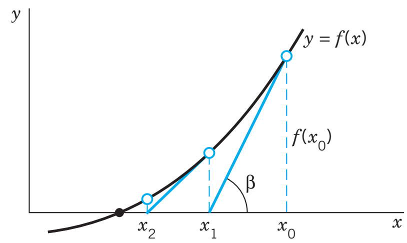


Fig. 428. Newton's method

Newton's Method for Solving Equations $f(x) = 0$

Newton's method, also known as **Newton–Raphson's method**,[1] is another iteration method for solving equations $f(x) = 0$, where f is assumed to have a continuous derivative f'. The method is commonly used because of its simplicity and great speed.

The underlying idea is that we approximate the graph of f by suitable tangents. Using an approximate value x_0 obtained from the graph of f, we let x_1 be the point of intersection of the x-axis and the tangent to the curve of f at x_0 (see Fig. 428). Then

$$\tan \beta = f'(x_0) = \frac{f(x_0)}{x_0 - x_1}, \qquad \text{hence} \qquad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

In the second step we compute $x_2 = x_1 - f(x_1)/f'(x_1)$, in the third step x_3 from x_2 again by the same formula, and so on. We thus have the algorithm shown in Table 19.1. Formula (5) in this algorithm can also be obtained if we algebraically solve Taylor's formula

$$f(x_{n+1}) \approx f(x_n) + (x_{n+1} - x_n)f'(x_n) = 0. \tag{5*}$$

[1]JOSEPH RAPHSON (1648–1715), English mathematician who published a method similar to Newton's method. For historical details, see Ref. [GenRef2], p. 203, listed in App. 1.

Table 19.1 Newton's Method for Solving Equations $f(x) = 0$

ALGORITHM NEWTON (f, f', x_0, ϵ, N)

This algorithm computes a solution of $f(x) = 0$ given an initial approximation x_0 (starting value of the iteration). Here the function $f(x)$ is continuous and has a continuous derivative $f'(x)$.

INPUT: f, f', initial approximation x_0, tolerance $\epsilon > 0$, maximum number of iterations N.

OUTPUT: Approximate solution x_n ($n \leqq N$) or message of failure.

For $n = 0, 1, 2, \cdots, N - 1$ do:

1 Compute $f'(x_n)$.

2 If $f'(x_n) = 0$ then OUTPUT "Failure." Stop.

[*Procedure completed unsuccessfully*]

3 Else compute

(5)
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

4 If $|x_{n+1} - x_n| \leqq \epsilon |x_{n+1}|$ then OUTPUT x_{n+1}. Stop.

[*Procedure completed successfully*]

End

5 OUTPUT "Failure". Stop.

[*Procedure completed unsuccessfully after N iterations*]

End NEWTON

If it happens that $f'(x_n) = 0$ for some n (see line 2 of the algorithm), then try another starting value x_0. Line 3 is the heart of Newton's method.

The inequality in line 4 is a **termination criterion**. If the sequence of the x_n converges and the criterion holds, we have reached the desired accuracy and stop. Note that this is just a form of the relative error test. It ensures that the result has the desired number of significant digits. If $|x_{n+1}| = 0$, the condition is satisfied if and only if $x_{n+1} = x_n = 0$, otherwise $|x_{n+1} - x_n|$ must be sufficiently small. The factor $|x_{n+1}|$ is needed in the case of zeros of very small (or very large) absolute value because of the high density (or of the scarcity) of machine numbers for those x.

WARNING! The criterion by itself does not imply convergence. *Example.* The harmonic series diverges, although its partial sums $x_n = \sum_{k=1}^{n} 1/k$ satisfy the criterion because $\lim (x_{n+1} - x_n) = \lim (1/(n + 1)) = 0$.

Line 5 gives another termination criterion and is needed because Newton's method may diverge or, due to a poor choice of x_0, may not reach the desired accuracy by a reasonable number of iterations. Then we may try another x_0. If $f(x) = 0$ has more than one solution, different choices of x_0 may produce different solutions. Also, an iterative sequence may sometimes converge to a solution different from the expected one.

EXAMPLE 3 Square Root

Set up a Newton iteration for computing the square root x of a given positive number c and apply it to $c = 2$.

Solution. We have $x = \sqrt{c}$, hence $f(x) = x^2 - c = 0, f'(x) = 2x$, and (5) takes the form

$$x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right).$$

For $c = 2$, choosing $x_0 = 1$, we obtain

$$x_1 = 1.500000, \qquad x_2 = 1.416667, \qquad x_3 = 1.414216, \qquad x_4 = 1.414214, \cdots.$$

x_4 is exact to 6D. ■

EXAMPLE 4 Iteration for a Transcendental Equation

Find the positive solution of $2 \sin x = x$.

Solution. Setting $f(x) = x - 2 \sin x$, we have $f'(x) = 1 - 2 \cos x$, and (5) gives

$$x_{n+1} = x_n - \frac{x_n - 2 \sin x_n}{1 - 2 \cos x_n} = \frac{2(\sin x_n - x_n \cos x_n)}{1 - 2 \cos x_n} = \frac{N_n}{D_n}.$$

n	x_n	N_n	D_n	x_{n+1}
0	2.00000	3.48318	1.83229	1.90100
1	1.90100	3.12470	1.64847	1.89552
2	1.89552	3.10500	1.63809	1.89550
3	1.89550	3.10493	1.63806	1.89549

From the graph of f we conclude that the solution is near $x_0 = 2$. We compute:
$x_4 = 1.89549$ is exact to 5D since the solution to 6D is 1.895494. ■

EXAMPLE 5 Newton's Method Applied to an Algebraic Equation

Apply Newton's method to the equation $f(x) = x^3 + x - 1 = 0$.

Solution. From (5) we have

$$x_{n+1} = x_n - \frac{x_n^3 + x_n - 1}{3x_n^2 + 1} = \frac{2x_n^3 + 1}{3x_n^2 + 1}.$$

Starting from $x_0 = 1$, we obtain

$$x_1 = 0.750000, \quad x_2 = 0.686047, \quad x_3 = 0.682340, \quad x_4 = 0.682328, \cdots$$

where x_4 has the error $-1 \cdot 10^{-6}$. A comparison with Example 2 shows that the present convergence is much more rapid. This may motivate the concept of the *order of an iteration process,* to be discussed next. ■

Order of an Iteration Method. Speed of Convergence

The quality of an iteration method may be characterized by the speed of convergence, as follows.

Let $x_{n+1} = g(x_n)$ define an iteration method, and let x_n approximate a solution s of $x = g(x)$. Then $x_n = s - \epsilon_n$, where ϵ_n is the error of x_n. Suppose that g is differentiable a number of times, so that the Taylor formula gives

$$\begin{aligned} x_{n+1} = g(x_n) &= g(s) + g'(s)(x_n - s) + \tfrac{1}{2}g''(s)(x_n - s)^2 + \cdots \\ &= g(s) - g'(s)\epsilon_n + \tfrac{1}{2}g''(s)\epsilon_n^2 + \cdots . \end{aligned} \tag{6}$$

The exponent of ϵ_n in the first nonvanishing term after $g(s)$ is called the **order** of the iteration process defined by g. The order measures the speed of convergence.

To see this, subtract $g(s) = s$ on both sides of (6). Then on the left you get $x_{n+1} - s = -\epsilon_{n+1}$, where ϵ_{n+1} is the error of x_{n+1}. And on the right the remaining expression equals approximately its first nonzero term because $|\epsilon_n|$ is small in the case of convergence. Thus

$$\begin{array}{lll} \text{(a)} & \epsilon_{n+1} \approx +g'(s)\epsilon_n & \text{in the case of first order,} \\ \text{(b)} & \epsilon_{n+1} \approx -\tfrac{1}{2}g''(s)\epsilon_n^2 & \text{in the case of second order,} \qquad \text{etc.} \end{array} \tag{7}$$

Thus if $\epsilon_n = 10^{-k}$ in some step, then for second order, $\epsilon_{n+1} = c \cdot (10^{-k})^2 = c \cdot 10^{-2k}$, so that the number of significant digits is about doubled in each step.

Convergence of Newton's Method

In Newton's method, $g(x) = x - f(x)/f'(x)$. By differentiation,

$$\begin{aligned} g'(x) &= 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} \\ &= \frac{f(x)f''(x)}{f'(x)^2} . \end{aligned} \tag{8}$$

Since $f(s) = 0$, this shows that also $g'(s) = 0$. Hence Newton's method is at least of second order. If we differentiate again and set $x = s$, we find that

$$g''(s) = \frac{f''(s)}{f'(s)} \tag{8*}$$

which will not be zero in general. This proves

THEOREM 2

Second-Order Convergence of Newton's Method

If $f(x)$ is three times differentiable and f' and f'' are not zero at a solution s of $f(x) = 0$, then for x_0 sufficiently close to s, Newton's method is of second order.

Comments. For Newton's method, (7b) becomes, by (8*),

$$\epsilon_{n+1} \approx -\frac{f''(s)}{2f'(s)}\epsilon_n^2. \tag{9}$$

For the rapid convergence of the method indicated in Theorem 2 it is important that s be a *simple* zero of $f(x)$ (thus $f'(s) \neq 0$) and that x_0 be close to s, because in Taylor's formula we took only the linear term [see (5*)], assuming the quadratic term to be negligibly small. (With a bad x_0 the method may even diverge!)

EXAMPLE 6 Prior Error Estimate of the Number of Newton Iteration Steps

Use $x_0 = 2$ and $x_1 = 1.901$ in Example 4 for estimating how many iteration steps we need to produce the solution to 5D-accuracy. This is an **a priori estimate** or **prior estimate** because we can compute it after only one iteration, prior to further iterations.

Solution. We have $f(x) = x - 2\sin x = 0$. Differentiation gives

$$\frac{f''(s)}{2f'(s)} \approx \frac{f''(x_1)}{2f'(x_1)} = \frac{2\sin x_1}{2(1 - 2\cos x_1)} \approx 0.57.$$

Hence (9) gives

$$|\epsilon_{n+1}| \approx 0.57\epsilon_n^2 \approx 0.57(0.57\epsilon_{n-1}^2)^2 = 0.57^3\epsilon_{n-1}^4 \approx \cdots \approx 0.57^M\epsilon_0^{M+1} \leqq 5 \cdot 10^{-6}$$

where $M = 2^n + 2^{n-1} + \cdots + 2 + 1 = 2^{n+1} - 1$. We show below that $\epsilon_0 \approx -0.11$. Consequently, our condition becomes

$$0.57^M 0.11^{M+1} \leqq 5 \cdot 10^{-6}.$$

Hence $n = 2$ is the smallest possible n, according to this crude estimate, in good agreement with Example 4.

$\epsilon_0 \approx -0.11$ is obtained from $\epsilon_1 - \epsilon_0 = (\epsilon_1 - s) - (\epsilon_0 - s) = -x_1 + x_0 \approx 0.10$, hence $\epsilon_1 = \epsilon_0 + 0.10 \approx -0.57\epsilon_0^2$ or $0.57\epsilon_0^2 + \epsilon_0 + 0.10 \approx 0$, which gives $\epsilon_0 \approx -0.11$. ■

Difficulties in Newton's Method. Difficulties may arise if $|f'(x)|$ is very small near a solution s of $f(x) = 0$. For instance, let s be a zero of $f(x)$ of second or higher order. Then Newton's method converges only linearly, as is shown by an application of l'Hopital's rule to (8). Geometrically, small $|f'(x)|$ means that the tangent of $f(x)$ near s almost coincides with the x-axis (so that double precision may be needed to get $f(x)$ and $f'(x)$ accurately enough). Then for values $x = \tilde{s}$ far away from s we can still have small function values

$$R(\tilde{s}) = f(\tilde{s}).$$

In this case we call the equation $f(x) = 0$ **ill-conditioned**. $R(\tilde{s})$ is called the **residual** of $f(x) = 0$ at $\tilde{s}$. Thus a small residual guarantees a small error of $\tilde{s}$ only if the equation is ***not*** ill-conditioned.

EXAMPLE 7 An Ill-Conditioned Equation

$f(x) = x^5 + 10^{-4}x = 0$ is ill-conditioned, $x = 0$ is a solution. $f'(0) = 10^{-4}$ is small. At $\tilde{s} = 0.1$ the residual $f(0.1) = 2 \cdot 10^{-5}$ is small, but the error -0.1 is larger in absolute value by a factor 5000. Invent a more drastic example of your own. ■

Secant Method for Solving $f(x) = 0$

Newton's method is very powerful but has the disadvantage that the derivative f' may sometimes be a far more difficult expression than f itself and its evaluation therefore computationally expensive. This situation suggests the idea of replacing the derivative with the difference quotient

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} .$$

Then instead of (5) we have the formula of the popular secant method

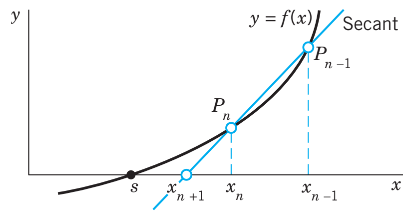


Fig. 429. Secant method

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} . \tag{10}$$

Geometrically, we intersect the x-axis at x_{n+1} with the secant of $f(x)$ passing through P_{n-1} and P_n in Fig. 429. We need two starting values x_0 and x_1. Evaluation of derivatives is now avoided. It can be shown that convergence is **superlinear** (that is, more rapid than linear, $|\epsilon_{n+1}| \approx \text{const} \cdot |\epsilon_n|^{1.62}$; see [E5] in App. 1), almost quadratic like Newton's method. The algorithm is similar to that of Newton's method, as the student may show.

CAUTION! It is ***not*** good to write (10) as

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} ,$$

because this may lead to loss of significant digits if x_n and x_{n-1} are about equal. (Can you see this from the formula?)

EXAMPLE 8 Secant Method

Find the positive solution of $f(x) = x - 2 \sin x = 0$ by the secant method, starting from $x_0 = 2, x_1 = 1.9$.

Solution. Here, (10) is

$$x_{n+1} = x_n - \frac{(x_n - 2\sin x_n)(x_n - x_{n-1})}{x_n - x_{n-1} + 2(\sin x_{n-1} - \sin x_n)} = x_n - \frac{N_n}{D_n} .$$

Numeric values are:

n	x_{n-1}	x_n	N_n	D_n	$x_{n+1} - x_n$
1	2.000000	1.900000	−0.000740	−0.174005	−0.004253
2	1.900000	1.895747	−0.000002	−0.006986	−0.000252
3	1.895747	1.895494	0		0

$x_3 = 1.895494$ is exact to 6D. See Example 4. ■

Summary of Methods. The methods for computing solutions s of $f(x) = 0$ with given continuous (or differentiable) $f(x)$ start with an initial approximation x_0 of s and generate a sequence $x_1, x_2, \cdots$ by **iteration**. **Fixed-point methods** solve $f(x) = 0$ written as $x = g(x)$, so that s is a *fixed point* of g, that is, $s = g(s)$. For $g(x) = x - f(x)/f'(x)$ this is **Newton's method**, which, for good x_0 and simple zeros, converges quadratically (and for multiple zeros linearly). From Newton's method the **secant method** follows by replacing $f'(x)$ by a difference quotient. The ***bisection method*** and the ***method of false position*** in Problem Set 19.2 always converge, but often slowly.

PROBLEM SET 19.2

1–13 FIXED-POINT ITERATION

Solve by fixed-point iteration and answer related questions where indicated. Show details.

1. **Monotone sequence.** Why is the sequence in Example 1 monotone? Why not in Example 2?
2. Do the iterations (b) in Example 2. Sketch a figure similar to Fig. 427. Explain what happens.
3. $f = x - 0.5 \cos x = 0$, $\quad x_0 = 1$. Sketch a figure.
4. $f = x - \operatorname{cosec} x$ the zero near $x = 1$.
5. Sketch $f(x) = x^3 - 5.00x^2 + 1.01x + 1.88$, showing roots near ± 1 and 5. Write $x = g(x) = (5.00x^2 - 1.01x + 1.88)/x^2$. Find a root by starting from $x_0 = 5, 4, 1, -1$. Explain the (perhaps unexpected) results.
6. Find a form $x = g(x)$ of $f(x) = 0$ in Prob. 5 that yields convergence to the root near $x = 1$.
7. Find the smallest positive solution of $\sin x = e^{-x}$.
8. Solve $x^4 - x - 0.12 = 0$ by starting from $x_0 = 1$.
9. Find the negative solution of $x^4 - x - 0.12 = 0$.
10. **Elasticity.** Solve $x \cosh x = 1$. (Similar equations appear in vibrations of beams; see Problem Set 12.3.)
11. **Drumhead. Bessel functions.** A partial sum of the Maclaurin series of $J_0(x)$ (Sec. 5.5) is $f(x) = 1 - \frac{1}{4}x^2 + \frac{1}{64}x^4 - \frac{1}{2304}x^6$. Conclude from a sketch that $f(x) = 0$ near $x = 2$. Write $f(x) = 0$ as $x = g(x)$ (by dividing $f(x)$ by $\frac{1}{4}x$ and taking the resulting x-term to the other side). Find the zero. (See Sec. 12.10 for the importance of these zeros.)
12. **CAS EXPERIMENT. Convergence.** Let $f(x) = x^3 + 2x^2 - 3x - 4 = 0$. Write this as $x = g(x)$, for g choosing (1) $(x^3 - f)^{1/3}$, (2) $(x^2 - \frac{1}{2}f)^{1/2}$, (3) $x + \frac{1}{3}f$, (4) $(x^3 - f)/x^2$, (5) $(2x^2 - f)/(2x)$, and (6) $x - f/f'$ and in each case $x_0 = 1.5$. Find out about convergence and divergence and the number of steps to reach 6S-values of a root.
13. **Existence of fixed point.** Prove that if g is continuous in a closed interval I and its range lies in I, then the equation $x = g(x)$ has at least one solution in I. Illustrate that it may have more than one solution in I.

14–23 NEWTON'S METHOD

Apply Newton's method (6S-accuracy). First sketch the function(s) to see what is going on.

14. **Cube root.** Design a Newton iteration. Compute $\sqrt[3]{7}$, $x_0 = 2$.
15. $f = 2x - \cos x$, $\quad x_0 = 1$. Compare with Prob. 3.
16. What happens in Prob. 15 for any other x_0?
17. **Dependence on x_0.** Solve Prob. 5 by Newton's method with $x_0 = 5, 4, 1, -3$. Explain the result.
18. **Legendre polynomials.** Find the largest root of the Legendre polynomial $P_5(x)$ given by $P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$ (Sec. 5.3) (to be needed in ***Gauss integration*** in Sec. 19.5) **(a)** by Newton's method, **(b)** from a quadratic equation.
19. **Associated Legendre functions.** Find the smallest positive zero of $P_4^2 = (1 - x^2)P_4'' = \frac{15}{2}(-7x^4 + 8x^2 - 1)$ (Sec. 5.3) **(a)** by Newton's method, **(b)** exactly, by solving a quadratic equation.
20. $x + \ln x = 2$, $\quad x_0 = 2$
21. $f = x^3 - 5x + 3 = 0$, $\quad x_0 = 2, \quad 0, -2$
22. **Heating, cooling.** At what time x (4S-accuracy only) will the processes governed by $f_1(x) = 100(1 - e^{-0.2x})$ and $f_2(x) = 40e^{-0.01x}$ reach the same temperature? Also find the latter.
23. **Vibrating beam.** Find the solution of $\cos x \cosh x = 1$ near $x = \frac{3}{2}\pi$. (This determines a frequency of a vibrating beam; see Problem Set 12.3.)
24. **Method of False Position (Regula falsi).** Figure 430 shows the idea. We assume that f is continuous. We compute the x-intercept c_0 of the line through $(a_0, f(a_0))$, $(b_0, f(b_0))$. If $f(c_0) = 0$, we are done. If $f(a_0)f(c_0) < 0$ (as in Fig. 430), we set $a_1 = a_0$, $b_1 = c_0$ and repeat to get c_1, etc. If $f(a_0)f(c_0) > 0$, then $f(c_0)f(b_0) < 0$ and we set $a_1 = c_0$, $b_1 = b_0$, etc.

    (a) Algorithm. Show that

    $$c_0 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)}$$

    and write an algorithm for the method.

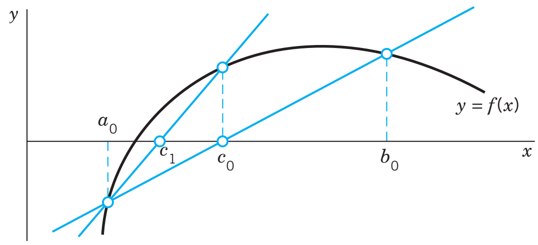


Fig. 430. Method of false position

(b) Solve $x^4 = 2$, $\cos x = \sqrt{x}$, and $x + \ln x = 2$, with $a = 1, b = 2$.

25. TEAM PROJECT. Bisection Method. This simple but slowly convergent method for finding a solution of $f(x) = 0$ with continuous f is based on the **intermediate value theorem**, which states that if a continuous function f has opposite signs at some $x = a$ and $x = b (> a)$, that is, either $f(a) < 0, f(b) > 0$ or $f(a) > 0, f(b) < 0$, then f must be 0 somewhere on $[a, b]$. The solution is found by repeated bisection of the interval and in each iteration picking that half which also satisfies that sign condition.

(a) Algorithm. Write an algorithm for the method.

(b) Comparison. Solve $x = \cos x$ by Newton's method and by bisection. Compare.

(c) Solve $e^{-x} = \ln x$ and $e^x + x^4 + x = 2$ by bisection.

26–29 SECANT METHOD

Solve, using x_0 and x_1 as indicated:

26. $e^{-x} - \tan x = 0, \quad x_0 = 1, \quad x_1 = 0.7$

27. Prob. 21, $x_0 = 1.0, \quad x_1 = 2.0$

28. $x = \cos x, \quad x_0 = 0.5, \quad x_1 = 1$

29. $\sin x = \cot x, \quad x_0 = 1, \quad x_1 = 0.5$

30. WRITING PROJECT. Solution of Equations. Compare the methods in this section and problem set, discussing advantages and disadvantages in terms of examples of your own. No proofs, just motivations and ideas.

19.3 Interpolation

We are given the values of a function $f(x)$ at different points $x_0, x_1, \cdots, x_n$. We want to find approximate values of the function $f(x)$ for "new" x's that lie between these points for which the function values are given. This process is called **interpolation**. The student should pay close attention to this section as interpolation forms the underlying foundation for both Secs. 19.4 and 19.5. Indeed, interpolation allows us to develop formulas for numeric integration and differentiation as shown in Sec. 19.5.

Continuing our discussion, we write these given values of a function f in the form

$$f_0 = f(x_0), \qquad f_1 = f(x_1), \qquad \cdots, \qquad f_n = f(x_n)$$

or as ordered pairs

$$(x_0, f_0), \qquad (x_1, f_1), \qquad \cdots, \quad (x_n, f_n).$$

Where do these given function values come from? They may come from a "mathematical" function, such as a logarithm or a Bessel function. More frequently, they may be measured or automatically recorded values of an "empirical" function, such as air resistance of a car or an airplane at different speeds. Other examples of functions that are "empirical" are the yield of a chemical process at different temperatures or the size of the U.S. population as it appears from censuses taken at 10-year intervals.

A standard idea in interpolation now is to find a polynomial $p_n(x)$ of degree n (or less) that assumes the given values; thus

$$p_n(x_0) = f_0, \qquad p_n(x_1) = f_1, \qquad \cdots, \qquad p_n(x_n) = f_n. \tag{1}$$

We call this p_n an **interpolation polynomial** and $x_0, \cdots, x_n$ the **nodes**. And if $f(x)$ is a mathematical function, we call p_n an **approximation** of f (or a **polynomial approximation**, because there are other kinds of approximations, as we shall see later). We use p_n to get (approximate) values of f for x's between x_0 and x_n ("**interpolation**") or sometimes outside this interval $x_0 \leqq x \leqq x_n$ ("**extrapolation**").

Motivation. Polynomials are convenient to work with because we can readily differentiate and integrate them, again obtaining polynomials. Moreover, they approximate *continuous* functions with any desired accuracy. That is, for any continuous $f(x)$ on an interval J: $a \leqq x \leqq b$ and error bound $\beta > 0$, there is a polynomial $p_n(x)$ (of sufficiently high degree n) such that

$$|f(x) - p_n(x)| < \beta \quad \text{for all } x \text{ on } J.$$

This is the famous **Weierstrass approximation theorem** (for a proof see Ref. [GenRef7], App. 1).

Existence and Uniqueness. Note that the interpolation polynomial p_n satisfying (1) for given data exists and we shall give formulas for it below. Furthermore, p_n is unique: Indeed, if another polynomial q_n also satisfies $q_n(x_0) = f_0, \cdots, q_n(x_n) = f_n$, then $p_n(x) - q_n(x) = 0$ at $x_0, \cdots, x_n$, but a polynomial $p_n - q_n$ of degree n (or less) with $n + 1$ roots must be identically zero, as we know from algebra; thus $p_n(x) = q_n(x)$ for all x, which means uniqueness. ■

How Do We Find p_n? We shall explain several standard methods that give us p_n. By the uniqueness proof above, we know that, for given data, the different methods *must* give us the same polynomial. However, the polynomials may be expressed in different forms suitable for different purposes.

Lagrange Interpolation

Given $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with arbitrarily spaced x_j, Lagrange had the idea of multiplying each f_j by a polynomial that is 1 at x_j and 0 at the other n nodes and then taking the sum of these $n + 1$ polynomials. Clearly, this gives the unique interpolation polynomial of degree n or less. Beginning with the simplest case, let us see how this works.

Linear interpolation is interpolation by the straight line through $(x_0, f_0), (x_1, f_1)$; see Fig. 431. Thus the linear Lagrange polynomial p_1 is a sum $p_1 = L_0 f_0 + L_1 f_1$ with L_0 the linear polynomial that is 1 at x_0 and 0 at x_1; similarly, L_1 is 0 at x_0 and 1 at x_1. Obviously,

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

This gives the linear Lagrange polynomial

$$p_1(x) = L_0(x) f_0 + L_1(x) f_1 = \frac{x - x_1}{x_0 - x_1} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1. \tag{2}$$

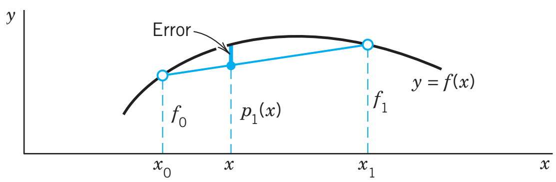


Fig. 431. Linear Interpolation

EXAMPLE 1 **Linear Lagrange Interpolation**

Compute a 4D-value of ln 9.2 from ln 9.0 = 2.1972, ln 9.5 = 2.2513 by linear Lagrange interpolation and determine the error, using ln 9.2 = 2.2192 (4D).

Solution. $x_0 = 9.0$, $x_1 = 9.5$, $f_0 = \ln 9.0$, $f_1 = \ln 9.5$. Ln (2) we need

$$L_0(x) = \frac{x - 9.5}{-0.5} = -2.0(x - 9.5), \qquad L_0(9.2) = -2.0(-0.3) = 0.6$$

$$L_1(x) = \frac{x - 9.0}{0.5} = 2.0(x - 9.0), \qquad L_1(9.2) = 2 \cdot 0.2 = 0.4$$

(see Fig. 432) and obtain the answer

$$\ln 9.2 \approx p_1(9.2) = L_0(9.2) f_0 + L_1(9.2) f_1 = 0.6 \cdot 2.1972 + 0.4 \cdot 2.2513 = 2.2188.$$

The error is $\epsilon = a - \tilde{a} = 2.2192 - 2.2188 = 0.0004$. Hence linear interpolation is not sufficient here to get 4D accuracy; it would suffice for 3D accuracy.

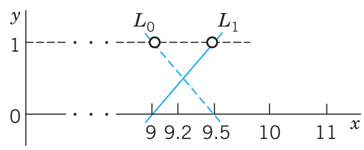


Fig. 432. L_0 and L_1 in Example 1

Quadratic interpolation is interpolation of given (x_0, f_0), (x_1, f_1), (x_2, f_2) by a second-degree polynomial $p_2(x)$, which by Lagrange's idea is

$$p_2(x) = L_0(x)f_0 + L_1(x)f_1 + L_2(x)f_2 \tag{3a}$$

with $L_0(x_0) = 1$, $L_1(x_1) = 1$, $L_2(x_2) = 1$, and $L_0(x_1) = L_0(x_2) = 0$, etc. We claim that

$$\begin{aligned} L_0(x) &= \frac{l_0(x)}{l_0(x_0)} = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \\ L_1(x) &= \frac{l_1(x)}{l_1(x_1)} = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \\ L_2(x) &= \frac{l_2(x)}{l_2(x_2)} = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}. \end{aligned} \tag{3b}$$

How did we get this? Well, the numerator makes $L_k(x_j) = 0$ if $j \neq k$. And the denominator makes $L_k(x_k) = 1$ because it equals the numerator at $x = x_k$.

EXAMPLE 2 **Quadratic Lagrange Interpolation**

Compute ln 9.2 by (3) from the data in Example 1 and the additional third value ln 11.0 = 2.3979.

Solution. In (3),

$$L_0(x) = \frac{(x - 9.5)(x - 11.0)}{(9.0 - 9.5)(9.0 - 11.0)} = x^2 - 20.5x + 104.5, \qquad L_0(9.2) = 0.5400,$$

$$L_1(x) = \frac{(x - 9.0)(x - 11.0)}{(9.5 - 9.0)(9.5 - 11.0)} = -\frac{1}{0.75}(x^2 - 20x + 99), \qquad L_1(9.2) = 0.4800,$$

$$L_2(x) = \frac{(x - 9.0)(x - 9.5)}{(11.0 - 9.0)(11.0 - 9.5)} = \frac{1}{3}(x^2 - 18.5x + 85.5), \qquad L_2(9.2) = -0.0200,$$

(see Fig. 433), so that (3a) gives, exact to 4D,

$$\ln 9.2 \approx p_2(9.2) = 0.5400 \cdot 2.1972 + 0.4800 \cdot 2.2513 - 0.0200 \cdot 2.3979 = 2.2192.$$

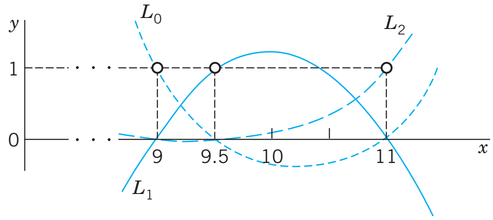


Fig. 433. L_0, L_1, L_2 in Example 2

General Lagrange Interpolation Polynomial. For general n we obtain

(4a)
$$f(x) \approx p_n(x) = \sum_{k=0}^{n} L_k(x) f_k = \sum_{k=0}^{n} \frac{l_k(x)}{l_k(x_k)} f_k$$

where $L_k(x_k) = 1$ and L_k is 0 at the other nodes, and the L_k are independent of the function f to be interpolated. We get (4a) if we take

(4b)
$$\begin{aligned} l_0(x) &= (x - x_1)(x - x_2) \cdots (x - x_n), \\ l_k(x) &= (x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n), \qquad 0 < k < n, \\ l_n(x) &= (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

We can easily see that $p_n(x_k) = f_k$. Indeed, inspection of (4b) shows that $l_k(x_j) = 0$ if $j \neq k$, so that for $x = x_k$, the sum in (4a) reduces to the single term $(l_k(x_k)/l_k(x_k)) f_k = f_k$.

Error Estimate. If f is itself a polynomial of degree n (or less), it must coincide with p_n because the $n + 1$ data $(x_0, f_0), \cdots, (x_n, f_n)$ determine a polynomial uniquely, so the error is zero. Now the special f has its $(n + 1)$st derivative identically zero. This makes it plausible that for a *general* f its $(n + 1)$st derivative $f^{(n+1)}$ should measure the error

$$\epsilon_n(x) = f(x) - p_n(x).$$

It can be shown that this is true if $f^{(n+1)}$ exists and is continuous. Then, with a suitable t between x_0 and x_n (or between x_0, x_n, and x if we extrapolate),

(5)
$$\epsilon_n(x) = f(x) - p_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(t)}{(n + 1)!}.$$

Thus $|\epsilon_n(x)|$ is 0 at the nodes and small near them, because of continuity. The product $(x - x_0) \cdots (x - x_n)$ is large for x away from the nodes. This makes extrapolation risky. And interpolation at an x will be best if we choose nodes on both sides of that x. Also, we get error bounds by taking the smallest and the largest value of $f^{(n+1)}(t)$ in (5) on the interval $x_0 \leqq t \leqq x_n$ (or on the interval also containing x if we *extra*polate).

Most importantly, since p_n is unique, as we have shown, we have

THEOREM 1

Error of Interpolation

Formula (5) *gives the error for* ***any*** *polynomial interpolation method if* $f(x)$ *has a continuous* $(n+1)$*st derivative.*

Practical error estimate. If the derivative in (5) is difficult or impossible to obtain, apply the Error Principle (Sec. 19.1), that is, take another node and the Lagrange polynomial $p_{n+1}(x)$ and regard $p_{n+1}(x) - p_n(x)$ as a (crude) error estimate for $p_n(x)$.

EXAMPLE 3 Error Estimate (5) of Linear Interpolation. Damage by Roundoff. Error Principle

Estimate the error in Example 1 first by (5) directly and then by the Error Principle (Sec. 19.1).

Solution. **(A)** ***Estimation by* (5).** We have $n = 1$, $f(t) = \ln t$, $f'(t) = 1/t$, $f''(t) = -1/t^2$. Hence

$$\epsilon_1(x) = (x - 9.0)(x - 9.5)\frac{(-1)}{2t^2}, \qquad \text{thus} \qquad \epsilon_1(9.2) = \frac{0.03}{t^2}.$$

$t = 0.9$ gives the maximum $0.03/9^2 = 0.00037$ and $t = 9.5$ gives the minimum $0.03/9.5^2 = 0.00033$, so that we get $0.00033 \leqq \epsilon_1(9.2) \leqq 0.00037$, or better, 0.00038 because $0.3/81 = 0.003703 \cdots$.

But the error 0.0004 in Example 1 disagrees, and we can learn something! Repetition of the computation there with 5D instead of 4D gives

$$\ln 9.2 \approx p_1(9.2) = 0.6 \cdot 2.19722 + 0.4 \cdot 2.25129 = 2.21885$$

with an actual error $\epsilon = 2.21920 - 2.21885 = 0.00035$, which lies nicely near the middle between our two error bounds.

This shows that the discrepancy (0.0004 vs. 0.00035) was caused by rounding, which is not taken into account in (5).

(B) ***Estimation by the Error Principle.*** We calculate $p_1(9.2) = 2.21885$ as before and then $p_2(9.2)$ as in Example 2 but with 5D, obtaining

$$p_2(9.2) = 0.54 \cdot 2.19722 + 0.48 \cdot 2.25129 - 0.02 \cdot 2.39790 = 2.21916.$$

The difference $p_2(9.2) - p_1(9.2) = 0.00031$ is the approximate error of $p_1(9.2)$ that we wanted to obtain; this is an approximation of the actual error 0.00035 given above. ■

Newton's Divided Difference Interpolation

For given data $(x_0, f_0), \cdots, (x_n, f_n)$ the interpolation polynomial $p_n(x)$ satisfying (1) is unique, as we have shown. But for different purposes we may use $p_n(x)$ in different forms. **Lagrange's form** just discussed is useful for deriving formulas in numeric differentiation (approximation formulas for derivatives) and integration (Sec. 19.5).

Practically more important are Newton's forms of $p_n(x)$, which we shall also use for solving ODEs (in Sec. 21.2). They involve fewer arithmetic operations than Lagrange's form. Moreover, it often happens that we have to increase the degree n to reach a required accuracy. Then in Newton's forms we can use all the previous work and just add another term, a possibility without counterpart for Lagrange's form. This also simplifies the application of the Error Principle (used in Example 3 for Lagrange). The details of these ideas are as follows.

Let $p_{n-1}(x)$ be the $(n-1)$st Newton polynomial (whose form we shall determine); thus $p_{n-1}(x_0) = f_0$, $p_{n-1}(x_1) = f_1, \cdots, p_{n-1}(x_{n-1}) = f_{n-1}$. Furthermore, let us write the nth Newton polynomial as

$$p_n(x) = p_{n-1}(x) + g_n(x); \tag{6}$$

hence

$$g_n(x) = p_n(x) - p_{n-1}(x). \tag{6'}$$

Here $g_n(x)$ is to be determined so that $p_n(x_0) = f_0, p_n(x_1) = f_1, \cdots, p_n(x_n) = f_n$.

Since p_n and p_{n-1} agree at $x_0, \cdots, x_{n-1}$, we see that g_n is zero there. Also, g_n will generally be a polynomial of nth degree because so is p_n, whereas p_{n-1} can be of degree $n - 1$ at most. Hence g_n must be of the form

$$g_n(x) = a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{6''}$$

We determine the constant a_n. For this we set $x = x_n$ and solve $(6'')$ algebraically for a_n. Replacing $g_n(x_n)$ according to $(6')$ and using $p_n(x_n) = f_n$, we see that this gives

$$a_n = \frac{f_n - p_{n-1}(x_n)}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})}. \tag{7}$$

We write a_k instead of a_n and show that a_k equals the **kth divided difference**, recursively denoted and defined as follows:

$$a_1 = f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0}$$

$$a_2 = f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

and in general

$$a_k = f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}. \tag{8}$$

If $n = 1$, then $p_{n-1}(x_n) = p_0(x_1) = f_0$ because $p_0(x)$ is constant and equal to f_0, the value of $f(x)$ at x_0. Hence (7) gives

$$a_1 = \frac{f_1 - p_0(x_1)}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = f[x_0, x_1],$$

and (6) and $(6'')$ give the Newton interpolation polynomial of the first degree

$$p_1(x) = f_0 + (x - x_0)f[x_0, x_1].$$

If $n = 2$, then this p_1 and (7) give

$$a_2 = \frac{f_2 - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)} = \frac{f_2 - f_0 - (x_2 - x_0)f[x_0, x_1]}{(x_2 - x_0)(x_2 - x_1)} = f[x_0, x_1, x_2]$$

where the last equality follows by straightforward calculation and comparison with the definition of the right side. (Verify it; be patient.) From (6) and $(6'')$ we thus obtain the second Newton polynomial

$$p_2(x) = f_0 + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2].$$

For $n = k$, formula (6) gives

$$p_k(x) = p_{k-1}(x) + (x - x_0)(x - x_1)\cdots(x - x_{k-1})f[x_0, \cdots, x_k]. \tag{9}$$

With $p_0(x) = f_0$ by repeated application with $k = 1, \cdots, n$ this finally gives **Newton's divided difference interpolation formula**

$$\begin{aligned} f(x) \approx f_0 &+ (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &+ \cdots + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f[x_0, \cdots, x_n]. \end{aligned} \tag{10}$$

An algorithm is shown in Table 19.2. The first do-loop computes the divided differences and the second the desired value $p_n(\hat{x})$.

Example 4 shows how to arrange differences near the values from which they are obtained; the latter always stand a half-line above and a half-line below in the preceding column. Such an arrangement is called a (divided) **difference table**.

Table 19.2 Newton's Divided Difference Interpolation

ALGORITHM INTERPOL ($x_0, \cdots, x_n$; $f_0, \cdots, f_n$; $\hat{x}$)

This algorithm computes an approximation $p_n(\hat{x})$ of $f(\hat{x})$ at $\hat{x}$.

INPUT: Data $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$; $\hat{x}$

OUTPUT: Approximation $p_n(\hat{x})$ of $f(\hat{x})$

Set $f[x_j] = f_j \quad (j = 0, \cdots, n)$.

For $m = 1, \cdots, n - 1$) do:

For $j = 0, \cdots, n - m$ do:

$$f[x_j, \cdots, x_{j+m}] = \frac{f[x_{j+1}, \cdots, x_{j+m}] - f[x_j, \cdots, x_{j+m-1}]}{x_{j+m} - x_j}$$

End

End

Set $p_0(x) = f_0$.

For $k = 1, \cdots, n$ do:

$$p_k(\hat{x}) = p_{k-1}(\hat{x}) + (\hat{x} - x_0)\cdots(\hat{x} - x_{k-1})f[x_0, \cdots, x_k]$$

End

OUTPUT $p_n(\hat{x})$

End INTERPOL

EXAMPLE 4 Newton's Divided Difference Interpolation Formula

Compute $f(9.2)$ from the values shown in the first two columns of the following table.

x_j	$f_j = f(x_j)$	$f[x_j, x_{j+1}]$	$f[x_j, x_{j+1}, x_{j+2}]$	$f[x_j, \cdots, x_{j+3}]$
8.0	2.079442			
		0.117783		
9.0	2.197225		−0.006433	
		0.108134		0.000411
9.5	2.251292		−0.005200	
		0.097735		
11.0	2.397895			

Solution. We compute the divided differences as shown. Sample computation:

$$(0.097735 - 0.108134)/(11 - 9) = -0.005200.$$

The values we need in (10) are circled. We have

$$f(x) \approx p_3(x) = 2.079442 + 0.117783(x - 8.0) - 0.006433(x - 8.0)(x - 9.0) + 0.000411(x - 8.0)(x - 9.0)(x - 9.5).$$

At $x = 9.2$,

$$f(9.2) \approx 2.079442 + 0.141340 - 0.001544 - 0.000030 = 2.219208.$$

The value exact to 6D is $f(9.2) = \ln 9.2 = 2.219203$. Note that we can nicely see how the accuracy increases from term to term:

$$p_1(9.2) = 2.220782, \qquad p_2(9.2) = 2.219238, \qquad p_3(9.2) = 2.219208.$$

Equal Spacing: Newton's Forward Difference Formula

Newton's formula (10) is valid for *arbitrarily spaced* nodes as they may occur in practice in experiments or observations. However, in many applications the x_j's are *regularly spaced*—for instance, in measurements taken at regular intervals of time. Then, denoting the distance by h, we can write

$$x_0, \quad x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \quad \cdots, \quad x_n = x_0 + nh. \tag{11}$$

We show how (8) and (10) now simplify considerably!

To get started, let us define the *first forward difference* of f at x_j by

$$\Delta f_j = f_{j+1} - f_j,$$

the *second forward difference* of f at x_j by

$$\Delta^2 f_j = \Delta f_{j+1} - \Delta f_j,$$

and, continuing in this way, the ***k*th forward difference** of f at x_j by

$$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 1, 2, \cdots). \tag{12}$$

Examples and an explanation of the name "forward" follow on the next page. What is the point of this? We show that if we have regular spacing (11), then

(13)
$$f[x_0, \cdots, x_k] = \frac{1}{k!h^k} \Delta^k f_0.$$

PROOF We prove (13) by induction. It is true for $k = 1$ because $x_1 = x_0 + h$, so that

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{1}{h}(f_1 - f_0) = \frac{1}{1!h} \Delta f_0.$$

Assuming (13) to be true for all forward differences of order k, we show that (13) holds for $k + 1$. We use (8) with $k + 1$ instead of k; then we use $(k + 1)h = x_{k+1} - x_0$, resulting from (11), and finally (12) with $j = 0$, that is, $\Delta^{k+1} f_0 = \Delta^k f_1 - \Delta^k f_0$. This gives

$$\begin{aligned} f[x_0, \cdots, x_{k+1}] &= \frac{f[x_1, \cdots, x_{k+1}] - f[x_0, \cdots, x_k]}{(k + 1)h} \\ &= \frac{1}{(k + 1)h} \left[\frac{1}{k!h^k} \Delta^k f_1 - \frac{1}{k!h^k} \Delta^k f_0 \right] \\ &= \frac{1}{(k + 1)!h^{k+1}} \Delta^{k+1} f_0 \end{aligned}$$

which is (13) with $k + 1$ instead of k. Formula (13) is proved. ■

In (10) we finally set $x = x_0 + rh$. Then $x - x_0 = rh$, $x - x_1 = (r - 1)h$ since $x_1 - x_0 = h$, and so on. With this and (13), formula (10) becomes **Newton's** (or *Gregory*[2]*–Newton's*) **forward difference interpolation formula**

(14)
$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r}{s} \Delta^s f_0 \qquad (x = x_0 + rh, \quad r = (x - x_0)/h) \\ &= f_0 + r\Delta f_0 + \frac{r(r - 1)}{2!} \Delta^2 f_0 + \cdots + \frac{r(r - 1) \cdots (r - n + 1)}{n!} \Delta^n f_0 \end{aligned}$$

where the **binomial coefficients** in the first line are defined by

(15)
$$\binom{r}{0} = 1, \quad \binom{r}{s} = \frac{r(r - 1)(r - 2) \cdots (r - s + 1)}{s!} \qquad (s > 0, \text{ integer})$$

and $s! = 1 \cdot 2 \cdots s$.

Error. From (5) we get, with $x - x_0 = rh$, $x - x_1 = (r - 1)h$, etc.,

(16)
$$\epsilon_n(x) = f(x) - p_n(x) = \frac{h^{n+1}}{(n + 1)!} r(r - 1) \cdots (r - n) f^{(n+1)}(t)$$

with t as characterized in (5).

[2]JAMES GREGORY (1638–1675), Scots mathematician, professor at St. Andrews and Edinburgh. Δ in (14) and ∇^2 (on p. 818) have nothing to do with the Laplacian.

Formula (16) is an exact formula for the error, but it involves the unknown t. In Example 5 (below) we show how to use (16) for obtaining an error estimate and an interval in which the true value of $f(x)$ must lie.

Comments on Accuracy. (A) The order of magnitude of the error $\epsilon_n(x)$ is about equal to that of the next difference not used in $p_n(x)$.

(B) One should choose $x_0, \cdots, x_n$ such that the x at which one interpolates is as well centered between $x_0, \cdots, x_n$ as possible.

The reason for (A) is that in (16),

$$f^{n+1}(t) \approx \frac{\Delta^{n+1} f(t)}{h^{n+1}}, \qquad \frac{|r(r-1)\cdots(r-n)|}{1 \cdot 2 \cdots (n+1)} \leqq 1 \quad \text{if} \quad |r| \leqq 1$$

(and actually for any r as long as we do not *extrapolate*). The reason for (B) is that $|r(r-1)\cdots(r-n)|$ becomes smallest for that choice.

EXAMPLE 5 **Newton's Forward Difference Formula. Error Estimation**

Compute cosh 0.56 from (14) and the four values in the following table and estimate the error.

j	x_j	$f_j = \cosh x_j$	Δf_j	$\Delta^2 f_j$	$\Delta^3 f_j$
0	0.5	1.127626			
			0.057839		
1	0.6	1.185465		0.011865	
			0.069704		0.000697
2	0.7	1.255169		0.012562	
			0.082266		
3	0.8	1.337435			

Solution. We compute the forward differences as shown in the table. The values we need are circled. In (14) we have $r = (0.56 - 0.50)/0.1 = 0.6$, so that (14) gives

$$\begin{aligned}\cosh 0.56 &\approx 1.127626 + 0.6 \cdot 0.057839 + \frac{0.6(-0.4)}{2} \cdot 0.011865 + \frac{0.6(-0.4)(-1.4)}{6} \cdot 0.000697 \\ &= 1.127626 + 0.034703 - 0.001424 + 0.000039 \\ &= 1.160944.\end{aligned}$$

Error estimate. From (16), since the fourth derivative is $\cosh^{(4)} t = \cosh t$,

$$\begin{aligned}\epsilon_3(0.56) &= \frac{0.1^4}{4!} \cdot 0.6(-0.4)(-1.4)(-2.4) \cosh t \\ &= A \cosh t,\end{aligned}$$

where $A = -0.00000336$ and $0.5 \leqq t \leqq 0.8$. We do not know t, but we get an inequality by taking the largest and smallest cosh t in that interval:

$$A \cosh 0.8 \leqq \epsilon_3(0.62) \leqq A \cosh 0.5.$$

Since

$$f(x) = p_3(x) + \epsilon_3(x),$$

this gives

$$p_3(0.56) + A \cosh 0.8 \leqq \cosh 0.56 \leqq p_3(0.56) + A \cosh 0.5.$$

Numeric values are

$$1.160939 \leqq \cosh 0.56 \leqq 1.160941.$$

The exact 6D-value is $\cosh 0.56 = 1.160941$. It lies within these bounds. Such bounds are not always so tight. Also, we did not consider roundoff errors, which will depend on the number of operations. ■

This example also explains the name "*forward* difference formula": we see that the differences in the formula slope forward in the difference table.

Equal Spacing: Newton's Backward Difference Formula

Instead of forward-sloping differences we may also employ backward-sloping differences. The difference table remains the same as before (same numbers, in the same positions), except for a very harmless change of the running subscript j (which we explain in Example 6, below). Nevertheless, purely for reasons of convenience it is standard to introduce a second name and notation for differences as follows. We define the *first backward difference* of f at x_j by

$$\nabla f_j = f_j - f_{j-1},$$

the *second backward difference* of f at x_j by

$$\nabla^2 f_j = \nabla f_j - \nabla f_{j-1},$$

and, continuing in this way, the ***k*th backward difference** of f at x_j by

$$\nabla^k f_j = \nabla^{k-1} f_j - \nabla^{k-1} f_{j-1} \qquad (k = 1, 2, \cdots). \tag{17}$$

A formula similar to (14) but involving backward differences is **Newton's** (or *Gregory–Newton's*) **backward difference interpolation formula**

$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r+s-1}{s} \nabla^s f_0 \qquad (x = x_0 + rh,\ r = (x - x_0)/h) \\ &= f_0 + r\nabla f_0 + \frac{r(r+1)}{2!} \nabla^2 f_0 + \cdots + \frac{r(r+1)\cdots(r+n-1)}{n!} \nabla^n f_0. \end{aligned} \tag{18}$$

EXAMPLE 6 **Newton's Forward and Backward Interpolations**

Compute a 7D-value of the Bessel function $J_0(x)$ for $x = 1.72$ from the four values in the following table, using (a) Newton's forward formula (14), (b) Newton's backward formula (18).

j_{for}	j_{back}	x_j	$J_0(x_j)$	1st Diff.	2nd Diff.	3rd Diff.
0	−3	1.7	0.3979849			
				−0.0579985		
1	−2	1.8	0.3399864		−0.0001693	
				−0.0581678		0.0004093
2	−1	1.9	0.2818186		0.0002400	
				−0.0579278		
3	0	2.0	0.2238908			

Solution. The computation of the differences is the same in both cases. Only their notation differs.

(a) Forward. In (14) we have $r = (1.72 - 1.70)/0.1 = 0.2$, and j goes from 0 to 3 (see first column). In each column we need the first given number, and (14) thus gives

$$
\begin{aligned}
J_0(1.72) &\approx 0.3979849 + 0.2(-0.0579985) + \frac{0.2(-0.8)}{2}(-0.0001693) + \frac{0.2(-0.8)(-1.8)}{6}\cdot 0.0004093 \\
&= 0.3979849 - 0.0115997 + 0.0000135 + 0.0000196 = 0.3864183,
\end{aligned}
$$

which is exact to 6D, the exact 7D-value being 0.3864185.

(b) Backward. For (18) we use j shown in the second column, and in each column the last number. Since $r = (1.72 - 2.00)/0.1 = -2.8$, we thus get from (18)

$$
\begin{aligned}
J_0(1.72) &\approx 0.2238908 - 2.8(-0.0579278) + \frac{-2.8(-1.8)}{2}\cdot 0.0002400 + \frac{-2.8(-1.8)(-0.8)}{6}\cdot 0.0004093 \\
&= 0.2238908 + 0.1621978 + 0.0006048 - 0.0002750 \\
&= 0.3864184.
\end{aligned}
$$

There is a third notation for differences, called the **central difference notation**. It is used in numerics for ODEs and certain interpolation formulas. See Ref. [E5] listed in App. 1.

PROBLEM SET 19.3

1. **Linear interpolation.** Calculate $p_1(x)$ in Example 1 and from it ln 9.3.
2. **Error estimate.** Estimate the error in Prob. 1 by (5).
3. **Quadratic interpolation. Gamma function.** Calculate the Lagrange polynomial $p_2(x)$ for the values $\Gamma(1.00) = 1.0000$, $\Gamma(1.02) = 0.9888$, $\Gamma(1.04) = 0.9784$ of the gamma function [(24) in App. A3.1] and from it approximations of $\Gamma(1.01)$ and $\Gamma(1.03)$.
4. **Error estimate for quadratic interpolation.** Estimate the error for $p_2(9.2)$ in Example 2 from (5).
5. **Linear and quadratic interpolation.** Find $e^{-0.25}$ and $e^{-0.75}$ by linear interpolation of e^{-x} with $x_0 = 0$, $x_1 = 0.5$ and $x_0 = 0.5$, $x_1 = 1$, respectively. Then find $p_2(x)$ by quadratic interpolation of e^{-x} with $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$ and from it $e^{-0.25}$ and $e^{-0.75}$. Compare the errors. Use 4S-values of e^{-x}.
6. **Interpolation and extrapolation.** Calculate $p_2(x)$ in Example 2. Compute from it approximations of ln 9.4, ln 10, ln 10.5, ln 11.5, and ln 12. Compute the errors by using exact 5S-values and comment.
7. **Interpolation and extrapolation.** Find the quadratic polynomial that agrees with $\sin x$ at $x = 0, \pi/4, \pi/2$ and use it for the interpolation and extrapolation of $\sin x$ at $x = -\pi/8, \pi/8, 3\pi/8, 5\pi/8$. Compute the errors.
8. **Extrapolation.** Does a sketch of the product of the $(x - x_j)$ in (5) for the data in Example 2 indicate that extrapolation is likely to involve larger errors than interpolation does?
9. **Error function** (35) in App. A3.1. Calculate the Lagrange polynomial $p_2(x)$ for the 5S-values $f(0.25) = 0.27633$, $f(0.5) = 0.52050$, $f(1.0) = 0.84270$ and from $p_2(x)$ an approximation of $f(0.75)$ $(= 0.71116)$.

10. Error bound. Derive an error bound in Prob. 9 from (5).

11. Cubic Lagrange interpolation. Bessel function J_0. Calculate and graph L_0, L_1, L_2, L_3 with $x_0 = 0$, $x_1 = 1, x_2 = 2, x_3 = 3$ on common axes. Find $p_3(x)$ for the data (0, 1), (1, 0.765198), (2, 0.223891), (3, -0.260052) [values of the Bessel function $J_0(x)$]. Find p_3 for $x = 0.5, 1.5, 2.5$ and compare with the 6S-exact values 0.938470, 0.511828, -0.048384.

12. Newton's forward formula (14). Sine integral. Using (14), find $f(1.25)$ by linear, quadratic, and cubic interpolation of the data (values of (40) in App. A31); 6S-value $\text{Si}(1.25) = 1.14645)$ $f(1.0) = 0.94608$, $f(1.5) = 1.32468$, $f(2.0) = 1.60541$, $f(2.5) = 1.77852$, and compute the errors. For the linear interpolation use $f(1.0)$ and $f(1.5)$, for the quadratic $f(1.0)$, $f(1.5)$, $f(2.0)$, etc.

13 Lower degree. Find the degree of the interpolation polynomial for the data $(-4, 50)$, $(-2, 18)$, (0, 2), (2, 2), (4, 18), using a difference table. Find the polynomial.

14. Newton's forward formula (14). Gamma function. Set up (14) for the data in Prob. 3 and compute $\Gamma(1.01)$, $\Gamma(1.03)$, $\Gamma(1.05)$.

15. Divided differences. Obtain p_2 in Example 2 from (10).

16. Divided differences. Error function. Compute $p_2(0.75)$ from the data in Prob. 9 and Newton's divided difference formula (10).

17. Backward difference formula (18). Use $p_2(x)$ in (18) and the values of erf x, $x = 0.2, 0.4, 0.6$ in Table A4 of App. 5, compute erf 0.3 and the error. (4S-exact erf 0.3 = 0.3286).

18. In Example 5 of the text, write down the difference table as needed for (18), then write (18) with general x and then with $x = 0.56$ to verify the answer in Example 5.

19. CAS EXPERIMENT. Adding Terms in Newton Formulas. Write a program for the forward formula (14). Experiment on the increase of accuracy by successively adding terms. As data use values of some function of your choice for which your CAS gives the values needed in determining errors.

20. TEAM PROJECT. Interpolation and Extrapolation. **(a) Lagrange practical error estimate** (after Theorem 1). Apply this to $p_1(9.2)$ and $p_2(9.2)$ for the data $x_0 = 9.0, x_1 = 9.5, x_2 = 11.0, f_0 = \ln x_0, f_1 = \ln x_1$, $f_2 = \ln x_2$ (6S-values).

(b) Extrapolation. Given $(x_j, f(x_j)) = (0.2, 0.9980)$, (0.4, 0.9686), (0.6, 0.8443), (0.8, 0.5358), (1.0, 0). Find $f(0.7)$ from the quadratic interpolation polynomials based on (α) 0.6, 0.8, 1.0, (β) 0.4, 0.6, 0.8, (γ) 0.2, 0.4, 0.6. Compare the errors and comment. [Exact $f(x) = \cos\left(\frac{1}{2}\pi x^2\right)$, $f(0.7) = 0.7181$ (4S).]

(c) Graph the product of factors $(x - x_j)$ in the error formula (5) for $n = 2, \cdots, 10$ separately. What do these graphs show regarding accuracy of interpolation and extrapolation?

21. WRITING PROJECT. Comparison of interpolation methods. List 4–5 ideas that you feel are most important in this section. Arrange them in best logical order. Discuss them in a 2–3 page report.

19.4 Spline Interpolation

Given data (function values, points in the xy-plane) $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ can be interpolated by a polynomial $P_n(x)$ of degree n or less so that the curve of $P_n(x)$ passes through these $n + 1$ points (x_j, f_j); here $f_0 = f(x_0), \cdots, f_n = f(x_n)$, See Sec. 19.3.

Now if n is large, there may be trouble: $P_n(x)$ may tend to oscillate for x between the **nodes** $x_0, \cdots, x_n$. Hence we must be prepared for ***numeric instability*** (Sec. 19.1). Figure 434 shows a famous example by C. Runge[3] for which the maximum error even approaches ∞ as $n \to \infty$ (with the nodes kept equidistant and their number increased). Figure 435 illustrates the increase of the oscillation with n for some other function that is piecewise linear.

Those undesirable oscillations are avoided by the method of splines initiated by I. J. Schoenberg in 1946 (*Quarterly of Applied Mathematics* **4**, pp. 45–99, 112–141). This method is widely used in practice. It also laid the foundation for much of modern **CAD (computer-aided design)**. Its name is borrowed from a *draftman's spline*, which is an elastic rod bent to pass through given points and held in place by weights. The mathematical idea of the method is as follows:

[3]CARL RUNGE (1856–1927), German mathematician, also known for his work on ODEs (Sec. 21.1).

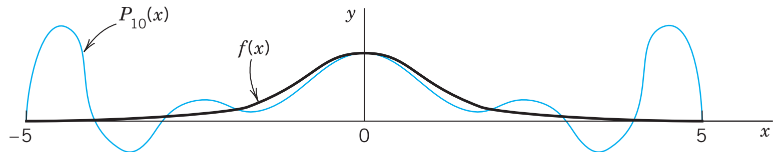


Fig. 434. Runge's example $f(x) = 1/(1 + x^2)$ and interpolating polynomial $P_{10}(x)$

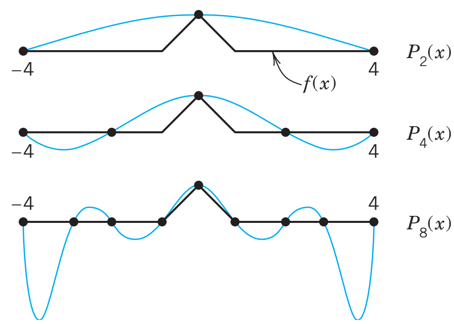


Fig. 435. Piecewise linear function $f(x)$ and interpolation polynomials of increasing degrees

Instead of using a single high-degree polynomial P_n over the entire interval $a \leqq x \leqq b$ in which the nodes lie, that is,

$$a = x_0 < x_1 < \cdots < x_n = b, \tag{1}$$

we use n low-degree, e.g., cubic, polynomials

$$q_0(x), \qquad q_1(x), \qquad \cdots, \qquad q_{n-1}(x),$$

one over each subinterval between adjacent nodes, hence q_0 from x_0 to x_1, then q_1 from x_1 to x_2, and so on. From this we compose an interpolation function $g(x)$, called a **spline**, by fitting these polynomials together into a single continuous curve passing through the data points, that is,

$$g(x_0) = f(x_0) = f_0, \qquad g(x_1) = f(x_1) = f_1, \qquad \cdots, \qquad g(x_n) = f(x_n) = f_n. \tag{2}$$

Note that $g(x) = q_0(x)$ when $x_0 \leqq x \leqq x_1$, then $g(x) = q_1(x)$ when $x_1 \leqq x \leqq x_2$, and so on, according to our construction of g.

Thus **spline interpolation** is *piecewise polynomial interpolation.*

The simplest q_j's would be linear polynomials. However, the curve of a piecewise linear continuous function has corners and would be of little interest in general—think of designing the body of a car or a ship.

We shall consider cubic splines because these are the most important ones in applications. By definition, a **cubic spline** $g(x)$ interpolating given data $(x_0, f_0), \cdots, (x_n, f_n)$ is a continuous function on the interval $a = x_0 \leqq x \leqq x_n = b$ that has continuous first and second derivatives and satisfies the interpolation condition (2); furthermore, between adjacent nodes, $g(x)$ is given by a polynomial $q_j(x)$ of degree 3 or less.

We claim that there is such a cubic spline. And if in addition to (2) we also require that

$$g'(x_0) = k_0, \qquad g'(x_n) = k_n \tag{3}$$

(given tangent directions of $g(x)$ at the two endpoints of the interval $a \leqq x \leqq b$), then we have a uniquely determined cubic spline. This is the content of the following existence and uniqueness theorem, whose proof will also suggest the actual determination of splines. (Condition (3) will be discussed after the proof.)

THEOREM 1

Existence and Uniqueness of Cubic Splines

Let $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with given (arbitrarily spaced) x_j [see (1)] and given $f_j = f(x_j), j = 0, 1, \cdots, n$. Let k_0 and k_n be any given numbers. Then there is one and only one cubic spline $g(x)$ corresponding to (1) *and satisfying* (2) *and* (3).

PROOF By definition, on every subinterval I_j given by $x_j \leqq x \leqq x_{j+1}$, the spline $g(x)$ must agree with a polynomial $q_j(x)$ of degree not exceeding 3 such that

$$q_j(x_j) = f(x_j), \qquad q_j(x_{j+1}) = f(x_{j+1}) \qquad (j = 0, 1, \cdots, n-1). \tag{4}$$

For the derivatives we write

$$q_j'(x_j) = k_j, \qquad q_j'(x_{j+1}) = k_{j+1} \qquad (j = 0, 1, \cdots, n-1) \tag{5}$$

with k_0 and k_n given and $k_1, \cdots, k_{n-1}$ to be determined later. Equations (4) and (5) are four conditions for each $q_j(x)$. By direct calculation, using the notation

$$c_j = \frac{1}{h_j} = \frac{1}{x_{j+1} - x_j} \qquad (j = 0, 1, \cdots, n-1) \tag{6*}$$

we can verify that the unique cubic polynomial $q_j(x)$ $(j = 0, 1, \cdots, n-1)$ satisfying (4) and (5) is

$$\begin{aligned} q_j(x) = {} & f(x_j)c_j^2(x - x_{j+1})^2[1 + 2c_j(x - x_j)] \\ & + f(x_{j+1})c_j^2(x - x_j)^2[1 - 2c_j(x - x_{j+1})] \\ & + k_j c_j^2(x - x_j)(x - x_{j+1})^2 \\ & + k_{j+1} c_j^2(x - x_j)^2(x - x_{j+1}). \end{aligned} \tag{6}$$

Differentiating twice, we obtain

$$q_j''(x_j) = -6c_j^2 f(x_j) + 6c_j^2 f(x_{j+1}) - 4c_j k_j - 2c_j k_{j+1} \tag{7}$$

$$q_j''(x_{j+1}) = 6c_j^2 f(x_j) - 6c_j^2 f(x_{j+1}) + 2c_j k_j + 4c_j k_{j+1}. \tag{8}$$

By definition, $g(x)$ has continuous second derivatives. This gives the conditions

$$q_{j-1}''(x_j) = q_j''(x_j) \qquad (j = 1, \cdots, n-1).$$

If we use (8) with j replaced by $j - 1$, and (7), these $n - 1$ equations become

$$c_{j-1}k_{j-1} + 2(c_{j-1} + c_j)k_j + c_jk_{j+1} = 3[c_{j-1}^2\nabla f_j + c_j^2\nabla f_{j+1}] \tag{9}$$

where $\nabla f_j = f(x_j) - f(x_{j-1})$ and $\nabla f_{j+1} = f(x_{j+1}) - f(x_j)$ and $j = 1, \cdots, n - 1$, as before. This linear system of $n - 1$ equations has a unique solution $k_1, \cdots, k_{n-1}$ since the coefficient matrix is strictly diagonally dominant (that is, in each row the (positive) diagonal entry is greater than the sum of the other (positive) entries). Hence the determinant of the matrix cannot be zero (as follows from Theorem 3 in Sec. 20.7), so that we may determine unique values $k_1, \cdots, k_{n-1}$ of the first derivative of $g(x)$ at the nodes. This proves the theorem. ■

Storage and Time Demands in solving (9) are modest, since the matrix of (9) is **sparse** (has few nonzero entries) and **tridiagonal** (may have nonzero entries only on the diagonal and on the two adjacent "parallels" above and below it). Pivoting (Sec. 7.3) is not necessary because of that dominance. This makes splines efficient in solving large problems with thousands of nodes or more. For some literature and some critical comments, see *American Mathematical Monthly* **105** (1998), 929–941.

Condition (3) includes the **clamped conditions**

$$g'(x_0) = f'(x_0), \qquad g'(x_n) = f'(x_n), \tag{10}$$

in which the tangent directions $f'(x_0)$ and $f'(x_n)$ at the ends are given. Other conditions of practical interest are the **free** or **natural conditions**

$$g''(x_0) = 0, \qquad g''(x_n) = 0 \tag{11}$$

(geometrically: zero curvature at the ends, as for the draftman's spline), giving a **natural spline**. These names are motivated by Fig. 293 in Problem Set 12.3.

Determination of Splines. Let k_0 and k_n be given. Obtain $k_1, \cdots, k_{n-1}$ by solving the linear system (9). Recall that the spline $g(x)$ to be found consists of n cubic polynomials $q_0, \cdots, q_{n-1}$. We write these polynomials in the form

$$q_j(x) = a_{j0} + a_{j1}(x - x_j) + a_{j2}(x - x_j)^2 + a_{j3}(x - x_j)^3 \tag{12}$$

where $j = 0, \cdots, n - 1$. Using Taylor's formula, we obtain

$$\begin{aligned}
a_{j0} &= q_j(x_j) = f_j && \text{by (2),}\\
a_{j1} &= q_j'(x_j) = k_j && \text{by (5),}\\
a_{j2} &= \frac{1}{2}\, q_j''(x_j) = \frac{3}{h_j^2}\,(f_{j+1} - f_j) - \frac{1}{h_j}\,(k_{j+1} + 2k_j) && \text{by (7),}\\
a_{j3} &= \frac{1}{6}\, q_j'''(x_j) = \frac{2}{h_j^3}\,(f_j - f_{j+1}) + \frac{1}{h_j^2}\,(k_{j+1} + k_j)
\end{aligned} \tag{13}$$

with a_{j3} obtained by calculating $q_j''(x_{j+1})$ from (12) and equating the result to (8), that is,

$$q_j''(x_{j+1}) = 2a_{j2} + 6a_{j3}h_j = \frac{6}{h_j^2}\,(f_j - f_{j+1}) + \frac{2}{h_j}\,(k_j + 2k_{j+1}),$$

and now subtracting from this $2a_{j2}$ as given in (13) and simplifying.

Note that for ***equidistant nodes*** of distance $h_j = h$ we can write $c_j = c = 1/h$ in (6*) and have from (9) simply

(14)
$$\boxed{k_{j-1} + 4k_j + k_{j+1} = \frac{3}{h}\,(f_{j+1} - f_{j-1})} \qquad (j = 1, \cdots, n-1).$$

EXAMPLE 1 **Spline Interpolation. Equidistant Nodes**

Interpolate $f(x) = x^4$ on the interval $-1 \leqq x \leqq 1$ by the cubic spline $g(x)$ corresponding to the nodes $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ and satisfying the clamped conditions $g'(-1) = f'(-1)$, $g'(1) = f'(1)$.

Solution. In our standard notation the given data are $f_0 = f(-1) = 1$, $f_1 = f(0) = 0$, $f_2 = f(1) = 1$. We have $h = 1$ and $n = 2$, so that our spline consists of $n = 2$ polynomials

$$q_0(x) = a_{00} + a_{01}(x+1) + a_{02}(x+1)^2 + a_{03}(x+1)^3 \qquad (-1 \leqq x \leqq 0),$$
$$q_1(x) = a_{10} + a_{11}x + a_{12}x^2 + a_{13}x^3 \qquad (0 \leqq x \leqq 1).$$

We determine the k_j from (14) (equidistance!) and then the coefficients of the spline from (13). Since $n = 2$, the system (14) is a single equation (with $j = 1$ and $h = 1$)

$$k_0 + 4k_1 + k_2 = 3(f_2 - f_0).$$

Here $f_0 = f_2 = 1$ (the value of x^4 at the ends) and $k_0 = -4$, $k_2 = 4$, the values of the derivative $4x^3$ at the ends -1 and 1. Hence

$$-4 + 4k_1 + 4 = 3(1 - 1) = 0, \qquad k_1 = 0.$$

From (13) we can now obtain the coefficients of q_0, namely, $a_{00} = f_0 = 1$, $a_{01} = k_0 = -4$, and

$$a_{02} = \frac{3}{1^2}(f_1 - f_0) - \frac{1}{1}(k_1 + 2k_0) = 3(0-1) - (0-8) = 5$$

$$a_{03} = \frac{2}{1^3}(f_0 - f_1) + \frac{1}{1^2}(k_1 + k_0) = 2(1-0) + (0-4) = -2.$$

Similarly, for the coefficients of q_1 we obtain from (13) the values $a_{10} = f_1 = 0$, $a_{11} = k_1 = 0$, and

$$a_{12} = 3(f_2 - f_1) - (k_2 + 2k_1) = 3(1-0) - (4+0) = -1$$
$$a_{13} = 2(f_1 - f_2) + (k_2 + k_1) = 2(0-1) + (4+0) = 2.$$

This gives the polynomials of which the spline $g(x)$ consists, namely,

$$g(x) = \begin{cases} q_0(x) = 1 - 4(x+1) + 5(x+1)^2 - 2(x+1)^3 = -x^2 - 2x^3 & \text{if} \quad -1 \leqq x \leqq 0 \\ q_1(x) = -x^2 + 2x^3 & \text{if} \quad 0 \leqq x \leqq 1. \end{cases}$$

Figure 436 shows $f(x)$ and this spline. Do you see that we could have saved over half of our work by using symmetry? ■

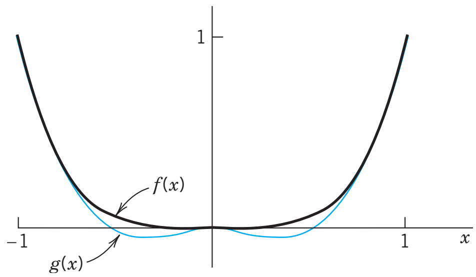


Fig. 436. Function $f(x) = x^4$ and cubic spline $g(x)$ in Example 1

EXAMPLE 2 Natural Spline. Arbitrarily Spaced Nodes

Find a spline approximation and a polynomial approximation for the curve of the cross section of the circular-shaped Shrine of the Book in Jerusalem shown in Fig. 437.

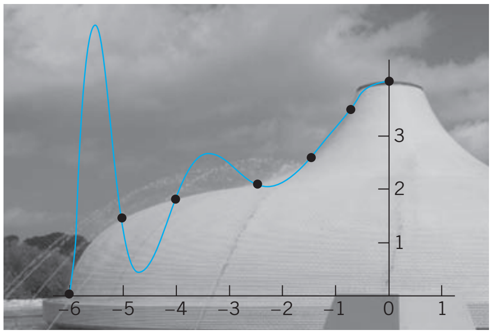


Fig. 437. Shrine of the Book in Jerusalem (Architects F. Kissler and A. M. Bartus)

Solution. Thirteen points, about equally distributed along the contour (not along the x-axis!), give these data:

x_j	−5.8	−5.0	−4.0	−2.5	−1.5	−0.8	0	0.8	1.5	2.5	4.0	5.0	5.8
f_j	0	1.5	1.8	2.2	2.7	3.5	3.9	3.5	2.7	2.2	1.8	1.5	0

The figure shows the corresponding interpolation polynomial of 12th degree, which is useless because of its oscillation. (Because of roundoff your software will also give you small error terms involving odd powers of x.) The polynomial is

$$P_{12}(x) = 3.9000 - 0.65083x^2 + 0.033858x^4 + 0.011041x^6 - 0.0014010x^8$$
$$+ 0.000055595x^{10} - 0.00000071867x^{12}.$$

The spline follows practically the contour of the roof, with a small error near the nodes −0.8 and 0.8. The spline is symmetric. Its six polynomials corresponding to positive x have the following coefficients of their representations (12). (Note well that (12) is in terms of powers of $x - x_j$, not x!)

j	x-interval	a_{j0}	a_{j1}	a_{j2}	a_{j3}
0	0.0...0.8	3.9	0.00	−0.61	−0.015
1	0.8...1.5	3.5	−1.01	−0.65	0.66
2	1.5...2.5	2.7	−0.95	0.73	−0.27
3	2.5...4.0	2.2	−0.32	−0.091	0.084
4	4.0...5.0	1.8	−0.027	0.29	−0.56
5	5.0...5.8	1.5	−1.13	−1.39	0.58

PROBLEM SET 19.4

1. WRITING PROJECT. Splines. In your own words, and using as few formulas as possible, write a short report on spline interpolation, its motivation, a comparison with polynomial interpolation, and its applications.

2–9 VERIFICATIONS. DERIVATIONS. COMPARISONS

2. Individual polynomial q_j. Show that $q_j(x)$ in (6) satisfies the interpolation condition (4) as well as the derivative condition (5).

3. Verify the differentiations that give (7) and (8) from (6).

4. System for derivatives. Derive the basic linear system (9) for $k_1, \cdots, k_{n-1}$ as indicated in the text.

5. Equidistant nodes. Derive (14) from (9).

6. Coefficients. Give the details of the derivation of a_{j2} and a_{j3} in (13).

7. Verify the computations in Example 1.

8. Comparison. Compare the spline g in Example 1 with the quadratic interpolation polynomial over the whole interval. Find the maximum deviations of g and p_2 from f. Comment.

9. Natural spline condition. Using the given coefficients, verify that the spline in Example 2 satisfies $g''(x) = 0$ at the ends.

10–16 DETERMINATION OF SPLINES

Find the cubic spline $g(x)$ for the given data with k_0 and k_n as given.

10. $f(-2) = f(-1) = f(1) = f(2) = 0, \quad f(0) = 1,$ $k_0 = k_4 = 0$

11. If we started from the piecewise linear function in Fig. 438, we would obtain $g(x)$ in Prob. 10 as the spline satisfying $g'(-2) = f'(-2) = 0, \quad g'(2) = f'(2) = 0.$ Find and sketch or graph the corresponding interpolation polynomial of 4th degree and compare it with the spline. Comment.

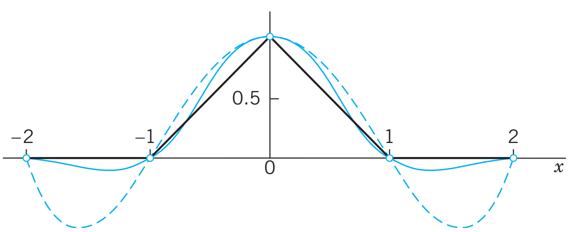


Fig. 438. Spline and interpolation polynomial in Probs. 10 and 11

12. $f_0 = f(0) = 1, \quad f_1 = f(2) = 9, \quad f_2 = f(4) = 41,$ $f_3 = f(6) = 41, \quad k_0 = 0, \quad k_3 = -12$

13. $f_0 = f(0) = 1, \quad f_1 = f(1) = 0, \quad f_2 = f(2) = -1,$ $f_3 = f(3) = 0, \quad k_0 = 0, \quad k_3 = -6$

14. $f_0 = f(0) = 2, \quad f_1 = f(1) = 3, \quad f_2 = f(2) = 8,$ $f_3 = f(3) = 12, \quad k_0 = k_3 = 0$

15. $f_0 = f(0) = 4, \quad f_1 = f(2) = 0, \quad f_2 = f(4) = 4,$ $f_3 = f(6) = 80, \quad k_0 = k_3 = 0$

16. $f_0 = f(0) = 2, \quad f_1 = f(2) = -2, \quad f_2 = f(4) = 2,$ $f_3 = f(6) = 78, \quad k_0 = k_3 = 0.$ Can you obtain the answer from that of Prob. 15?

17. If a cubic spline is three times continuously differentiable (that is, it has continuous first, second, and third derivatives), show that it must be a single polynomial.

18. CAS EXPERIMENT. Spline versus Polynomial. If your CAS gives natural splines, find the natural splines when x is integer from $-m$ to m, and $y(0) = 1$ and all other y equal to 0. Graph each such spline along with the interpolation polynomial p_{2m}. Do this for $m = 2$ to 10 (or more). What happens with increasing m?

19. Natural conditions. Explain the remark after (11).

20. TEAM PROJECT. Hermite Interpolation and Bezier Curves. In **Hermite interpolation** we are looking for a polynomial $p(x)$ (of degree $2n + 1$ or less) such that $p(x)$ and its derivative $p'(x)$ have given values at $n + 1$ nodes. (More generally, $p(x), p'(x), p''(x), \cdots$ may be required to have given values at the nodes.)

(a) Curves with given endpoints and tangents. Let C be a curve in the xy-plane parametrically represented by $\mathbf{r}(t) = [x(t), y(t)], 0 \leqq t \leqq 1$ (see Sec. 9.5). Show that for given initial and terminal points of a curve and given initial and terminal tangents, say,

$$
\begin{aligned}
A\colon \quad \mathbf{r}_0 &= [x(0), y(0)] \\
&= [x_0, y_0], \\
B\colon \quad \mathbf{r}_1 &= [x(1), y(1)] \\
&= [x_1, y_1] \\
\mathbf{v}_0 &= [x'(0), y'(0)] \\
&= [x_0', y_0'], \\
\mathbf{v}_1 &= [x'(1), y'(1)] \\
&= [x_1', y_1']
\end{aligned}
$$

we can find a curve C, namely,

$$
\begin{aligned}
\mathbf{r}(t) = \mathbf{r}_0 &+ \mathbf{v}_0 t \\
&+ (3(\mathbf{r}_1 - \mathbf{r}_0) - (2\mathbf{v}_0 + \mathbf{v}_1))t^2 \qquad (15) \\
&+ (2(\mathbf{r}_0 - \mathbf{r}_1) + \mathbf{v}_0 + \mathbf{v}_1)t^3;
\end{aligned}
$$

in components,

$$x(t) = x_0 + x_0't + (3(x_1 - x_0) - (2x_0' + x_1'))t^2$$
$$+ (2(x_0 - x_1) + x_0' + x_1')t^3$$

$$y(t) = y_0 + y_0't + (3(y_1 - y_0) - (2y_0' + y_1'))t^2$$
$$+ (2(y_0 - y_1) + y_0' + y_1')t^3.$$

Note that this is a cubic Hermite interpolation polynomial, and $n = 1$ because we have two nodes (the endpoints of C). (This has nothing to do with the Hermite polynomials in Sec. 5.8.) The two points

$$G_A\text{:}\ \mathbf{g}_0 = \mathbf{r}_0 + \mathbf{v}_0 = [x_0 + x_0',\ y_0 + y_0']$$

and

$$G_B\text{:}\ \mathbf{g}_1 = \mathbf{r}_1 - \mathbf{v}_1 = [x_1 - x_1',\ y_1 - y_1']$$

are called **guidepoints** because the segments AG_A and BG_B specify the tangents graphically. A, B, G_A, G_B determine C, and C can be changed quickly by moving the points. A curve consisting of such Hermite interpolation polynomials is called a **Bezier curve**, after the French engineer P. Bezier of the Renault Automobile Company, who introduced them in the early 1960s in designing car bodies. Bezier curves (and surfaces) are used in computer-aided design (CAD) and computer-aided manufacturing (CAM). (For more details, see Ref. [E21] in App. 1.)

(b) Find and graph the Bezier curve and its guidepoints if A: [0, 0], B: [1, 0], $\mathbf{v}_0 = [\frac{1}{2}, \frac{1}{2}]$, $\mathbf{v}_1 = [-\frac{1}{2}, -\frac{1}{4}\sqrt{3}]$.

(c) Changing guidepoints changes C. Moving guidepoints farther away results in C "staying near the tangents for a longer time." Confirm this by changing $\mathbf{v}_0$ and $\mathbf{v}_1$ in (b) to $2\mathbf{v}_0$ and $2\mathbf{v}_1$ (see Fig. 439).

(d) Make experiments of your own. What happens if you change $\mathbf{v}_1$ in (b) to $-\mathbf{v}_1$. If you rotate the tangents? If you multiply $\mathbf{v}_0$ and $\mathbf{v}_1$ by positive factors less than 1?

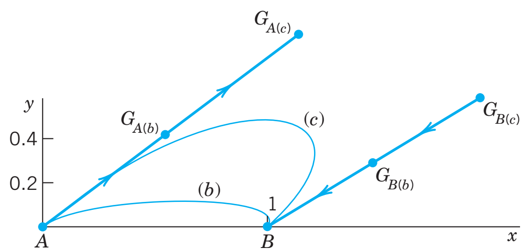


Fig. 439. Team Project 20(b) and (c): Bezier curves

19.5 Numeric Integration and Differentiation

In applications, the engineer often encounters integrals that are very difficult or even impossible to solve analytically. For example, the error function, the Fresnel integrals (see Probs. 16–25 on nonelementary integrals in this section), and others cannot be evaluated by the usual methods of calculus (see App. 3, (24)–(44) for such "difficult" integrals). We then need methods from numerical analysis to evaluate such integrals. We also need numerics when the integrand of the integral to be evaluated consists of an empirical function, where we are given some recorded values of that function. Methods that address these kinds of problems are called methods of numeric integration.

Numeric integration means the numeric evaluation of integrals

$$J = \int_a^b f(x)\,dx$$

where a and b are given and f is a function given analytically by a formula or empirically by a table of values. Geometrically, J is the area under the curve of f between a and b (Fig. 440), taken with a minus sign where f is negative.

We know that if f is such that we can find a differentiable function F whose derivative is f, then we can evaluate J directly, i.e., without resorting to numeric integration, by applying the familiar formula

$$J = \int_a^b f(x)\,dx = F(b) - F(a) \qquad [F'(x) = f(x)].$$

Your CAS (Mathematica, Maple, etc.) or tables of integrals may be helpful for this purpose.

Rectangular Rule. Trapezoidal Rule

Numeric integration methods are obtained by approximating the integrand f by functions that can easily be integrated.

The simplest formula, the **rectangular rule**, is obtained if we subdivide the interval of integration $a \leqq x \leqq b$ into n subintervals of equal length $h = (b - a)/n$ and in each subinterval approximate f by the constant $f(x_j^*)$, the value of f at the midpoint x_j^* of the jth subinterval (Fig. 441). Then f is approximated by a **step function** (piecewise constant function), the n rectangles in Fig. 441 have the areas $f(x_1^*)h, \cdots, f(x_n^*)h$, and the **rectangular rule** is

(1)
$$J = \int_a^b f(x)\,dx \approx h[f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)] \qquad \left(h = \frac{b-a}{n}\right).$$

The **trapezoidal rule** is generally more accurate. We obtain it if we take the same subdivision as before and approximate f by a broken line of segments (chords) with endpoints $[a, f(a)], [x_1, f(x_1)], \cdots, [b, f(b)]$ on the curve of f (Fig. 442). Then the area under the curve of f between a and b is approximated by n trapezoids of areas

$$\tfrac{1}{2}[f(a) + f(x_1)]h, \qquad \tfrac{1}{2}[f(x_1) + f(x_2)]h, \qquad \cdots, \qquad \tfrac{1}{2}[f(x_{n-1}) + f(b)]h.$$

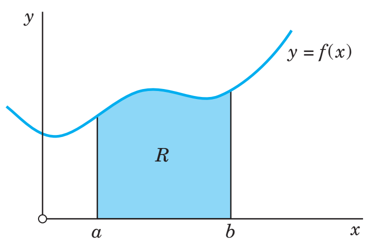


Fig. 440. Geometric interpretation of a definite integral

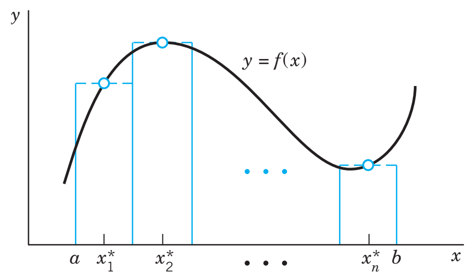


Fig. 441. Rectangular rule

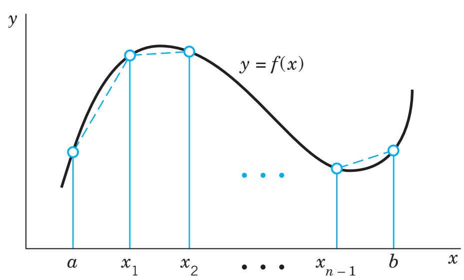


Fig. 442. Trapezoidal rule

By taking their sum we obtain the **trapezoidal rule**

$$J = \int_a^b f(x)\,dx \approx h\left[\frac{1}{2}f(a) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(b)\right] \tag{2}$$

where $h = (b - a)/n$, as in (1). The x_j's and a and b are called **nodes**.

EXAMPLE 1 Trapezoidal Rule

Evaluate $J = \int_0^1 e^{-x^2}\,dx$ by means of (2) with $n = 10$.

Note that this integral cannot be evaluated by elementary calculus, but leads to the error function (see Eq. (35), App. 3).

Solution. $J \approx 0.1(0.5 \cdot 1.367879 + 6.778167) = 0.746211$ from Table 19.3. ■

Table 19.3 Computations in Example 1

j	x_j	x_j^2	$e^{-x_j^2}$	
0	0	0	1.000000	
1	0.1	0.01		0.990050
2	0.2	0.04		0.960789
3	0.3	0.09		0.913931
4	0.4	0.16		0.852144
5	0.5	0.25		0.778801
6	0.6	0.36		0.697676
7	0.7	0.49		0.612626
8	0.8	0.64		0.527292
9	0.9	0.81		0.444858
10	1.0	1.00	0.367879	
Sums			1.367879	6.778167

Error Bounds and Estimate for the Trapezoidal Rule

An error estimate for the trapezoidal rule can be derived from (5) in Sec. 19.3 with $n = 1$ by integration as follows. For a single subinterval we have

$$f(x) - p_1(x) = (x - x_0)(x - x_1)\,\frac{f''(t)}{2}$$

with a suitable t depending on x, between x_0 and x_1. Integration over x from $a = x_0$ to $x_1 = x_0 + h$ gives

$$\int_{x_0}^{x_0+h} f(x)\,dx - \frac{h}{2}\,[f(x_0) + f(x_1)] = \int_{x_0}^{x_0+h} (x - x_0)(x - x_0 - h)\,\frac{f''(t(x))}{2}\,dx.$$

Setting $x - x_0 = v$ and applying the mean value theorem of integral calculus, which we can use because $(x - x_0)(x - x_0 - h)$ does not change sign, we find that the right side equals

$$(3^*) \qquad \int_0^h v(v-h)\,dv\,\frac{f''(\tilde{t})}{2} = \left(\frac{h^3}{3} - \frac{h^3}{2}\right)\frac{f''(\tilde{t})}{2} = -\frac{h^3}{12}\,f''(\tilde{t})$$

where $\tilde{t}$ is a (suitable, unknown) value between x_0 and x_1. This is the error for the trapezoidal rule with $n = 1$, often called the **local error**.

Hence the **error** ϵ of (2) with any n is the sum of such contributions from the n subintervals; since $h = (b - a)/n$, $nh^3 = n(b - a)^3/n^3$, and $(b - a)^2 = n^2h^2$, we obtain

$$\textbf{(3)} \qquad \epsilon = -\frac{(b-a)^3}{12n^2}\,f''(\hat{t}) = -\frac{b-a}{12}\,h^2 f''(\hat{t})$$

with (suitable, unknown) $\hat{t}$ between a and b.

Because of (3) the trapezoidal rule (2) is also written

$$(2^*) \quad J = \int_a^b f(x)\,dx \approx h\left[\frac{1}{2}f(a) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(b)\right] - \frac{b-a}{12}\,h^2 f''(\hat{t}).$$

Error Bounds are now obtained by taking the largest value for f'', say, M_2, and the smallest value, M_2^*, in the interval of integration. Then (3) gives (note that K is negative)

$$\textbf{(4)} \qquad KM_2 \leqq \epsilon \leqq KM_2^* \qquad \text{where} \qquad K = -\frac{(b-a)^3}{12n^2} = -\frac{b-a}{12}\,h^2.$$

Error Estimation by Halving $\boldsymbol{h}$ is advisable if f'' is very complicated or unknown, for instance, in the case of experimental data. Then we may apply the Error Principle of Sec. 19.1. That is, we calculate by (2), first with h, obtaining, say, $J = J_h + \epsilon_h$, and then with $\frac{1}{2}h$, obtaining $J = J_{h/2} + \epsilon_{h/2}$. Now if we replace h^2 in (3) with $(\frac{1}{2}h)^2$, the error is multiplied by $\frac{1}{4}$. Hence $\epsilon_{h/2} \approx \frac{1}{4}\epsilon_h$ (not exactly because $\hat{t}$ may differ). Together, $J_{h/2} + \epsilon_{h/2} = J_h + \epsilon_h \approx J_h + 4\epsilon_{h/2}$. Thus $J_{h/2} - J_h = (4 - 1)\epsilon_{h/2}$. Division by 3 gives the error formula for $J_{h/2}$

$$\textbf{(5)} \qquad \epsilon_{h/2} \approx \tfrac{1}{3}\,(J_{h/2} - J_h).$$

EXAMPLE 2 Error Estimation for the Trapezoidal Rule by (4) and (5)

Estimate the error of the approximate value in Example 1 by (4) and (5).

Solution. **(A)** ***Error bounds by*** **(4).** By differentiation, $f''(x) = 2(2x^2 - 1)e^{-x^2}$. Also, $f'''(x) > 0$ if $0 < x < 1$, so that the minimum and maximum occur at the ends of the interval. We compute $M_2 = f''(1) = 0.735759$ and $M_2^* = f''(0) = -2$. Furthermore, $K = -1/1200$, and (4) gives

$$-0.000614 \leqq \epsilon \leqq 0.001667.$$

Hence the exact value of J must lie between

$$0.746211 - 0.000614 = 0.745597 \qquad \text{and} \qquad 0.746211 + 0.001667 = 0.747878.$$

Actually, $J = 0.746824$, exact to 6D.

(B) ***Error estimate by*** **(5).** $J_h = 0.746211$ in Example 1. Also,

$$J_{h/2} = 0.05\left[\sum_{j=1}^{19} e^{-(j/20)^2} + \frac{1}{2}(1 + 0.367879)\right] = 0.746671.$$

Hence $\epsilon_{h/2} = \frac{1}{3}(J_{h/2} - J_h) = 0.000153$ and $J_{h/2} + \epsilon_{h/2} = 0.746824$, exact to 6D.

Simpson's Rule of Integration

Piecewise constant approximation of f led to the rectangular rule (1), piecewise linear approximation to the trapezoidal rule (2), and piecewise quadratic approximation will lead to Simpson's rule, which is of great practical importance because it is sufficiently accurate for most problems, but still sufficiently simple.

To derive Simpson's rule, we divide the interval of integration $a \leqq x \leqq b$ into an ***even number*** of equal subintervals, say, into $n = 2m$ subintervals of length $h = (b - a)/(2m)$, with endpoints $x_0 (= a), x_1, \cdots, x_{2m-1}, x_{2m} (= b)$; see Fig. 443. We now take the first two subintervals and approximate $f(x)$ in the interval $x_0 \leqq x \leqq x_2 = x_0 + 2h$ by the Lagrange polynomial $p_2(x)$ through $(x_0, f_0), (x_1, f_1), (x_2, f_2)$, where $f_j = f(x_j)$. From (3) in Sec. 19.3 we obtain

$$(6) \quad p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2.$$

The denominators in (6) are $2h^2$, $-h^2$, and $2h^2$, respectively. Setting $s = (x - x_1)/h$, we have

$$x - x_1 = sh, \qquad x - x_0 = x - (x_1 - h) = (s + 1)h$$
$$x - x_2 = x - (x_1 + h) = (s - 1)h$$

and we obtain

$$p_2(x) = \tfrac{1}{2}s(s - 1)f_0 - (s + 1)(s - 1)f_1 + \tfrac{1}{2}(s + 1)sf_2.$$

We now integrate with respect to x from x_0 to x_2. This corresponds to integrating with respect to s from -1 to 1. Since $dx = h\,ds$, the result is

$$(7^*) \qquad \int_{x_0}^{x_2} f(x)\,dx \approx \int_{x_0}^{x_2} p_2(x)\,dx = h\left(\frac{1}{3}f_0 + \frac{4}{3}f_1 + \frac{1}{3}f_2\right).$$

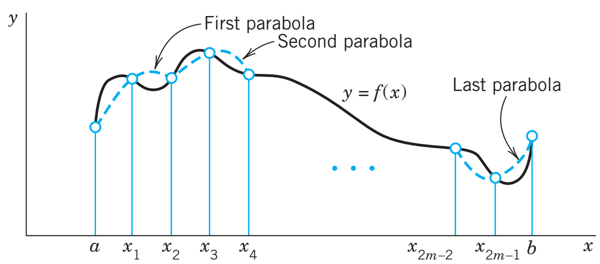


Fig. 443. Simpson's rule

A similar formula holds for the next two subintervals from x_2 to x_4, and so on. By summing all these m formulas we obtain **Simpson's rule**[4]

$$\text{(7)} \qquad \int_a^b f(x)\,dx \approx \frac{h}{3}\,(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}),$$

where $h = (b - a)/(2m)$ and $f_j = f(x_j)$. Table 19.4 shows an algorithm for Simpson's rule.

Table 19.4 Simpson's Rule of Integration

ALGORITHM SIMPSON $(a, b, m, f_0, f_1, \cdots, f_{2m})$

This algorithm computes the integral $J = \int_a^b f(x)\,dx$ from given values $f_j = f(x_j)$ at equidistant $x_0 = a,\ x_1 = x_0 + h, \cdots, x_{2m} = x_0 + 2mh = b$ by Simpson's rule (7), where $h = (b - a)/(2m)$.

INPUT: $a, b, m, f_0, \cdots, f_{2m}$

OUTPUT: Approximate value $\tilde{J}$ of J

Compute $s_0 = f_0 + f_{2m}$

$s_1 = f_1 + f_3 + \cdots + f_{2m-1}$

$s_2 = f_2 + f_4 + \cdots + f_{2m-2}$

$h = (b - a)/2m$

$\tilde{J} = \dfrac{h}{3}\,(s_0 + 4s_1 + 2s_2)$

OUTPUT $\tilde{J}$. Stop.

End SIMPSON

Error of Simpson's Rule (7). If the fourth derivative $f^{(4)}$ exists and is continuous on $a \leqq x \leqq b$, the **error** of (7), call it ϵ_S, is

$$\text{(8)} \qquad \epsilon_S = -\frac{(b - a)^5}{180\,(2m)^4}\, f^{(4)}(\hat{t}) = -\frac{b - a}{180}\, h^4 f^{(4)}(\hat{t});$$

here $\hat{t}$ is a suitable unknown value between a and b. This is obtained similarly to (3). With this we may also write Simpson's rule (7) as

$$\text{(7**)} \qquad \int_a^b f(x)\,dx = \frac{h}{3}\,(f_0 + 4f_1 + \cdots + f_{2m}) - \frac{b - a}{180}\, h^4 f^{(4)}(\hat{t}).$$

[4]THOMAS SIMPSON (1710–1761), self-taught English mathematician, author of several popular textbooks. Simpson's rule was used much earlier by Torricelli, Gregory (in 1668), and Newton (in 1676).

Error Bounds. By taking for $f^{(4)}$ in (8) the maximum M_4 and minimum M_4^* on the interval of integration we obtain from (8) the error bounds (note that C is negative)

$$\textbf{(9)} \qquad CM_4 \leqq \epsilon_S \leqq CM_4^* \qquad \text{where} \qquad C = -\frac{(b-a)^5}{180(2m)^4} = -\frac{b-a}{180}h^4.$$

Degree of Precision (DP) *of an integration formula.* This is the maximum degree of arbitrary polynomials for which the formula gives exact values of integrals over any intervals.

Hence for the trapezoidal rule,

$$\text{DP} = 1$$

because we approximate the curve of f by portions of straight lines (linear polynomials).

For Simpson's rule we might expect DP = 2 (why?). Actually,

$$\text{DP} = 3$$

by (9) because $f^{(4)}$ is identically zero for a cubic polynomial. This makes Simpson's rule sufficiently accurate for most practical problems and accounts for its popularity.

Numeric Stability *with respect to rounding* is another important property of Simpson's rule. Indeed, for the sum of the roundoff errors ϵ_j of the $2m + 1$ values f_j in (7) we obtain, since $h = (b - a)/2m$,

$$\frac{h}{3}\left|\epsilon_0 + 4\epsilon_1 + \cdots + \epsilon_{2m}\right| \leqq \frac{b-a}{3.2m}\, 6mu = (b-a)u$$

where u is the rounding unit ($u = \frac{1}{2} \cdot 10^{-6}$ if we round off to 6D; see Sec. 19.1). Also $6 = 1 + 4 + 1$ is the sum of the coefficients for a pair of intervals in (7); take $m = 1$ in (7) to see this. The bound $(b - a)u$ is independent of m, so that it cannot increase with increasing m, that is, with decreasing h. This proves stability. ■

Newton–Cotes Formulas. We mention that the trapezoidal and Simpson rules are special *closed Newton–Cotes formulas*, that is, integration formulas in which $f(x)$ is interpolated at equally spaced nodes by a polynomial of degree n ($n = 1$ for trapezoidal, $n = 2$ for Simpson), and **closed** means that a and b are nodes ($a = x_0$, $b = x_n$). $n = 3$ and higher n are used occasionally. From $n = 8$ on, some of the coefficients become negative, so that a positive f_j could make a negative contribution to an integral, which is absurd. For more on this topic see Ref. [E25] in App. 1.

EXAMPLE 3 Simpson's Rule. Error Estimate

Evaluate $J = \int_0^1 e^{-x^2} dx$ by Simpson's rule with $2m = 10$ and estimate the error.

Solution. Since $h = 0.1$, Table 19.5 gives

$$J \approx \frac{0.1}{3}(1.367879 + 4 \cdot 3.740266 + 2 \cdot 3.037901) = 0.746825.$$

Estimate of error. Differentiation gives $f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2}$. By considering the derivative $f^{(5)}$ of $f^{(4)}$ we find that the largest value of $f^{(4)}$ in the interval of integration occurs at 0 and the smallest value at $x^* = (2.5 - 0.5\sqrt{10})^{1/2}$. Computation gives the values $M_4 = f^{(4)}(0) = 12$ and $M_4^* = f^{(4)}(x^*) = -7.419$. Since $2m = 10$ and $b - a = 1$, we obtain $C = -1/1800000 = -0.00000056$. Therefore, from (9),

$$-0.000007 \leqq \epsilon_S \leqq 0.000005.$$

Hence J must lie between $0.746825 - 0.000007 = 0.746818$ and $0.746825 + 0.000005 = 0.746830$, so that at least four digits of our approximate value are exact. Actually, the value 0.746825 is exact to 5D because $J = 0.746824$ (exact to 6D).

Thus our result is much better than that in Example 1 obtained by the trapezoidal rule, whereas the number of operations is nearly the same in both cases.

Table 19.5 Computations in Example 3

j	x_j	x_j^2	$e^{-x_j^2}$		
0	0	0	1.000000		
1	0.1	0.01		0.990050	
2	0.2	0.04			0.960789
3	0.3	0.09		0.913931	
4	0.4	0.16			0.852144
5	0.5	0.25		0.778801	
6	0.6	0.36			0.697676
7	0.7	0.49		0.612626	
8	0.8	0.64			0.527292
9	0.9	0.81		0.444858	
10	1.0	1.00	0.367879		
Sums			1.367879	3.740266	3.037901

Instead of picking an $n = 2m$ and then estimating the error by (9), as in Example 3, it is better to require an accuracy (e.g., 6D) and then determine $n = 2m$ from (9).

EXAMPLE 4 Determination of $n = 2m$ in Simpson's Rule from the Required Accuracy

What n should we choose in Example 3 to get 6D-accuracy?

Solution. Using $M_4 = 12$ (which is bigger in absolute value than M_4^*, we get from (9), with $b - a = 1$ and the required accuracy,

$$|CM_4| = \frac{12}{180(2m)^4} = \frac{1}{2} \cdot 10^{-6}, \qquad \text{thus} \qquad m = \left[\frac{2 \cdot 10^6 \cdot 12}{180 \cdot 2^4}\right]^{1/4} = 9.55.$$

Hence we should choose $n = 2m = 20$. Do the computation, which parallels that in Example 3.

Note that the error bounds in (4) or (9) may sometimes be loose, so that in such a case a smaller $n = 2m$ may already suffice.

Error Estimation for Simpson's Rule by Halving h. The idea is the same as in (5) and gives

(10)
$$\epsilon_{h/2} \approx \tfrac{1}{15}(J_{h/2} - J_h).$$

J_h is obtained by using h and $J_{h/2}$ by using $\frac{1}{2}h$, and $\epsilon_{h/2}$ is the error of $J_{h/2}$.

Derivation. In (5) we had $\frac{1}{3}$ as the reciprocal of $3 = 4 - 1$ and $\frac{1}{4} = (\frac{1}{2})^2$ resulted from h^2 in (3) by replacing h with $\frac{1}{2}h$. In (10) we have $\frac{1}{15}$ as the reciprocal of $15 = 16 - 1$ and $\frac{1}{16} = (\frac{1}{2})^4$ results from h^4 in (8) by replacing h with $\frac{1}{2}h$.

EXAMPLE 5 Error Estimation for Simpson's Rule by Halving

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from 0 to 2 with $h = 1$ and apply (10).

Solution. The exact 5D-value of the integral is $J = 1.25953$. Simpson's rule gives

$$J_h = \tfrac{1}{3}[f(0) + 4f(1) + f(2)] = \tfrac{1}{3}(0 + 4 \cdot 0.555360 + 0) = 0.740480,$$

$$J_{h/2} = \tfrac{1}{6}\,[f(0) + 4f(\tfrac{1}{2}) + 2f(1) + 4f(\tfrac{3}{2}) + f(2)]$$

$$= \tfrac{1}{6}[0 + 4 \cdot 0.045351 + 2 \cdot 0.555361 + 4 \cdot 1.521579 + 0] = 1.22974.$$

Hence (10) gives $\epsilon_{h/2} = \frac{1}{15}(1.22974 - 0.74048) = 0.032617$ and thus $J \approx J_{h/2} + \epsilon_{h/2} = 1.26236$, with an error -0.00283 which is less in absolute value than $\frac{1}{10}$ of the error 0.02979 of $J_{h/2.}$ Hence the use of (10) was well worthwhile. ■

Adaptive Integration

The idea is to adapt step h to the variability of $f(x)$. That is, where f varies but little, we can proceed in large steps without causing a substantial error in the integral, but where f varies rapidly, we have to take small steps in order to stay everywhere close enough to the curve of f.

Changing h is done systematically, usually by halving h, and automatically (not "by hand") depending on the size of the (estimated) error over a subinterval. The subinterval is halved if the corresponding error is still too large, that is, larger than a given **tolerance** TOL (maximum admissible absolute error), or is not halved if the error is less than or equal to TOL (or doubled if the error is very small).

Adapting is one of the techniques typical of modern software. In connection with integration it can be applied to various methods. We explain it here for Simpson's rule. In Table 19.6 an asterisk means that for that subinterval, TOL has been reached.

EXAMPLE 6 Adaptive Integration with Simpson's Rule

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 by adaptive integration and with Simpson's rule and TOL[0, 2] = 0.0002.

Solution. Table 19.6 shows the calculations. Figure 444 shows the integrand $f(x)$ and the adapted intervals used. The first two intervals ([0, 0.5], [0.5, 1.0]) have length 0.5, hence $h = 0.25$ [because we use $2m = 2$ subintervals in Simpson's rule (7**)]. The next two intervals ([1.00, 1.25], [1.25, 1.50]) have length 0.25 (hence $h = 0.125$) and the last four intervals have length 0.125. *Sample computations.* For 0.740480 see Example 5. Formula (10) gives $(0.123716 - 0.122794)/15 = 0.000061$. Note that 0.123716 refers to [0, 0.5] and [0.5, 1], so that we must subtract the value corresponding to [0, 1] in the line before. Etc. TOL[0, 2] = 0.0002 gives 0.0001 for subintervals of length 1, 0.00005 for length 0.5, etc. The value of the integral obtained is the sum of the values marked by an asterisk (for which the error estimate has become less than TOL). This gives

$$J \approx 0.123716 + 0.528895 + 0.388263 + 0.218483 = 1.25936.$$

The exact 5D-value is $J = 1.25953$. Hence the error is 0.00017. This is about 1/200 of the absolute value of that in Example 5. Our more extensive computation has produced a much better result. ■

Table 19.6 Computations in Example 6

Interval	Integral	Error (10)	TOL	Comment
[0, 2]	0.740480		0.0002	
[0, 1] [1, 2]	0.122794 1.10695 Sum = 1.22974	0.032617	0.0002	Divide further
[0.0, 0.5] [0.5, 1.0]	0.004782 0.118934 Sum = 0.123716*	0.000061	0.0001	TOL reached
[1.0, 1.5] [1.5, 2.0]	0.528176 0.605821 Sum = 1.13300	0.001803	0.0001	Divide further
[1.00, 1.25] [1.25, 1.50]	0.200544 0.328351 Sum = 0.528895*	0.000048	0.00005	TOL reached
[1.50, 1.75] [1.75, 2.00]	0.388235 0.218457 Sum = 0.606692	0.000058	0.00005	Divide further
[1.500, 1.625] [1.625, 1.750]	0.196244 0.192019 Sum = 0.388263*	0.000002	0.000025	TOL reached
[1.750, 1.875] [1.875, 2.000]	0.153405 0.065078 Sum = 0.218483*	0.000002	0.000025	TOL reached

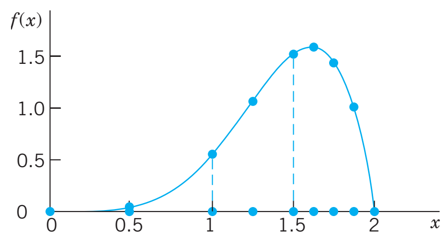


Fig. 444. Adaptive integration in Example 6

Gauss Integration Formulas Maximum Degree of Precision

Our integration formulas discussed so far use function values at *predetermined* (equidistant) x-values (nodes) and give exact results for polynomials not exceeding a certain degree [called the *degree of precision*; see after (9)]. But we can get much more accurate integration formulas as follows. We set

$$\int_{-1}^{1} f(t)\,dt \approx \sum_{j=1}^{n} A_j f_j \qquad [f_j = f(t_j)] \tag{11}$$

with fixed n, and $t = \pm 1$ obtained from $x = a, b$ by setting $x = \frac{1}{2}[a(t-1) + b(t+1)]$. Then we determine the n coefficients $A_1, \cdots, A_n$ and n nodes $t_1, \cdots, t_n$ so that (11) gives exact results for polynomials of degree k as high as possible. Since $n + n = 2n$ is the number of coefficients of a polynomial of degree $2n - 1$, it follows that $k \leqq 2n - 1$.

Gauss has shown that exactness for polynomials of degree not exceeding $2n - 1$ (instead of $n - 1$ for predetermined nodes) can be attained, and he has given the location of the t_j (= the jth zero of the Legendre polynomial P_n in Sec. 5.3) and the coefficients A_j which depend on n but not on $f(t)$, and are obtained by using Lagrange's interpolation polynomial, as shown in Ref. [E5] listed in App. 1. With these t_j and A_j, formula (11) is called a **Gauss integration formula** or *Gauss quadrature formula*. Its degree of precision is $2n - 1$, as just explained. Table 19.7 gives the values needed for $n = 2, \cdots, 5$. (For larger n, see pp. 916–919 of Ref. [GenRef1] in App. 1.)

Table 19.7 Gauss Integration: Nodes t_j and Coefficients A_j

n	Nodes t_j	Coefficients A_j	Degree of Precision
2	−0.5773502692	1	3
	0.5773502692	1	
3	−0.7745966692	0.5555555556	5
	0	0.8888888889	
	0.7745966692	0.5555555556	
4	−0.8611363116	0.3478548451	7
	−0.3399810436	0.6521451549	
	0.3399810436	0.6521451549	
	0.8611363116	0.3478548451	
5	−0.9061798459	0.2369268851	9
	−0.5384693101	0.4786286705	
	0	0.5688888889	
	0.5384693101	0.4786286705	
	0.9061798459	0.2369268851	

EXAMPLE 7 Gauss Integration Formula with $n = 3$

Evaluate the integral in Example 3 by the Gauss integration formula (11) with $n = 3$.

Solution. We have to convert our integral from 0 to 1 into an integral from -1 to 1. We set $x = \frac{1}{2}(t + 1)$. Then $dx = \frac{1}{2}\,dt$, and (11) with $n = 3$ and the above values of the nodes and the coefficients yields

$$\int_0^1 \exp(-x^2)\,dx = \frac{1}{2}\int_{-1}^{1} \exp\left(-\frac{1}{4}(t+1)^2\right)dt$$

$$\approx \frac{1}{2}\left[\frac{5}{9}\exp\left(-\frac{1}{4}\left(1-\sqrt{\frac{3}{5}}\right)^2\right) + \frac{8}{9}\exp\left(-\frac{1}{4}\right) + \frac{5}{9}\exp\left(-\frac{1}{4}\left(1+\sqrt{\frac{3}{5}}\right)^2\right)\right] = 0.746815$$

(exact to 6D: 0.746825), which is almost as accurate as the Simpson result obtained in Example 3 with a much larger number of arithmetic operations. With 3 function values (as in this example) and Simpson's rule we would get $\frac{1}{6}(1 + 4e^{-0.25} + e^{-1}) = 0.747180$, with an error over 30 times that of the Gauss integration. ■

EXAMPLE 8 **Gauss Integration Formula with $n = 4$ and 5**

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 by Gauss. Compare with the adaptive integration in Example 6 and comment.

Solution. $x = t + 1$ gives $f(t) = \frac{1}{4}\pi(t+1)^4 \cos(\frac{1}{4}\pi(t+1))$, as needed in (11). For $n = 4$ we calculate (6S)

$$J \approx A_1 f_1 + \cdots + A_4 f_4 = A_1(f_1 + f_4) + A_2(f_2 + f_3)$$

$$= 0.347855(0.000290309 + 1.02570) + 0.652145(0.129464 + 1.25459) = 1.25950.$$

The error is 0.00003 because $J = 1.25953$ (6S). Calculating with 10S and $n = 4$ gives the same result; so the error is due to the formula, not rounding. For $n = 5$ and 10S we get $J \approx 1.259526185$, too large by the amount 0.000000250 because $J = 1.259525935$ (10S). The accuracy is impressive, particularly if we compare the amount of work with that in Example 6. ■

Gauss integration is of considerable practical importance. Whenever the integrand f is given by a formula (not just by a table of numbers) or when experimental measurements can be set at times t_j (or whatever t represents) shown in Table 19.7 or in Ref. [GenRef1], then the great accuracy of Gauss integration outweighs the disadvantage of the complicated t_j and A_j (which may have to be stored). Also, Gauss coefficients A_j are positive for all n, in contrast with some of the Newton–Cotes coefficients for larger n.

Of course, there are frequent applications with equally spaced nodes, so that Gauss integration does not apply (or has no great advantage if one first has to get the t_j in (11) by interpolation).

Since the endpoints -1 and 1 of the interval of integration in (11) are not zeros of P_n, they do not occur among $t_0, \cdots, t_n$, and the Gauss formula (11) is called, therefore, an **open formula**, in contrast with a **closed formula**, in which the endpoints of the interval of integration are t_0 and t_n. [For example, (2) and (7) are closed formulas.]

Numeric Differentiation

Numeric differentiation is the computation of values of the derivative of a function f from given values of f. Numeric differentiation should be avoided whenever possible. Whereas *integration* is a smoothing process and is not very sensitive to small inaccuracies in function values, *differentiation* tends to make matters rough and generally gives values of f' that are much less accurate than those of f. The difficulty with differentiation is tied in with the definition of the derivative, which is the limit of the difference quotient, and, in that quotient, you usually have the difference of a large quantity divided by a small quantity. This can cause numerical instability. While being aware of this caveat, we must still develop basic differentiation formulas for use in numeric solutions of differential equations.

We use the notations $f_j' = f'(x_j)$, $f_j'' = f''(x_j)$, etc., and may obtain rough approximation formulas for derivatives by remembering that

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

This suggests

$$f'_{1/2} \approx \frac{\delta f_{1/2}}{h} = \frac{f_1 - f_0}{h}. \tag{12}$$

Similarly, for the second derivative we obtain

$$f''_1 \approx \frac{\delta^2 f_1}{h^2} = \frac{f_2 - 2f_1 + f_0}{h^2}, \qquad \text{etc.} \tag{13}$$

More accurate approximations are obtained by differentiating suitable Lagrange polynomials. Differentiating (6) and remembering that the denominators in (6) are $2h^2$, $-h^2$, $2h^2$, we have

$$f'(x) \approx p'_2(x) = \frac{2x - x_1 - x_2}{2h^2} f_0 - \frac{2x - x_0 - x_2}{h^2} f_1 + \frac{2x - x_0 - x_1}{2h^2} f_2.$$

Evaluating this at x_0, x_1, x_2, we obtain the "three-point formulas"

$$\begin{aligned} &\text{(a)} \quad f'_0 \approx \frac{1}{2h}(-3f_0 + 4f_1 - f_2), \\ &\text{(b)} \quad f'_1 \approx \frac{1}{2h}(-f_0 + f_2), \\ &\text{(c)} \quad f'_2 \approx \frac{1}{2h}(f_0 - 4f_1 + 3f_2). \end{aligned} \tag{14}$$

Applying the same idea to the Lagrange polynomial $p_4(x)$, we obtain similar formulas, in particular,

$$f'_2 \approx \frac{1}{12h}(f_0 - 8f_1 + 8f_3 - f_4). \tag{15}$$

Some examples and further formulas are included in the problem set as well as in Ref. [E5] listed in App. 1.

PROBLEM SET 19.5

1–6 RECTANGULAR AND TRAPEZOIDAL RULES

1. **Rectangular rule.** Evaluate the integral in Example 1 by the rectangular rule (1) with subintervals of length 0.1. Compare with Example 1. (6S-exact: 0.746824)
2. **Bounds for (1).** Derive a formula for lower and upper bounds for the rectangular rule. Apply it to Prob. 1.
3. **Trapezoidal rule.** To get a feel for increase in accuracy, integrate x^2 from 0 to 1 by (2) with $h = 1, 0.5, 0.25, 0.1$.
4. **Error estimation by halfing.** Integrate $f(x) = x^4$ from 0 to 1 by (2) with $h = 1$, $h = 0.5$, $h = 0.25$ and estimate the error for $h = 0.5$ and $h = 0.25$ by (5).
5. **Error estimation.** Do the tasks in Prob. 4 for $f(x) = \sin \frac{1}{2}\pi x$.

6. Stability. Prove that the trapezoidal rule is stable with respect to rounding.

7–15 SIMPSON'S RULE

Evaluate the integrals $A = \int_1^2 \frac{dx}{x}$, $B = \int_0^{0.4} xe^{-x^2}\, dx$, $J = \int_0^1 \frac{dx}{1 + x^2}$ by Simpson's rule with $2m$ as indicated, and compare with the exact value known from calculus.

7. A, $2m = 4$

8. A, $2m = 10$

9. B, $2m = 4$

10. B, $2m = 10$

11. J, $2m = 4$

12. J, $2m = 10$

13. Error estimate. Compute the integral J by Simpson's rule with $2m = 8$ and use the value and that in Prob. 11 to estimate the error by (10).

14. Error bounds and estimate. Integrate e^{-x} from 0 to 2 by (7) with $h = 1$ and with $h = 0.5$. Give error bounds for the $h = 0.5$ value and an error estimate by (10).

15. Given TOL. Find the smallest n in computing A (see Probs. 7 and 8) such that 5S-accuracy is guaranteed **(a)** by (4) in the use of (2), **(b)** by (9) in the use of (7).

16–21 NONELEMENTARY INTEGRALS

The following integrals cannot be evaluated by the usual methods of calculus. Evaluate them as indicated. Compare your value with that possibly given by your CAS. Si(x) is the sine integral. S(x) and C(x) are the Fresnel integrals. See App. A3.1. They occur in optics.

$$\mathrm{Si}(x) = \int_0^x \frac{\sin x^*}{x^*}\, dx^*,$$

$$\mathrm{S}(x) = \int_0^x \sin(x^{*2})\, dx^*, \quad \mathrm{C}(x) = \int_0^x \cos(x^{*2})\, dx^*$$

16. Si(1) by (2), $n = 5$, $n = 10$, and apply (5).

17. Si(1) by (7), $2m = 2$, $2m = 4$

18. Obtain a better value in Prob. 17. *Hint.* Use (10).

19. Si(1) by (7), $2m = 10$

20. S(1.25) by (7), $2m = 10$

21. C(1.25) by (7), $2m = 10$

22–25 GAUSS INTEGRATION

Integrate by (11) with $n = 5$:

22. $\cos x$ from 0 to $\frac{1}{2}\pi$

23. xe^{-x} from 0 to 1

24. $\sin(x^2)$ from 0 to 1.25

25. $\exp(-x^2)$ from 0 to 1

26. TEAM PROJECT. Romberg Integration (W. Romberg, *Norske Videnskab. Trondheim, Fφrh.* 28, Nr. 7, 1955). This method uses the trapezoidal rule and gains precision stepwise by halving h and adding an error estimate. Do this for the integral of $f(x) = e^{-x}$ from $x = 0$ to $x = 2$ with TOL $= 10^{-3}$, as follows.

Step 1. Apply the trapezoidal rule (2) with $h = 2$ (hence $n = 1$) to get an approximation J_{11}. Halve h and use (2) to get J_{21} and an error estimate

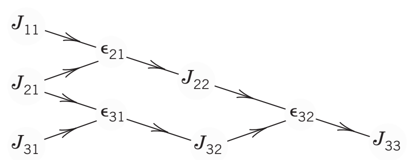
$$\epsilon_{21} = \frac{1}{2^2 - 1}(J_{21} - J_{11}).$$

If $|\epsilon_{21}| \leqq$ TOL, stop. The result is $J_{22} = J_{21} + \epsilon_{21}$.

Step 2. Show that $\epsilon_{21} = -0.066596$, hence $|\epsilon_{21}| >$ TOL and go on. Use (2) with $h/4$ to get J_{31} and add to it the error estimate $\epsilon_{31} = \frac{1}{3}(J_{31} - J_{21})$ to get the better $J_{32} = J_{31} + \epsilon_{31}$. Calculate

$$\epsilon_{32} = \frac{1}{2^4 - 1}(J_{32} - J_{22}) = \frac{1}{15}(J_{32} - J_{22}).$$

If $|\epsilon_{32}| \leqq$ TOL, stop. The result is $J_{33} = J_{32} + \epsilon_{32}$. (Why does $2^4 = 16$ come in?) Show that we obtain $\epsilon_{32} = -0.000266$, so that we can stop. Arrange your J- and ϵ-values in a kind of "difference table."



If $|\epsilon_{32}|$ were greater than TOL, you would have to go on and calculate in the next step J_{41} from (2) with $h = \frac{1}{4}$; then

$$J_{42} = J_{41} + \epsilon_{41} \quad \text{with} \quad \epsilon_{41} = \tfrac{1}{3}(J_{41} - J_{31})$$
$$J_{43} = J_{42} + \epsilon_{42} \quad \text{with} \quad \epsilon_{42} = \tfrac{1}{15}(J_{42} - J_{32})$$
$$J_{44} = J_{43} + \epsilon_{43} \quad \text{with} \quad \epsilon_{43} = \tfrac{1}{63}(J_{43} - J_{33})$$

where $63 = 2^6 - 1$. (How does this come in?)

Apply the Romberg method to the integral of $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 with TOL $= 10^{-4}$.

27–30 DIFFERENTIATION

27. Consider $f(x) = x^4$ for $x_0 = 0, x_1 = 0.2, x_2 = 0.4, x_3 = 0.6, x_4 = 0.8$. Calculate f_2' from (14a), (14b), (14c), (15). Determine the errors. Compare and comment.

28. A "**four-point formula**" for the derivative is

$$f_2' \approx \frac{1}{6h}\,(-2f_1 - 3f_2 + 6f_3 - f_4).$$

Apply it to $f(x) = x^4$ with $x_1, \cdots, x_4$ as in Prob. 27, determine the error, and compare it with that in the case of (15).

29. The derivative $f'(x)$ can also be approximated in terms of first-order and higher order differences (see Sec. 19.3):

$$f'(x_0) \approx \frac{1}{h}\bigg(\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + - \cdots\bigg).$$

Compute $f'(0.4)$ in Prob. 27 from this formula, using differences up to and including first order, second order, third order, fourth order.

30. Derive the formula in Prob. 29 from (14) in Sec. 19.3.

CHAPTER 19 REVIEW QUESTIONS AND PROBLEMS

1. What is a numeric method? How has the computer influenced numerics?

2. What is an error? A relative error? An error bound?

3. Why are roundoff errors important? State the rounding rules.

4. What is an algorithm? Which of its properties are important in software implementation?

5. What do you know about stability?

6. Why is the selection of a *good* method at least as important on a large computer as it is on a small one?

7. Can the Newton (–Raphson) method diverge? Is it fast? Same questions for the bisection method.

8. What is fixed-point iteration?

9. What is the advantage of Newton's interpolation formulas over Lagrange's?

10. What is spline interpolation? Its advantage over polynomial interpolation?

11. List and compare the integration methods we have discussed.

12. How did we use an interpolation polynomial in deriving Simpson's rule?

13. What is adaptive integration? Why is it useful?

14. In what sense is Gauss integration optimal?

15. How did we obtain formulas for numeric differentiation?

16. Write $-46.9028104, 0.000317399, 54/7, -890/3$ in floating-point form with 5S (5 significant digits, properly rounded).

17. Compute $(5.346 - 3.644)/(3.444 - 3.055)$ as given and then rounded stepwise to 3S, 2S, 1S. Comment. ("Stepwise" means rounding the rounded numbers, not the given ones.)

18. Compute $0.38755/(5.6815 - 0.38419)$ as given and then rounded stepwise to 4S, 3S, 2S, 1S. Comment.

19. Let 19.1 and 25.84 be correctly rounded. Find the shortest interval in which the sum s of the true (unrounded) numbers must lie.

20. Do the same task as in Prob. 19 for the difference $3.2 - 6.29$.

21. What is the relative error of $n\tilde{a}$ in terms of that of $\tilde{a}$?

22. Show that the relative error of $\tilde{a}^2$ is about twice that of $\tilde{a}$.

23. Solve $x^2 - 40x + 2 = 0$ in two ways (cf. Sec. 19.1). Use 4S-arithmetic.

24. Solve $x^2 - 100x + 1 = 0$. Use 5S-arithmetic.

25. Compute the solution of $x^4 = x + 0.1$ near $x = 0$ by transforming the equation algebraically to the form $x = g(x)$ and starting from $x_0 = 0$.

26. Solve $\cos x = x^2$ by Newton's method, starting from $x = 0.5$.

27. Solve Prob. 25 by bisection (3S-accuracy).

28. Compute $\sinh 0.4$ from $\sinh 0$, $\sinh 0.5 = 0.521$, $\sinh 1.0 = 1.175$ by quadratic interpolation.

29. Find the cubic spline for the data $f(0) = 0$, $f(1) = 0$, $f(2) = 4$, $k_0 = -1$, $k_2 = 5$.

30. Find the cubic spline q and the interpolation polynomial p for the data (0, 0), (1, 1), (2, 6), (3, 10), with $q'(0) = 0$, $q'(3) = 0$ and graph p and q on common axes.

31. Compute the integral of x^3 from 0 to 1 by the trapezoidal rule with $n = 5$. What error bounds are obtained from (4) in Sec. 19.5? What is the actual error of the result?

32. Compute the integral of $\cos(x^2)$ from 0 to 1 by Simpson's rule with $2m = 4$.

33. Solve Prob. 32 by Gauss integration with $n = 3$ and $n = 5$.

34. Compute $f'(0.2)$ for $f(x) = x^3$ using (14b) in Sec. 19.5 with **(a)** $h = 0.2$, **(b)** $h = 0.1$. Compare the accuracy.

35. Compute $f''(0.2)$ for $f(x) = x^3$ using (13) in Sec. 19.5 with **(a)** $h = 0.2$, **(b)** $h = 0.1$.

SUMMARY OF CHAPTER 19
Numerics in General

In this chapter we discussed concepts that are relevant throughout numeric work as a whole and methods of a general nature, as opposed to methods for linear algebra (Chap. 20) or differential equations (Chap. 21).

In scientific computations we use the ***floating-point*** representation of numbers (Sec. 19.1); fixed-point representation is less suitable in most cases.

Numeric methods give approximate values $\tilde{a}$ of quantities. The **error** ϵ of $\tilde{a}$ is

$$\epsilon = a - \tilde{a} \qquad \text{(Sec. 19.1)} \tag{1}$$

where a is the exact value. The *relative error* of $\tilde{a}$ is ϵ/a. Errors arise from rounding, inaccuracy of measured values, truncation (that is, replacement of integrals by sums, series by partial sums), and so on.

An algorithm is called **numerically stable** if small changes in the initial data give only correspondingly small changes in the final results. Unstable algorithms are generally useless because errors may become so large that results will be very inaccurate. The numeric instability of algorithms must not be confused with the mathematical instability of problems ("*ill-conditioned problems*," Sec. 19.2).

Fixed-point iteration is a method for solving equations $f(x) = 0$ in which the equation is first transformed algebraically to $x = g(x)$, an initial guess x_0 for the solution is made, and then approximations $x_1, x_2, \cdots$, are successively computed by iteration from (see Sec. 19.2)

$$x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots). \tag{2}$$

Newton's method for solving equations $f(x) = 0$ is an iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad \text{(Sec. 19.2).} \tag{3}$$

Here x_{n+1} is the x-intercept of the tangent of the curve $y = f(x)$ at the point x_n. This method is of second order (Theorem 2, Sec. 19.2). If we replace f' in (3) by a difference quotient (geometrically: we replace the tangent by a secant), we obtain the **secant method;** see (10) in Sec. 19.2. For the *bisection method* (which converges slowly) and the *method of false position,* see Problem Set 19.2.

Polynomial interpolation means the determination of a polynomial $p_n(x)$ such that $p_n(x_j) = f_j$, where $j = 0, \cdots, n$ and $(x_0, f_0), \cdots, (x_n, f_n)$ are measured or observed values, values of a function, etc. $p_n(x)$ is called an *interpolation polynomial.* For given data, $p_n(x)$ of degree n (or less) is unique. However, it can be written in different forms, notably in **Lagrange's form** (4), Sec. 19.3, or in **Newton's divided difference form** (10), Sec. 19.3, which requires fewer operations. For regularly spaced $x_0, x_1 = x_0 + h, \cdots, x_n = x_0 + nh$ the latter becomes **Newton's forward difference formula** (formula (14) in Sec. 19.3):

$$(4) \qquad f(x) \approx p_n(x) = f_0 + r\,\Delta f_0 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!}\,\Delta^n f_0$$

where $r = (x - x_0)/h$ and the forward differences are $\Delta f_j = f_{j+1} - f_j$ and

$$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 2, 3, \cdots).$$

A similar formula is *Newton's backward difference interpolation formula* (formula (18) in Sec. 19.3).

Interpolation polynomials may become numerically unstable as n increases, and instead of interpolating and approximating by a single high-degree polynomial it is preferable to use a cubic **spline** $g(x)$, that is, a twice continuously differentiable interpolation function [thus, $g(x_j) = f_j$], which in each subinterval $x_j \leqq x \leqq x_{j+1}$ consists of a cubic polynomial $q_j(x)$; see Sec. 19.4.

Simpson's rule of numeric integration is [see (7), Sec. 19.5]

$$(5) \quad \int_a^b f(x)\,dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m})$$

with equally spaced nodes $x_j = x_0 + jh, j = 1, \cdots, 2m, h = (b - a)/(2m)$, and $f_j = f(x_j)$. It is simple but accurate enough for many applications. Its degree of precision is DP $= 3$ because the error (8), Sec. 19.5, involves h^4. A more practical error estimate is (10), Sec. 19.5,

$$\epsilon_{h/2} = \tfrac{1}{15}(J_{h/2} - J_h),$$

obtained by first computing with step h, then with step $h/2$, and then taking $\frac{1}{15}$ of the difference of the results.

Simpson's rule is the most important of the **Newton–Cotes formulas**, which are obtained by integrating Lagrange interpolation polynomials, linear ones for the **trapezoidal rule** (2), Sec. 19.5, quadratic for Simpson's rule, cubic for the *three-eights rule* (see the Chap. 19 Review Problems), etc.

Adaptive integration (Sec. 19.5, Example 6) is integration that adjusts (*"adapts"*) the step (automatically) to the variability of $f(x)$.

Romberg integration (Team Project 26, Problem Set 19.5) starts from the trapezoidal rule (2), Sec. 19.5, with $h, h/2, h/4$, etc. and improves results by systematically adding error estimates.

Gauss integration (11), Sec. 19.5, is important because of its great accuracy (DP $= 2n - 1$, compared to Newton–Cotes's DP $= n - 1$ or n). This is achieved by an optimal choice of the nodes, which are not equally spaced; see Table 19.7, Sec. 19.5.

Numeric differentiation is discussed at the end of Sec. 19.5. (Its main application (to differential equations) follows in Chap. 21.)

CHAPTER 20

Numeric Linear Algebra

This chapter deals with two main topics. The first topic is how to solve linear systems of equations numerically. We start with Gauss elimination, which may be familiar to some readers, but this time in an algorithmic setting with partial pivoting. Variants of this method (Doolittle, Crout, Cholesky, Gauss–Jordan) are discussed in Sec. 20.2. All these methods are direct methods, that is, methods of numerics where we know in advance how many steps they will take until they arrive at a solution. However, small pivots and roundoff error magnification may produce nonsensical results, such as in the Gauss method. A shift occurs in Sec. 20.3, where we discuss numeric iteration methods or indirect methods to address our first topic. Here we cannot be totally sure how many steps will be needed to arrive at a good answer. Several factors—such as how far is the starting value from our initial solution, how is the problem structure influencing speed of convergence, how accurate would we like our result to be—determine the outcome of these methods. Moreover, our computation cycle may not converge. Gauss–Seidel iteration and Jacobi iteration are discussed in Sec. 20.3. Section 20.4 is at the heart of addressing the pitfalls of numeric linear algebra. It is concerned with problems that are ill-conditioned. We learn to estimate how "bad" such a problem is by calculating the condition number of its matrix.

The second topic (Secs. 20.6–20.9) is how to solve eigenvalue problems numerically. Eigenvalue problems appear throughout engineering, physics, mathematics, economics, and many areas. For large or very large matrices, determining the eigenvalues is difficult as it involves finding the roots of the characteristic equations, which are high-degree polynomials. As such, there are different approaches to tackling this problem. Some methods, such as Gerschgorin's method and Collatz's method only provide a range in which eigenvalues lie and thus are known as inclusion methods. Others such as tridiagonalization and QR-factorization actually find all the eigenvalues. The area is quite ingeneous and should be fascinating to the reader.

COMMENT. ***This chapter is independent of Chap. 19 and can be studied immediately after Chap. 7 or 8.***

Prerequisite: Secs. 7.1, 7.2, 8.1.
Sections that may be omitted in a shorter course: 20.4, 20.5, 20.9.
References and Answers to Problems: App. 1 Part E, App. 2.

20.1 Linear Systems: Gauss Elimination

The basic method for solving systems of linear equations by Gauss elimination and back substitution was explained in Sec. 7.3. If you covered Sec. 7.3, you may wonder why we cover Gauss elimination again. The reason is that *here we cover Gauss elimination in the* *setting of numerics* and introduce new material such as pivoting, row scaling, and operation count. Furthermore, we give an algorithmic representation of Gauss elimination in Table 20.1 that can be readily converted into software. We also show when Gauss elimination runs into difficulties with small pivots and what to do about it. The reader should pay close attention to the material as variants of Gauss elimination are covered in Sec. 20.2 and, furthermore, the general problem of solving linear systems is the focus of the first half of this chapter.

A **linear system of n equations** in n *unknowns* $x_1, \cdots, x_n$ is a set of equations $\mathrm{E}_1, \cdots, \mathrm{E}_n$ of the form

$$
\begin{array}{ll}
\mathrm{E}_1: & a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\mathrm{E}_2: & a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\
 & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\
\mathrm{E}_n: & a_{n1}x_1 + \cdots + a_{nn}x_n = b_n
\end{array} \tag{1}
$$

where the **coefficients** a_{jk} and the b_j are given numbers. The system is called **homogeneous** if all the b_j are zero; otherwise it is called **nonhomogeneous.** Using matrix multiplication (Sec. 7.2), we can write (1) as a single vector equation

$$
\mathbf{Ax} = \mathbf{b} \tag{2}
$$

where the **coefficient matrix $\mathbf{A}$** $= [a_{jk}]$ is the $n \times n$ matrix

$$
\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}
$$

are column vectors. The following matrix $\widetilde{\mathbf{A}}$ is called the **augmented matrix** of the system (1):

$$
\widetilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & a_{2n} & b_2 \\ \cdot & \cdots & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix}.
$$

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfy all the n equations, and a **solution vector** of (1) is a vector $\mathbf{x}$ whose components constitute a solution of (1).

The method of solving such a system by determinants (Cramer's rule in Sec. 7.7) is not practical, even with efficient methods for evaluating the determinants.

A practical method for the solution of a linear system is the so-called *Gauss elimination*, which we shall now discuss (***proceeding independently of Sec. 7.3***).

Gauss Elimination

This standard method for solving linear systems (1) is a systematic process of elimination that reduces (1) to **triangular form** because the system can then be easily solved by **back substitution**. For instance, a triangular system is

$$\begin{aligned} 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7 \\ 6x_3 &= 3 \end{aligned}$$

and back substitution gives $x_3 = \frac{3}{6} = \frac{1}{2}$ from the third equation, then

$$x_2 = \tfrac{1}{8}(-7 - 2x_3) = -1$$

from the second equation, and finally from the first equation

$$x_1 = \tfrac{1}{3}(8 - 5x_2 - 2x_3) = 4.$$

How do we reduce a given system (1) to triangular form? In the first step we ***eliminate*** x_1 from equations E_2 to E_n in (1). We do this by adding (or subtracting) suitable multiples of E_1 to (from) equations $\mathrm{E}_2, \cdots, \mathrm{E}_n$ and taking the resulting equations, call them $\mathrm{E}_2^*, \cdots, \mathrm{E}_n^*$ as the new equations. The first equation, E_1, is called the **pivot equation** in this step, and a_{11} is called the **pivot**. This equation is left unaltered. In the second step we take the new second equation E_2^* (which no longer contains x_1) as the pivot equation and use it to ***eliminate*** $\boldsymbol{x_2}$ from E_3^* to E_n^*. And so on. After $n - 1$ steps this gives a triangular system that can be solved by back substitution as just shown. In this way we obtain precisely all solutions of the *given* system (as proved in Sec. 7.3).

The pivot a_{kk} (in step k) ***must be*** different from zero and ***should be*** large in absolute value to avoid roundoff magnification by the multiplication in the elimination. For this we choose as our pivot equation one that has the absolutely largest a_{jk} in column k on or below the main diagonal (actually, the uppermost if there are several such equations). This popular method is called **partial pivoting**. It is used in CASs (e.g., in Maple).

Partial pivoting distinguishes it from **total pivoting**, which involves both row and column interchanges but is hardly used in practice.

Let us illustrate this method with a simple example.

EXAMPLE 1 Gauss Elimination. Partial Pivoting

Solve the system

$$\begin{aligned} \mathrm{E}_1: &\qquad 8x_2 + 2x_3 = -7 \\ \mathrm{E}_2: &\qquad 3x_1 + 5x_2 + 2x_3 = 8 \\ \mathrm{E}_3: &\qquad 6x_1 + 2x_2 + 8x_3 = 26. \end{aligned}$$

Solution. We must pivot since E_1 has no x_1-term. In Column 1, equation E_3 has the largest coefficient. Hence we interchange E_1 and E_3,

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7. \end{aligned}$$

Step 1. Elimination of x_1

It would suffice to show the augmented matrix and operate on it. We show both the equations and the augmented matrix. In the first step, the first equation is the pivot equation. Thus

$$\begin{array}{lr} \text{Pivot 6} \longrightarrow & 6x_1 + 2x_2 + 8x_3 = \ \ 26 \\ \text{Eliminate} \longrightarrow & \boxed{3x_1} + 5x_2 + 2x_3 = \ \ \ 8 \\ & 8x_2 + 2x_3 = -7 \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 3 & 5 & 2 & 8 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

To eliminate x_1 from the other equations (here, from the second equation), do:

Subtract $\frac{3}{6} = \frac{1}{2}$ times the pivot equation from the second equation.

The result is

$$\begin{array}{r} 6x_1 + 2x_2 + 8x_3 = \ \ 26 \\ 4x_2 - 2x_3 = -5 \\ 8x_2 + 2x_3 = -7 \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 4 & -2 & -5 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

Step 2. Elimination of x_2

The largest coefficient in Column 2 is 8. Hence we take the *new* third equation as the pivot equation, interchanging equations 2 and 3,

$$\begin{array}{lr} & 6x_1 + 2x_2 + 8x_3 = \ \ 26 \\ \text{Pivot 8} \longrightarrow & 8x_2 + 2x_3 = -7 \\ \text{Eliminate} \longrightarrow & \boxed{4x_2} - 2x_3 = -5 \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 4 & -2 & -5 \end{array}\right].$$

To eliminate x_2 from the third equation, do:

Subtract $\frac{1}{2}$ times the pivot equation from the third equation.

The resulting triangular system is shown below. This is the end of the forward elimination. Now comes the back substitution.

Back substitution. ***Determination of x_3, x_2, x_1***

The triangular system obtained in Step 2 is

$$\begin{array}{r} 6x_1 + 2x_2 + 8x_3 = \ \ 26 \\ 8x_2 + 2x_3 = -7 \\ -3x_3 = -\frac{3}{2} \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 0 & -3 & -\frac{3}{2} \end{array}\right].$$

From this system, taking the last equation, then the second equation, and finally the first equation, we compute the solution

$$\begin{aligned} x_3 &= \tfrac{1}{2} \\ x_2 &= \tfrac{1}{8}(-7 - 2x_3) = -1 \\ x_1 &= \tfrac{1}{6}(26 - 2x_2 - 8x_3) = 4. \end{aligned}$$

This agrees with the values given above, before the beginning of the example. ■

The general algorithm for the Gauss elimination is shown in Table 20.1. To help explain the algorithm, we have numbered some of its lines. b_j is denoted by $a_{j,n+1}$, for uniformity. In lines 1 and 2 we look for a possible pivot. [For $k = 1$ we can always find one; otherwise x_1 would not occur in (1).] In line 2 we do pivoting if necessary, picking an a_{jk} of greatest absolute value (the one with the smallest j if there are several) and interchange the corresponding rows. If $|a_{kk}|$ is greatest, we do no pivoting. m_{jk} in line 4 suggests *multiplier*, since these are the factors by which we have to multiply the pivot equation E_k^* in Step k before subtracting it from an equation E_j^* below E_k^* from which we want to eliminate x_k. Here we have written E_k^* and E_j^* to indicate that after Step 1 these are no longer the equations given in (1), but these underwent a change in each step, as indicated in line 5. Accordingly, a_{jk} etc. in all lines refer to the most recent equations, and $j \geqq k$ in line 1 indicates that we leave untouched all the equations that have served as pivot equations in previous steps. For $p = k$ in line 5 we get 0 on the right, as it should be in the elimination,

$$a_{jk} - m_{jk}a_{kk} = a_{jk} - \frac{a_{jk}}{a_{kk}}a_{kk} = 0.$$

In line 3, if the last equation in the *triangular* system is $0 = b_n^* \neq 0$, we have no solution. If it is $0 = b_n^* = 0$, we have no unique solution because we then have fewer equations than unknowns.

EXAMPLE 2 Gauss Elimination in Table 20.1, Sample Computation

In Example 1 we had $a_{11} = 0$, so that pivoting was necessary. The greatest coefficient in Column 1 was a_{31}. Thus $\tilde{j} = 3$ in line 2, and we interchanged E_1 and E_3. Then in lines 4 and 5 we computed $m_{21} = \frac{3}{6} = \frac{1}{2}$ and

$$a_{22} = 5 - \tfrac{1}{2} \cdot 2 = 4, \qquad a_{23} = 2 - \tfrac{1}{2} \cdot 8 = -2, \qquad a_{24} = 8 - \tfrac{1}{2} \cdot 26 = -5,$$

and then $m_{31} = \frac{0}{6} = 0$, so that the third equation $8x_2 + 2x_3 = -7$ did not change in Step 1. In Step 2 ($k = 2$) we had 8 as the greatest coefficient in Column 2, hence $\tilde{j} = 3$. We interchanged equations 2 and 3, computed $m_{32} = -\frac{4}{8} = -\frac{1}{2}$ in line 5, and the $a_{33} = -2 - \frac{1}{2} \cdot 2 = -3$, $a_{34} = -5 - \frac{1}{2}(-7) = -\frac{3}{2}$. This produced the triangular form used in the back substitution. ■

If $a_{kk} = 0$ in Step k, ***we must pivot***. If $|a_{kk}|$ is small, ***we should pivot*** because of roundoff error magnification that may seriously affect accuracy or even produce nonsensical results.

EXAMPLE 3 Difficulty with Small Pivots

The solution of the system

$$\begin{aligned} 0.0004x_1 + 1.402x_2 &= 1.406 \\ 0.4003x_1 - 1.502x_2 &= 2.501 \end{aligned}$$

is $x_1 = 10$, $x_2 = 1$. We solve this system by the Gauss elimination, using four-digit floating-point arithmetic. (4D is for simplicity. Make an 8D-arithmetic example that shows the same.)

(a) Picking the first of the given equations as the pivot equation, we have to multiply this equation by $m = 0.4003/0.0004 = 1001$ and subtract the result from the second equation, obtaining

$$-1405x_2 = -1404.$$

Hence $x_2 = -1404/(-1405) = 0.9993$, and from the first equation, instead of $x_1 = 10$, we get

$$x_1 = \frac{1}{0.0004}(1.406 - 1.402 \cdot 0.9993) = \frac{0.005}{0.0004} = 12.5.$$

This failure occurs because $|a_{11}|$ is small compared with $|a_{12}|$, so that a small roundoff error in x_2 leads to a large error in x_1.

(b) Picking the second of the given equations as the pivot equation, we have to multiply this equation by $0.0004/0.4003 = 0.0009993$ and subtract the result from the first equation, obtaining

$$1.404x_2 = 1.404.$$

Hence $x_2 = 1$, and from the pivot equation $x_1 = 10$. This success occurs because $|a_{21}|$ is not very small compared to $|a_{22}|$, so that a small roundoff error in x_2 would not lead to a large error in x_1. Indeed, for instance, if we had the value $x_2 = 1.002$, we would still have from the pivot equation the good value $x_1 = (2.501 + 1.505)/0.4003 = 10.01$. ■

Table 20.1 Gauss Elimination

ALGORITHM GAUSS ($\tilde{\mathbf{A}} = [a_{jk}] = [\mathbf{A} \quad \mathbf{b}]$)

This algorithm computes a unique solution $\mathbf{x} = [x_j]$ of the system (1) or indicates that (1) has no unique solution.

INPUT: Augmented $n \times (n+1)$ matrix $\tilde{\mathbf{A}} = [a_{jk}]$, where $a_{j,n+1} = b_j$

OUTPUT: Solution $\mathbf{x} = [x_j]$ of (1) or message that the system (1) has no unique solution

For $k = 1, \cdots, n-1$, do:

1 $\quad m = k$

For $j = k+1, \cdots, n$, do:

If $(|a_{mk}| < |a_{jk}|)$ then $m = j$

End

If $a_{mk} = 0$ then OUTPUT "No unique solution exists"

Stop

[*Procedure completed unsuccessfully*]

2 $\quad$ Else exchange row k and row m

3 $\quad$ If $a_{nn} = 0$ then OUTPUT "No unique solution exists."

Stop

Else

4 $\quad$ For $j = k+1, \cdots, n$, do:

$$m_{jk} := \frac{a_{jk}}{a_{kk}}$$

5 $\quad$ For $p = k+1, \cdots, n+1$, do:

$$a_{jp} := a_{jp} - m_{jk}a_{kp}$$

End

End

End

6 $$x_n = \frac{a_{n,n+1}}{a_{nn}} \qquad [\textit{Start back substitution}]$$

For $i = n-1, \cdots, 1$, do:

7 $$x_i = \frac{1}{a_{ii}}\left(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right)$$

End

OUTPUT $\mathbf{x} = [x_j]$. Stop

End GAUSS

Error estimates for the Gauss elimination are discussed in Ref. [E5] listed in App. 1.

Row scaling means the multiplication of each Row j by a suitable scaling factor s_j. It is done in connection with partial pivoting to get more accurate solutions. Despite much research (see Refs. [E9], [E24] in App. 1) and the proposition of several principles, scaling is still not well understood. As a possibility, one can scale for pivot choice only (not in the calculation, to avoid additional roundoff) and take as first pivot the entry a_{j1} for which $|a_{j1}|/|A_j|$ is largest; here A_j is an entry of largest absolute value in Row j. Similarly in the further steps of the Gauss elimination.

For instance, for the system

$$\begin{aligned} 4.0000x_1 + 14020x_2 &= 14060 \\ 0.4003x_1 - 1.502x_2 &= 2.501 \end{aligned}$$

we might pick 4 as pivot, but dividing the first equation by 10^4 gives the system in Example 3, for which the second equation is a better pivot equation.

Operation Count

Quite generally, important factors in judging the quality of a numeric method are

Amount of storage

Amount of time ($\equiv$ number of operations)

Effect of roundoff error

For the Gauss elimination, the operation count for a full matrix (a matrix with relatively many nonzero entries) is as follows. In Step k we eliminate x_k from $n - k$ equations. This needs $n - k$ divisions in computing the m_{jk} (line 3) and $(n - k)(n - k + 1)$ multiplications and as many subtractions (both in line 4). Since we do $n - 1$ steps, k goes from 1 to $n - 1$ and thus the total number of operations in this forward elimination is

$$\begin{aligned} f(n) &= \sum_{k=1}^{n-1} (n-k) + 2\sum_{k=1}^{n-1} (n-k)(n-k+1) \qquad \text{(write } n - k = s\text{)} \\ &= \sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s(s+1) = \tfrac{1}{2}(n-1)n + \tfrac{2}{3}(n^2-1)n \approx \tfrac{2}{3}n^3 \end{aligned}$$

where $2n^3/3$ is obtained by dropping lower powers of n. We see that $f(n)$ grows about proportional to n^3. We say that $f(n)$ is of *order* n^3 and write

$$f(n) = O(n^3)$$

where O suggests **order**. The general definition of O is as follows. We write

$$f(n) = O(h(n))$$

if the quotients $|f(n)/h(n)|$ and $|h(n)/f(n)|$ remain bounded (do not trail off to infinity) as $n \to \infty$. In our present case, $h(n) = n^3$ and, indeed, $f(n)/n^3 \to \frac{2}{3}$ because the omitted terms divided by n^3 go to zero as $n \to \infty$.

In the back substitution of x_i we make $n - i$ multiplications and as many subtractions, as well as 1 division. Hence the number of operations in the back substitution is

$$b(n) = 2\sum_{i=1}^{n}(n-i) + n = 2\sum_{s=1}^{n} s + n = n(n+1) + n = n^2 + 2n = O(n^2).$$

We see that it grows more slowly than the number of operations in the forward elimination of the Gauss algorithm, so that it is negligible for large systems because it is smaller by a factor n, approximately. For instance, if an operation takes 10^{-9} sec, then the times needed are:

Algorithm	$n = 1000$	$n = 10000$
Elimination	0.7 sec	11 min
Back substitution	0.001 sec	0.1 sec

PROBLEM SET 20.1

APPLICATIONS of linear systems see Secs. 7.1 and 8.2.

1–3 GEOMETRIC INTERPRETATION

Solve graphically and explain geometrically.

1. $x_1 - 4x_2 = 20.1$
$3x_1 + 5x_2 = 5.9$

2. $-5.00x_1 + 8.40x_2 = 0$
$10.25x_1 - 17.22x_2 = 0$

3. $7.2x_1 - 3.5x_2 = 16.0$
$-14.4x_1 + 7.0x_2 = 31.0$

4–16 GAUSS ELIMINATION

Solve the following linear systems by Gauss elimination, with partial pivoting if necessary (but without scaling). Show the intermediate steps. Check the result by substitution. If no solution or more than one solution exists, give a reason.

4. $6x_1 + x_2 = -3$
$4x_1 - 2x_2 = 6$

5. $2x_1 - 8x_2 = -4$
$3x_1 + x_2 = 7$

6. $25.38x_1 - 15.48x_2 = 30.60$
$-14.10x_1 + 8.60x_2 = -17.00$

7. $-3x_1 + 6x_2 - 9x_3 = -46.725$
$x_1 - 4x_2 + 3x_3 = 19.571$
$2x_1 + 5x_2 - 7x_3 = -20.073$

8. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

9. $6x_2 + 13x_3 = 137.86$
$6x_1 - 8x_3 = -85.88$
$13x_1 - 8x_2 = 178.54$

10. $4x_1 + 4x_2 + 2x_3 = 0$
$3x_1 - x_2 + 2x_3 = 0$
$3x_1 + 7x_2 + x_3 = 0$

11. $3.4x_1 - 6.12x_2 - 2.72x_3 = 0$
$-x_1 + 1.80x_2 + 0.80x_3 = 0$
$2.7x_1 - 4.86x_2 + 2.16x_3 = 0$

12. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

13.
$$\begin{aligned} 3x_2 + 5x_3 &= 1.20736 \\ 3x_1 - 4x_2 \qquad &= -2.34066 \\ 5x_1 \qquad + 6x_3 &= -0.329193 \end{aligned}$$

14.
$$\begin{aligned} -47x_1 + 4x_2 - 7x_3 &= -118 \\ 19x_1 - 3x_2 + 2x_3 &= 43 \\ -15x_1 + 5x_2 \qquad &= -25 \end{aligned}$$

15.
$$\begin{aligned} 2.2x_2 + 1.5x_3 - 3.3x_4 &= -9.30 \\ 0.2x_1 + 1.8x_2 \qquad + 4.2x_4 &= 9.24 \\ -x_1 - 3.1x_2 + 2.5x_3 \qquad &= -8.70 \\ 0.5x_1 \qquad - 3.8x_3 + 1.5x_4 &= 11.94 \end{aligned}$$

16.
$$\begin{aligned} 3.2x_1 + 1.6x_2 \qquad &= -0.8 \\ 1.6x_1 - 0.8x_2 + 2.4x_3 \qquad &= 16.0 \\ 2.4x_2 - 4.8x_3 + 3.6x_4 &= -39.0 \\ 3.6x_3 + 2.4x_4 &= 10.2 \end{aligned}$$

17. CAS EXPERIMENT. Gauss Elimination. Write a program for the Gauss elimination with pivoting. Apply it to Probs. 13–16. Experiment with systems whose coefficient determinant is small in absolute value. Also investigate the performance of your program for larger systems of your choice, including sparse systems.

18. TEAM PROJECT. Linear Systems and Gauss Elimination. (a) Existence and uniqueness. Find a and b such that $ax_1 + x_2 = b, x_1 + x_2 = 3$ has (i) a unique solution, (ii) infinitely many solutions, (iii) no solutions.

(b) Gauss elimination and nonexistence. Apply the Gauss elimination to the following two systems and compare the calculations step by step. Explain why the elimination fails if no solution exists.

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 13 \end{aligned}$$

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 12. \end{aligned}$$

(c) Zero determinant. Why may a computer program give you the result that a homogeneous linear system has only the trivial solution although you know its coefficient determinant to be zero?

(d) Pivoting. Solve System (A) (below) by the Gauss elimination first without pivoting. Show that for any fixed machine word length and sufficiently small $\epsilon > 0$ the computer gives $x_2 = 1$ and then $x_1 = 0$. What is the exact solution? Its limit as $\epsilon \to 0$? Then solve the system by the Gauss elimination with pivoting. Compare and comment.

(e) Pivoting. Solve System (B) by the Gauss elimination and three-digit rounding arithmetic, choosing (i) the first equation, (ii) the second equation as pivot equation. (Remember to round to 3S after each operation before doing the next, just as would be done on a computer!) Then use four-digit rounding arithmetic in those two calculations. Compare and comment.

(A)
$$\begin{aligned} \epsilon x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

(B)
$$\begin{aligned} 4.03x_1 + 2.16x_2 &= -4.61 \\ 6.21x_1 + 3.35x_2 &= -7.19 \end{aligned}$$

20.2 Linear Systems: LU-Factorization, Matrix Inversion

We continue our discussion of numeric methods for solving linear systems of n equations in n unknowns $x_1, \cdots, x_n$,

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

where $\mathbf{A} = [a_{jk}]$ is the $n \times n$ given coefficient matrix and $\mathbf{x}^\mathsf{T} = [x_1, \cdots, x_n]$ and $\mathbf{b}^\mathsf{T} = [b_1, \cdots, b_n]$. We present three related methods that are modifications of the Gauss elimination, which require fewer arithmetic operations. They are named after Doolittle, Crout, and Cholesky and use the idea of the LU-factorization of **A**, which we explain first.

An **LU-factorization** of a given square matrix **A** is of the form

$$\text{(2)} \qquad \boxed{\mathbf{A} = \mathbf{LU}}$$

where **L** is ***lower triangular*** and **U** is ***upper triangular***. For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 8 & 5 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & -7 \end{bmatrix}.$$

It can be proved that for any nonsingular matrix (see Sec. 7.8) the rows can be reordered so that the resulting matrix **A** has an LU-factorization (2) in which **L** turns out to be the matrix of the ***multipliers*** m_{jk} of the Gauss elimination, with main diagonal $1, \cdots, 1$, and **U** is the matrix of the triangular system at the end of the Gauss elimination. (See Ref. [E5], pp. 155–156, listed in App. 1.)

The ***crucial idea*** now is that **L** and **U** in (2) can be computed directly, without solving simultaneous equations (thus, without using the Gauss elimination). As a count shows, this needs about $n^3/3$ operations, about half as many as the Gauss elimination, which needs about $2n^3/3$ (see Sec. 20.1). And once we have (2), we can use it for solving $\mathbf{Ax} = \mathbf{b}$ in two steps, involving only about n^2 operations, simply by noting that $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ may be written

$$\text{(3)} \qquad \boxed{\text{(a)} \quad \mathbf{Ly} = \mathbf{b} \qquad \text{where} \qquad \text{(b)} \quad \mathbf{Ux} = \mathbf{y}}$$

and solving first (3a) for **y** and then (3b) for **x**. Here we can require that **L** have main diagonal $1, \cdots, 1$ as stated before; then this is called **Doolittle's method**.[1] Both systems (3a) and (3b) are triangular, so we can solve them as in the back substitution for the Gauss elimination.

A similar method, **Crout's method**,[2] is obtained from (2) if **U** (instead of **L**) is required to have main diagonal $1, \cdots, 1$. In either case the factorization (2) is unique.

EXAMPLE 1 Doolittle's Method

Solve the system in Example 1 of Sec. 20.1 by Doolittle's method.

Solution. The decomposition (2) is obtained from

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

[1]MYRICK H. DOOLITTLE (1830–1913). American mathematician employed by the U.S. Coast and Geodetic Survey Office. His method appeared in *U.S. Coast and Geodetic Survey*, 1878, 115–120.

[2]PRESCOTT DURAND CROUT (1907–1984), American mathematician, professor at MIT, also worked at General Electric.

by determining the m_{jk} and u_{jk}, using matrix multiplication. By going through **A** row by row we get successively

$a_{11} = 3 = 1 \cdot u_{11} = u_{11}$	$a_{12} = 5 = 1 \cdot u_{12} = u_{12}$	$a_{13} = 2 = 1 \cdot u_{13} = u_{13}$
$a_{21} = 0 = m_{21}u_{11}$ $m_{21} = 0$	$a_{22} = 8 = m_{21}u_{12} + u_{22}$ $u_{22} = 8$	$a_{23} = 2 = m_{21}u_{13} + u_{23}$ $u_{23} = 2$
$a_{31} = 6 = m_{31}u_{11}$ $= m_{31} \cdot 3$ $m_{31} = 2$	$a_{32} = 2 = m_{31}u_{12} + m_{32}u_{22}$ $= 2 \cdot 5 + m_{32} \cdot 8$ $m_{32} = -1$	$a_{33} = 8 = m_{31}u_{13} + m_{32}u_{23} + u_{33}$ $= 2 \cdot 2 - 1 \cdot 2 + u_{33}$ $u_{33} = 6$

Thus the factorization (2) is

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix}.$$

We first solve $\mathbf{Ly} = \mathbf{b}$, determining $y_1 = 8$, then $y_2 = -7$, then y_3 from $2y_1 - y_2 + y_3 = 16 + 7 + y_3 = 26$; thus (note the interchange in **b** because of the interchange in **A**!)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 26 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}.$$

Then we solve $\mathbf{Ux} = \mathbf{y}$, determining $x_3 = \frac{3}{6}$ then x_2, then x_1, that is,

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 4 \\ -1 \\ \frac{1}{2} \end{bmatrix}.$$

This agrees with the solution in Example 1 of Sec. 20.1. ■

Our formulas in Example 1 suggest that for general n the entries of the matrices $\mathbf{L} = [m_{jk}]$ (with main diagonal $1, \cdots, 1$ and m_{jk} suggesting "multiplier") and $\mathbf{U} = [u_{jk}]$ in the **Doolittle method** are computed from

$$\begin{aligned}
u_{1k} &= a_{1k} && k = 1, \cdots, n \\
m_{j1} &= \frac{a_{j1}}{u_{11}} && j = 2, \cdots, n \\
u_{jk} &= a_{jk} - \sum_{s=1}^{j-1} m_{js}u_{sk} && k = j, \cdots, n; \quad j \geqq 2 \\
m_{jk} &= \frac{1}{u_{kk}}\left(a_{jk} - \sum_{s=1}^{k-1} m_{js}u_{sk}\right) && j = k+1, \cdots, n; \quad k \geqq 2.
\end{aligned} \tag{4}$$

Row Interchanges. Matrices, such as

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

have no LU-factorization (try!). This indicates that for obtaining an LU-factorization, row interchanges of $\mathbf{A}$ (and corresponding interchanges in $\mathbf{b}$) may be necessary.

Cholesky's Method

For a ***symmetric, positive definite*** matrix $\mathbf{A}$ (thus $\mathbf{A} = \mathbf{A}^\mathsf{T}$, $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$) we can in (2) even choose $\mathbf{U} = \mathbf{L}^\mathsf{T}$, thus $u_{jk} = m_{kj}$ (but cannot impose conditions on the main diagonal entries). For example,

$$\text{(5)} \qquad \mathbf{A} = \begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \mathbf{L}\mathbf{L}^\mathsf{T} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix}.$$

The popular method of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ based on this factorization $\mathbf{A} = \mathbf{L}\mathbf{L}^\mathsf{T}$ is called **Cholesky's method**.[3] In terms of the entries of $\mathbf{L} = [l_{jk}]$ the formulas for the factorization are

$$\text{(6)} \qquad \begin{aligned} l_{11} &= \sqrt{a_{11}} & & \\ l_{j1} &= \frac{a_{j1}}{l_{11}} & & j = 2, \cdots, n \\ l_{jj} &= \sqrt{a_{jj} - \sum_{s=1}^{j-1} l_{js}^2} & & j = 2, \cdots, n \\ l_{pj} &= \frac{1}{l_{jj}} \left(a_{pj} - \sum_{s=1}^{j-1} l_{js} l_{ps} \right) & & p = j+1, \cdots, n; \quad j \geqq 2. \end{aligned}$$

If $\mathbf{A}$ is symmetric but not positive definite, this method could still be applied, but then leads to a *complex* matrix $\mathbf{L}$, so that the method becomes impractical.

EXAMPLE 2 Cholesky's Method

Solve by Cholesky's method:

$$\begin{aligned} 4x_1 + 2x_2 + 14x_3 &= 14 \\ 2x_1 + 17x_2 - 5x_3 &= -101 \\ 14x_1 - 5x_2 + 83x_3 &= 155. \end{aligned}$$

[3] ANDRÉ-LOUIS CHOLESKY (1875–1918), French military officer, geodecist, and mathematician. Surveyed Crete and North Africa. Died in World War I. His method was published posthumously in *Bulletin Géodésique* in 1924 but received little attention until JOHN TODD (1911–2007) — Irish-American mathematician, numerical analysist, and early pioneer of computer methods in numerics, professor at Caltech, and close personal friend and collaborator of ERWIN KREYSZIG, see [E20]—taught Cholesky's method in his analysis course at King's College, London, in the 1940s.

Solution. From (6) or from the form of the factorization

$$\begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

we compute, in the given order,

$$l_{11} = \sqrt{a_{11}} = 2 \qquad l_{21} = \frac{a_{21}}{l_{11}} = \frac{2}{2} = 1 \qquad l_{31} = \frac{a_{31}}{l_{11}} = \frac{14}{2} = 7$$

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{17 - 1} = 4$$

$$l_{32} = \frac{1}{l_{23}}(a_{32} - l_{31}l_{21}) = \frac{1}{4}(-5 - 7 \cdot 1) = -3$$

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{83 - 7^2 - (-3)^2} = 5.$$

This agrees with (5). We now have to solve $\mathbf{Ly} = \mathbf{b}$, that is,

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -101 \\ 155 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}.$$

As the second step, we have to solve $\mathbf{Ux} = \mathbf{L}^{\mathsf{T}}\mathbf{x} = \mathbf{y}$, that is,

$$\begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 3 \\ -6 \\ 1 \end{bmatrix}.$$ ■

THEOREM 1

Stability of the Cholesky Factorization

The Cholesky LL^{T}-factorization is numerically stable (as defined in Sec. 19.1).

PROOF We have $a_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2$ by squaring the third formula in (6) and solving it for a_{jj}. Hence for all l_{jk} (note that $l_{jk} = 0$ for $k > j$) we obtain (the inequality being trivial)

$$l_{jk}^2 \leqq l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2 = a_{jj}.$$

That is, l_{jk}^2 is bounded by an entry of $\mathbf{A}$, which means stability against rounding. ■

Gauss–Jordan Elimination. Matrix Inversion

Another variant of the Gauss elimination is the **Gauss–Jordan elimination**, introduced by W. Jordan in 1920, in which back substitution is avoided by additional computations that reduce the matrix to diagonal form, instead of the triangular form in the Gauss elimination. But this reduction from the Gauss triangular to the diagonal form requires more operations than back substitution does, so that the method is *disadvantageous* for solving systems $\mathbf{Ax} = \mathbf{b}$. But it may be used for matrix inversion, where the situation is as follows.

The **inverse** of a nonsingular square matrix $\mathbf{A}$ may be determined in principle by solving the n systems

$$\mathbf{A}\mathbf{x} = \mathbf{b}_j \qquad (j = 1, \cdots, n) \tag{7}$$

where $\mathbf{b}_j$ is the jth column of the $n \times n$ unit matrix.

However, it is preferable to produce $\mathbf{A}^{-1}$ by operating on the unit matrix $\mathbf{I}$ in the same way as the Gauss–Jordan algorithm, reducing $\mathbf{A}$ to $\mathbf{I}$. A typical illustrative example of this method is given in Sec. 7.8.

PROBLEM SET 20.2

1–5 DOOLITTLE'S METHOD

Show the factorization and solve by Doolittle's method.

1. $\begin{aligned} 4x_1 + 5x_2 &= 14 \\ 12x_1 + 14x_2 &= 36 \end{aligned}$

2. $\begin{aligned} 2x_1 + 9x_2 &= 82 \\ 3x_1 - 5x_2 &= -62 \end{aligned}$

3. $\begin{aligned} 5x_1 + 4x_2 + x_3 &= 6.8 \\ 10x_1 + 9x_2 + 4x_3 &= 17.6 \\ 10x_1 + 13x_2 + 15x_3 &= 38.4 \end{aligned}$

4. $\begin{aligned} 2x_1 + x_2 + 2x_3 &= 0 \\ -2x_1 + 2x_2 + x_3 &= 0 \\ x_1 + 2x_2 - 2x_3 &= 18 \end{aligned}$

5. $\begin{aligned} 3x_1 + 9x_2 + 6x_3 &= 4.6 \\ 18x_1 + 48x_2 + 39x_3 &= 27.2 \\ 9x_1 - 27x_2 + 42x_3 &= 9.0 \end{aligned}$

6. TEAM PROJECT. Crout's method factorizes $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is lower triangular and $\mathbf{U}$ is upper triangular with diagonal entries $u_{jj} = 1, j = 1, \cdots, n$.

(a) Formulas. Obtain formulas for Crout's method similar to (4).

(b) Examples. Solve Prob. 5 by Crout's method.

(c) Factor the following matrix by the Doolittle, Crout, and Cholesky methods.

$$\begin{bmatrix} 1 & -4 & 2 \\ -4 & 25 & 4 \\ 2 & 4 & 24 \end{bmatrix}$$

(d) Give the formulas for factoring a tridiagonal matrix by Crout's method.

(e) When can you obtain Crout's factorization from Doolittle's by transposition?

7–12 CHOLESKY'S METHOD

Show the factorization and solve.

7. $\begin{aligned} 9x_1 + 6x_2 + 12x_3 &= 17.4 \\ 6x_1 + 13x_2 + 11x_3 &= 23.6 \\ 12x_1 + 11x_2 + 26x_3 &= 30.8 \end{aligned}$

8. $\begin{aligned} 4x_1 + 6x_2 + 8x_3 &= 0 \\ 6x_1 + 34x_2 + 52x_3 &= -160 \\ 8x_1 + 52x_2 + 129x_3 &= -452 \end{aligned}$

9. $\begin{aligned} 0.01x_1 \qquad\qquad + 0.03x_3 &= 0.14 \\ 0.16x_2 + 0.08x_3 &= 0.16 \\ 0.03x_1 + 0.08x_2 + 0.14x_3 &= 0.54 \end{aligned}$

10. $\begin{aligned} 4x_1 \qquad + 2x_3 &= 1.5 \\ 4x_2 + x_3 &= 4.0 \\ 2x_1 + x_2 + 2x_3 &= 2.5 \end{aligned}$

11. $\begin{aligned} x_1 - x_2 + 3x_3 + 2x_4 &= 15 \\ -x_1 + 5x_2 - 5x_3 - 2x_4 &= -35 \\ 3x_1 - 5x_2 + 19x_3 + 3x_4 &= 94 \\ 2x_1 - 2x_2 + 3x_3 + 21x_4 &= 1 \end{aligned}$

12. $\begin{aligned} 4x_1 + 2x_2 + 4x_3 \qquad &= 20 \\ 2x_1 + 2x_2 + 3x_3 + 2x_4 &= 36 \\ 4x_1 + 3x_2 + 6x_3 + 3x_4 &= 60 \\ 2x_2 + 3x_3 + 9x_4 &= 122 \end{aligned}$

13. Definiteness. Let $\mathbf{A}$, $\mathbf{B}$ be $n \times n$ and positive definite. Are $-\mathbf{A}$, $\mathbf{A}^\mathsf{T}$, $\mathbf{A} + \mathbf{B}$, $\mathbf{A} - \mathbf{B}$ positive definite?

14. CAS PROJECT. Cholesky's Method. (a) Write a program for solving linear systems by Cholesky's method and apply it to Example 2 in the text, to Probs. 7–9, and to systems of your choice.

(b) Splines. Apply the factorization part of the program to the following matrices (as they occur in (9), Sec. 19.4 (with $c_j = 1$), in connection with splines).

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

15–19 INVERSE

Find the inverse by the Gauss–Jordan method, showing the details.

15. In Prob. 1

16. In Prob. 4

17. In Team Project 6(c)

18. In Prob. 9

19. In Prob. 12

20. Rounding. For the following matrix **A** find det **A**. What happens if you roundoff the given entries to **(a)** 5S, **(b)** 4S, **(c)** 3S, **(d)** 2S, **(e)** 1S? What is the practical implication of your work?

$$\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} & 2 \\ -\frac{1}{9} & 1 & \frac{1}{7} \\ \frac{4}{63} & -\frac{3}{28} & \frac{13}{49} \end{bmatrix}$$

20.3 Linear Systems: Solution by Iteration

The Gauss elimination and its variants in the last two sections belong to the **direct methods** for solving linear systems of equations; these are methods that give solutions after an amount of computation that can be specified in advance. In contrast, in an **indirect** or **iterative method** we start from an approximation to the true solution and, if successful, obtain better and better approximations from a computational cycle repeated as often as may be necessary for achieving a required accuracy, so that the amount of arithmetic depends upon the accuracy required and varies from case to case.

We apply iterative methods if the convergence is rapid (if matrices have large main diagonal entries, as we shall see), so that we save operations compared to a direct method. We also use iterative methods if a large system is **sparse**, that is, has very many zero coefficients, so that one would waste space in storing zeros, for instance, 9995 zeros per equation in a potential problem of 10^4 equations in 10^4 unknowns with typically only 5 nonzero terms per equation (more on this in Sec. 21.4).

Gauss–Seidel Iteration Method[4]

This is an iterative method of great practical importance, which we can simply explain in terms of an example.

EXAMPLE 1 Gauss–Seidel Iteration

We consider the linear system

$$\begin{aligned} x_1 - 0.25x_2 - 0.25x_3 \qquad\qquad &= 50 \\ -0.25x_1 + x_2 \qquad\qquad - 0.25x_4 &= 50 \\ -0.25x_1 \qquad\qquad + x_3 - 0.25x_4 &= 25 \\ - 0.25x_2 - 0.25x_3 + x_4 &= 25. \end{aligned} \tag{1}$$

[4]PHILIPP LUDWIG VON SEIDEL (1821–1896), German mathematician. For Gauss see footnote 5 in Sec. 5.4.

(Equations of this form arise in the numeric solution of PDEs and in spline interpolation.) We write the system in the form

$$
\begin{aligned}
x_1 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 50 \\
x_2 &= 0.25x_1 \phantom{{}+ 0.25x_2 + 0.25x_3} + 0.25x_4 + 50 \\
x_3 &= 0.25x_1 \phantom{{}+ 0.25x_2 + 0.25x_3} + 0.25x_4 + 25 \\
x_4 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 25.
\end{aligned}
\tag{2}
$$

These equations are now used for iteration; that is, we start from a (possibly poor) approximation to the solution, say $x_1^{(0)} = 100$, $x_2^{(0)} = 100$, $x_3^{(0)} = 100$, $x_4^{(0)} = 100$, and compute from (2) a perhaps better approximation

Use "old" values
("New" values here not yet available)

$$
\begin{aligned}
x_1^{(1)} &= \phantom{0.25x_1^{(1)} +{}} 0.25x_2^{(0)} + 0.25x_3^{(0)} && + 50.00 = 100.00 \\
x_2^{(1)} &= 0.25x_1^{(1)} \phantom{{}+ 0.25x_2^{(0)} + 0.25x_3^{(0)}} + 0.25x_4^{(0)} && + 50.00 = 100.00 \\
x_3^{(1)} &= 0.25x_1^{(1)} \phantom{{}+ 0.25x_2^{(0)} + 0.25x_3^{(0)}} + 0.25x_4^{(0)} && + 25.00 = 75.00 \\
x_4^{(1)} &= \phantom{0.25x_1^{(1)} +{}} 0.25x_2^{(1)} + 0.25x_3^{(1)} && + 25.00 = 68.75
\end{aligned}
\tag{3}
$$

Use "new" values

These equations (3) are obtained from (2) by substituting on the right the ***most recent*** approximation for each unknown. In fact, corresponding values replace previous ones as soon as they have been computed, so that in the second and third equations we use $x_1^{(1)}$ (not $x_1^{(0)}$), and in the last equation of (3) we use $x_2^{(1)}$ and $x_3^{(1)}$ (not $x_2^{(0)}$ and $x_3^{(0)}$). Using the same principle, we obtain in the next step

$$
\begin{aligned}
x_1^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(1)} + 0.25x_3^{(1)} && + 50.00 = 93.750 \\
x_2^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(1)} + 0.25x_3^{(1)}} + 0.25x_4^{(1)} && + 50.00 = 90.625 \\
x_3^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(1)} + 0.25x_3^{(1)}} + 0.25x_4^{(1)} && + 25.00 = 65.625 \\
x_4^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(2)} + 0.25x_3^{(2)} && + 25.00 = 64.062
\end{aligned}
$$

Further steps give the values

x_1	x_2	x_3	x_4
89.062	88.281	63.281	62.891
87.891	87.695	62.695	62.598
87.598	87.549	62.549	62.524
87.524	87.512	62.512	62.506
87.506	87.503	62.503	62.502

Hence convergence to the exact solution $x_1 = x_2 = 87.5$, $x_3 = x_4 = 62.5$ (verify!) seems rather fast. ■

An algorithm for the Gauss–Seidel iteration is shown in Table 20.2. To obtain the algorithm, let us derive the general formulas for this iteration.

We assume that $a_{jj} = 1$ for $j = 1, \cdots, n$. (Note that this can be achieved if we can rearrange the equations so that no diagonal coefficient is zero; then we may divide each equation by the corresponding diagonal coefficient.) We now write

$$\mathbf{A} = \mathbf{I} + \mathbf{L} + \mathbf{U} \qquad (a_{jj} = 1) \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{L}$ and $\mathbf{U}$ are, respectively, lower and upper triangular matrices with zero main diagonals. If we substitute (4) into $\mathbf{Ax} = \mathbf{b}$, we have

$$\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}.$$

Taking $\mathbf{Lx}$ and $\mathbf{Ux}$ to the right, we obtain, since $\mathbf{Ix} = \mathbf{x}$,

$$\mathbf{x} = \mathbf{b} - \mathbf{Lx} - \mathbf{Ux}. \tag{5}$$

Remembering from (3) in Example 1 that below the main diagonal we took "new" approximations and above the main diagonal "old" ones, we obtain from (5) the desired iteration formulas

$$\boxed{\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\overset{\text{“New”}}{\mathbf{x}^{(m+1)}} - \mathbf{U}\overset{\text{“Old”}}{\mathbf{x}^{(m)}}} \qquad (a_{jj} = 1) \tag{6}$$

where $\mathbf{x}^{(m)} = [x_j^{(m)}]$ is the mth approximation and $\mathbf{x}^{(m+1)} = [x_j^{(m+1)}]$ is the $(m + 1)$st approximation. In components this gives the formula in line 1 in Table 20.2. The matrix $\mathbf{A}$ must satisfy $a_{jj} \neq 0$ for all j. In Table 20.2 our assumption $a_{jj} = 1$ is no longer required, but is automatically taken care of by the factor $1/a_{jj}$ in line 1.

Table 20.2 Gauss–Seidel Iteration

ALGORITHM GAUSS–SEIDEL ($\mathbf{A}, \mathbf{b}, \mathbf{x}^{(0)}, \epsilon, N$)

This algorithm computes a solution $\mathbf{x}$ of the system $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$, where $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix with $a_{jj} \neq 0$, $j = 1, \cdots, n$.

INPUT: $\mathbf{A}, \mathbf{b}$, initial approximation $\mathbf{x}^{(0)}$, tolerance $\epsilon > 0$, maximum number of iterations N

OUTPUT: Approximate solution $\mathbf{x}^{(m)} = [x_j^{(m)}]$ or failure message that $\mathbf{x}^{(N)}$ does not satisfy the tolerance condition

For $m = 0, \cdots, N - 1$, do:

For $j = 1, \cdots, n$, do:

1 $$x_j^{(m+1)} = \frac{1}{a_{jj}}\left(b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(m+1)} - \sum_{k=j+1}^{n} a_{jk}x_k^{(m)}\right)$$

End

2 If $\max_j |x_j^{(m+1)} - x_j^{(m)}| < \epsilon\, |x_j^{(m+1)}|$ then OUTPUT $\mathbf{x}^{(m+1)}$. Stop

[*Procedure completed successfully*]

End

OUTPUT: "No solution satisfying the tolerance condition obtained after N iteration steps." Stop

[*Procedure completed unsuccessfully*]

End GAUSS–SEIDEL

Convergence and Matrix Norms

An iteration method for solving $\mathbf{Ax} = \mathbf{b}$ is said to **converge** for an initial $\mathbf{x}^{(0)}$ if the corresponding iterative sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ converges to a solution of the given system. Convergence depends on the relation between $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(m+1)}$. To get this relation for the Gauss–Seidel method, we use (6). We first have

$$(\mathbf{I} + \mathbf{L})\,\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(m)}$$

and by multiplying by $(\mathbf{I} + \mathbf{L})^{-1}$ from the left,

$$\mathbf{x}^{(m+1)} = \mathbf{C}\mathbf{x}^{(m)} + (\mathbf{I} + \mathbf{L})^{-1}\mathbf{b} \qquad \text{where} \qquad \mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}. \tag{7}$$

The Gauss–Seidel iteration converges for every $\mathbf{x}^{(0)}$ if and only if all the eigenvalues (Sec. 8.1) of the "iteration matrix" $\mathbf{C} = [c_{jk}]$ have absolute value less than 1. (Proof in Ref. [E5], p. 191, listed in App. 1.)

CAUTION! If you want to get $\mathbf{C}$, first divide the rows of $\mathbf{A}$ by a_{jj} to have main diagonal $1, \cdots, 1$. If the **spectral radius** of $\mathbf{C}$ (= maximum of those absolute values) is small, then the convergence is rapid.

Sufficient Convergence Condition. A sufficient condition for convergence is

$$\|\mathbf{C}\| < 1. \tag{8}$$

Here $\|\mathbf{C}\|$ is some **matrix norm**, such as

$$\|\mathbf{C}\| = \sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} c_{jk}^2} \qquad \textbf{(Frobenius norm)} \tag{9}$$

or the greatest of the sums of the $|c_{jk}|$ in a *column* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_k \sum_{j=1}^{n} |c_{jk}| \qquad \textbf{(Column "sum" norm)} \tag{10}$$

or the greatest of the sums of the $|c_{jk}|$ in a *row* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_j \sum_{k=1}^{n} |c_{jk}| \qquad \textbf{(Row "sum" norm)}. \tag{11}$$

These are the most frequently used matrix norms in numerics.

In most cases the choice of one of these norms is a matter of computational convenience. However, the following example shows that sometimes one of these norms is preferable to the others.

EXAMPLE 2 Test of Convergence of the Gauss–Seidel Iteration

Test whether the Gauss–Seidel iteration converges for the system

$$\begin{aligned} 2x + y + z &= 4 \\ x + 2y + z &= 4 \\ x + y + 2z &= 4 \end{aligned} \qquad \text{written} \qquad \begin{aligned} x &= 2 - \tfrac{1}{2}y - \tfrac{1}{2}z \\ y &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}z \\ z &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}y. \end{aligned}$$

Solution. The decomposition (multiply the matrix by $\frac{1}{2}$ – why?) is

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} = \mathbf{I} + \mathbf{L} + \mathbf{U} = \mathbf{I} + \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}.$$

It shows that

$$\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U} = -\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{8} & \frac{3}{8} \end{bmatrix}.$$

We compute the Frobenius norm of $\mathbf{C}$

$$\|\mathbf{C}\| = \left(\tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{64} + \tfrac{9}{64}\right)^{1/2} = \left(\tfrac{50}{64}\right)^{1/2} = 0.884 < 1$$

and conclude from (8) that this Gauss–Seidel iteration converges. It is interesting that the other two norms would permit no conclusion, as you should verify. Of course, this points to the fact that (8) is sufficient for convergence rather than necessary. ■

Residual. Given a system $\mathbf{Ax} = \mathbf{b}$, the **residual r** of $\mathbf{x}$ with respect to this system is defined by

$$\mathbf{r} = \mathbf{b} - \mathbf{Ax}. \tag{12}$$

Clearly, $\mathbf{r} = \mathbf{0}$ if and only if $\mathbf{x}$ is a solution. Hence $\mathbf{r} \neq \mathbf{0}$ for an approximate solution. In the Gauss–Seidel iteration, at each stage we modify or *relax* a component of an approximate solution in order to reduce a component of $\mathbf{r}$ to zero. Hence the Gauss–Seidel iteration belongs to a class of methods often called **relaxation methods**. More about the residual follows in the next section.

Jacobi Iteration

The Gauss–Seidel iteration is a method of **successive corrections** because for each component we successively replace an approximation of a component by a corresponding new approximation as soon as the latter has been computed. An iteration method is called a method of **simultaneous corrections** if no component of an approximation $\mathbf{x}^{(m)}$ is used until ***all*** the components of $\mathbf{x}^{(m)}$ have been computed. A method of this type is the **Jacobi iteration**, which is similar to the Gauss–Seidel iteration but involves *not* using improved values until a step has been completed and then replacing $\mathbf{x}^{(m)}$ by $\mathbf{x}^{(m+1)}$ at once, directly before the beginning of the next step. Hence if we write $\mathbf{Ax} = \mathbf{b}$ (***with*** $a_{jj} = 1$ ***as before!***) in the form $\mathbf{x} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}$, the Jacobi iteration in matrix notation is

$$\mathbf{x}^{(m+1)} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}^{(m)} \qquad (a_{jj} = 1). \tag{13}$$

This method converges for every choice of $\mathbf{x}^{(0)}$ if and only if the spectral radius of $\mathbf{I} - \mathbf{A}$ is less than 1. It has recently gained greater practical interest since on parallel processors all n equations can be solved simultaneously at each iteration step.

For Jacobi, see Sec. 10.3. For exercises, see the problem set.

PROBLEM SET 20.3

1. Verify the solution in Example 1 of the text.

2. Show that for the system in Example 2 the Jacobi iteration diverges. *Hint.* Use eigenvalues.

3. Verify the claim at the end of Example 2.

4–10 GAUSS–SEIDEL ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]^{\mathsf{T}}$ and using 6S in the computation. *Hint.* Make sure that you solve each equation for the variable that has the largest coefficient (why?). Show the details.

4.
$$\begin{aligned} 4x_1 - x_2 \phantom{{}-x_3} &= 21 \\ -x_1 + 4x_2 - x_3 &= -45 \\ -x_2 + 4x_3 &= 33 \end{aligned}$$

5.
$$\begin{aligned} 10x_1 + x_2 + x_3 &= 6 \\ x_1 + 10x_2 + x_3 &= 6 \\ x_1 + x_2 + 10x_3 &= 6 \end{aligned}$$

6.
$$\begin{aligned} x_2 + 7x_3 &= 25.5 \\ 5x_1 + x_2 &= 0 \\ x_1 + 6x_2 + x_3 &= -10.5 \end{aligned}$$

7.
$$\begin{aligned} 5x_1 - 2x_2 &= 18 \\ -2x_1 + 10x_2 - 2x_3 &= -60 \\ -2x_2 + 15x_3 &= 128 \end{aligned}$$

8.
$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 7 \\ x_1 + 3x_2 + 2x_3 &= 4 \\ 2x_1 + x_2 + 3x_3 &= 7 \end{aligned}$$

9.
$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 19 \\ x_1 + 4x_2 - 2x_3 &= -2 \\ 2x_1 + 3x_2 + 8x_3 &= 39 \end{aligned}$$

10.
$$\begin{aligned} 4x_1 + 5x_3 &= 12.5 \\ x_1 + 6x_2 + 2x_3 &= 18.5 \\ 8x_1 + 2x_2 + x_3 &= -11.5 \end{aligned}$$

11. Apply the Gauss–Seidel iteration (3 steps) to the system in Prob. 5, starting from **(a)** 0, 0, 0 **(b)** 10, 10, 10. Compare and comment.

12. In Prob. 5, compute $\mathbf{C}$ **(a)** if you solve the first equation for x_1, the second for x_2, the third for x_3, proving convergence; **(b)** if you nonsensically solve the third equation for x_1, the first for x_2, the second for x_3, proving divergence.

13. CAS Experiment. Gauss–Seidel Iteration. (a) Write a program for Gauss–Seidel iteration.

(b) Apply the program $\mathbf{A}(t)\mathbf{x} = \mathbf{b}$, to starting from $[0 \quad 0 \quad 0]^{\mathsf{T}}$, where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t & t \\ t & 1 & t \\ t & t & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

For $t = 0.2, 0.5, 0.8, 0.9$ determine the number of steps to obtain the exact solution to 6S and the corresponding spectral radius of $\mathbf{C}$. Graph the number of steps and the spectral radius as functions of t and comment.

(c) Successive overrelaxation (SOR). Show that by adding and subtracting $\mathbf{x}^{(m)}$ on the right, formula (6) can be written

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)} \qquad (a_{jj} = 1).$$

Anticipation of further corrections motivates the introduction of an **overrelaxation factor** $\omega > 1$ to get the **SOR formula for Gauss–Seidel**

$$\text{(14)} \qquad \begin{aligned} \mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} &+ \omega(\mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} \\ &- (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)}) \qquad (a_{jj} = 1) \end{aligned}$$

intended to give more rapid convergence. A recommended value is $\omega = 2/(1 + \sqrt{1 - \rho})$, where ρ is the spectral radius of $\mathbf{C}$ in (7). Apply SOR to the matrix in (b) for $t = 0.5$ and 0.8 and notice the improvement of convergence. (Spectacular gains are made with larger systems.)

14–17 JACOBI ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]$. Compare with the Gauss–Seidel iteration. Which of the two seems to converge faster? Show the details of your work.

14. The system in Prob. 4

15. The system in Prob. 9

16. The system in Prob. 10

17. Show convergence in Prob. 16 by verifying that $\mathbf{I} - \mathbf{A}$, where $\mathbf{A}$ is the matrix in Prob. 16 with the rows divided by the corresponding main diagonal entries, has the eigenvalues -0.519589 and $0.259795 \pm 0.246603i$.

18–20 NORMS

Compute the norms (9), (10), (11) for the following (square) matrices. Comment on the reasons for greater or smaller differences among the three numbers.

18. The matrix in Prob. 10

19. The matrix in Prob. 5

20. $\begin{bmatrix} 2k & -k & -k \\ k & -2k & k \\ -k & -k & 2k \end{bmatrix}$

20.4 Linear Systems: Ill-Conditioning, Norms

One does not need much experience to observe that some systems $\mathbf{Ax} = \mathbf{b}$ are good, giving accurate solutions even under roundoff or coefficient inaccuracies, whereas others are bad, so that these inaccuracies affect the solution strongly. We want to see what is going on and whether or not we can "trust" a linear system. Let us first formulate the two relevant concepts (ill- and well-conditioned) for general numeric work and then turn to linear systems and matrices.

A computational problem is called **ill-conditioned** (or *ill-posed*) if "small" changes in the data (the input) cause "large" changes in the solution (the output). On the other hand, a problem is called **well-conditioned** (or *well-posed*) if "small" changes in the data cause only "small" changes in the solution.

These concepts are qualitative. We would certainly regard a magnification of inaccuracies by a factor 100 as "large," but could debate where to draw the line between "large" and "small," depending on the kind of problem and on our viewpoint. Double precision may sometimes help, but if data are measured inaccurately, one should attempt ***changing the mathematical setting*** of the problem to a well-conditioned one.

Let us now turn to linear systems. Figure 445 explains that ill-conditioning occurs if and only if the two equations give two nearly parallel lines, so that their intersection point (the solution of the system) moves substantially if we raise or lower a line just a little. For larger systems the situation is similar in principle, although geometry no longer helps. We shall see that we may regard ill-conditioning as an approach to singularity of the matrix.

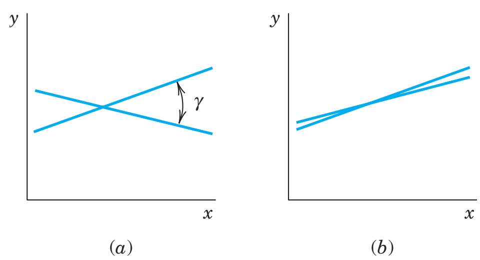


Fig. 445. (a) Well-conditioned and (b) ill-conditioned linear system of two equations in two unknowns

EXAMPLE 1 **An Ill-Conditioned System**

You may verify that the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \quad\quad y &= 1 \end{aligned}$$

has the solution $x = 0.5$, $y = -0.5$, whereas the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \quad\quad y &= 1 + \epsilon \end{aligned}$$

has the solution $x = 0.5 + 5000.5\epsilon$, $y = -0.5 + 4999.5\epsilon$. This shows that the system is ill-conditioned because a change on the right of magnitude ϵ produces a change in the solution of magnitude 5000ϵ, approximately. We see that the lines given by the equations have nearly the same slope.

Well-conditioning can be asserted if the main diagonal entries of $\mathbf{A}$ have large absolute values compared to those of the other entries. Similarly if $\mathbf{A}^{-1}$ and $\mathbf{A}$ have maximum entries of about the same absolute value.

Ill-conditioning is indicated if $\mathbf{A}^{-1}$ has entries of large absolute value compared to those of the solution (about 5000 in Example 1) and if poor approximate solutions may still produce small residuals.

Residual. The *residual* $\mathbf{r}$ of an approximate solution $\tilde{\mathbf{x}}$ of $\mathbf{Ax} = \mathbf{b}$ is defined as

$$\boxed{\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}.} \tag{1}$$

Now $\mathbf{b} = \mathbf{Ax}$, so that

$$\mathbf{r} = \mathbf{A}(\mathbf{x} - \mathbf{A}\tilde{\mathbf{x}}). \tag{2}$$

Hence $\mathbf{r}$ is small if $\tilde{\mathbf{x}}$ has high accuracy, but the converse may be false:

EXAMPLE 2 **Inaccurate Approximate Solution with a Small Residual**

The system

$$\begin{aligned} 1.0001x_1 + \quad\quad x_2 &= 2.0001 \\ x_1 + 1.0001x_2 &= 2.0001 \end{aligned}$$

has the exact solution $x_1 = 1$, $x_2 = 1$. Can you see this by inspection? The very inaccurate approximation $\tilde{x}_1 = 2.0000$, $\tilde{x}_2 = 0.0001$ has the very small residual (to 4D)

$$\mathbf{r} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 1.0001 & 1.0000 \\ 1.0000 & 1.0001 \end{bmatrix} \begin{bmatrix} 2.0000 \\ 0.0001 \end{bmatrix} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 2.0003 \\ 2.0001 \end{bmatrix} = \begin{bmatrix} -0.0002 \\ 0.0000 \end{bmatrix}.$$

From this, a naive person might draw the false conclusion that the approximation should be accurate to 3 or 4 decimals.

Our result is probably unexpected, but we shall see that it has to do with the fact that the system is ill-conditioned.

Our goal is to show that ill-conditioning of a linear system and of its coefficient matrix $\mathbf{A}$ can be measured by a number, the *condition number* $\kappa(\mathbf{A})$. Other measures for ill-conditioning have also been proposed, but $\kappa(\mathbf{A})$ is probably the most widely used one. $\kappa(\mathbf{A})$ is defined in terms of norm, a concept of great general interest throughout numerics (and in modern mathematics in general!). We shall reach our goal in three steps, discussing

1. **Vector norms**
2. **Matrix norms**
3. **Condition number** κ of a square matrix

Vector Norms

A **vector norm** for column vectors $\mathbf{x} = [x_j]$ with n components (n fixed) is a generalized length or distance. It is denoted by $\|\mathbf{x}\|$ and is defined by four properties of the usual length of vectors in three-dimensional space, namely,

$$
\text{(3)} \qquad
\begin{aligned}
&\text{(a)} \quad \|\mathbf{x}\| \text{ is a nonnegative real number.}\\
&\text{(b)} \quad \|\mathbf{x}\| = 0 \quad \text{if and only if} \quad \mathbf{x} = \mathbf{0}.\\
&\text{(c)} \quad \|k\mathbf{x}\| = |k|\,\|\mathbf{x}\| \qquad \text{for all } k.\\
&\text{(d)} \quad \|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad \text{(Triangle inequality).}
\end{aligned}
$$

If we use several norms, we label them by a subscript. Most important in connection with computations is the ***p*-norm** defined by

$$
\text{(4)} \qquad \|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}
$$

where p is a fixed number and $p \geqq 1$. In practice, one usually takes $p = 1$ or 2 and, as a third norm, $\|\mathbf{x}\|_\infty$ (the latter as defined below), that is,

$$
\textbf{(5)} \qquad \|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n| \qquad (\text{“}l_1\text{-norm”})
$$

$$
\textbf{(6)} \qquad \|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \qquad (\text{“Euclidean” or “}l_2\text{-norm”})
$$

$$
\textbf{(7)} \qquad \|\mathbf{x}\|_\infty = \max_j |x_j| \qquad (\text{“}l_\infty\text{-norm”}).
$$

For $n = 3$ the l_2-norm is the usual length of a vector in three-dimensional space. The l_1-norm and l_∞-norm are generally more convenient in computation. But all three norms are in common use.

EXAMPLE 3 Vector Norms

If $\mathbf{x}^\mathsf{T} = [2 \quad -3 \quad 0 \quad 1 \quad -4]$, then $\|\mathbf{x}\|_1 = 10$, $\|\mathbf{x}\|_2 = \sqrt{30}$, $\|\mathbf{x}\|_\infty = 4$. ■

In three-dimensional space, two points with position vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have distance $|\mathbf{x} - \tilde{\mathbf{x}}|$ from each other. For a linear system $\mathbf{Ax} = \mathbf{b}$, this suggests that we take $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ as a measure of inaccuracy and call it the **distance** between an exact and an approximate solution, or the **error** of $\tilde{\mathbf{x}}$.

Matrix Norm

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ any vector with n components, then $\mathbf{Ax}$ is a vector with n components. We now take a vector norm and consider $\|\mathbf{x}\|$ and $\|\mathbf{Ax}\|$. One can prove (see

Ref. [E17]. pp. 77, 92–93, listed in App. 1) that there is a number c (depending on $\mathbf{A}$) such that

$$\|\mathbf{Ax}\| \leqq c\|\mathbf{x}\| \qquad \text{for all } \mathbf{x}. \tag{8}$$

Let $\mathbf{x} \neq 0$. Then $\|\mathbf{x}\| > 0$ by (3b) and division gives $\|\mathbf{Ax}\|/\|\mathbf{x}\| \leqq c$. We obtain the smallest possible c valid for *all* $\mathbf{x}$ $(\neq \mathbf{0})$ by taking the maximum on the left. This smallest c is called the **matrix norm of A** *corresponding to the vector norm we picked* and is denoted by $\|\mathbf{A}\|$. Thus

$$\|\mathbf{A}\| = \max \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \qquad (\mathbf{x} \neq \mathbf{0}), \tag{9}$$

the maximum being taken over all $\mathbf{x} \neq \mathbf{0}$. Alternatively [see (c) in Team Project 24],

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|. \tag{10}$$

The maximum in (10) and thus also in (9) exists. And the name "matrix ***norm***" is justified because $\|\mathbf{A}\|$ satisfies (3) with $\mathbf{x}$ and $\mathbf{y}$ replaced by $\mathbf{A}$ and $\mathbf{B}$. (Proofs in Ref. [E17] pp. 77, 92–93.)

Note carefully that $\|\mathbf{A}\|$ depends on the vector norm that we selected. In particular, one can show that

for the l_1-norm (5) one gets the column "sum" norm (10), Sec. 20.3,

for the l_∞-norm (7) one gets the row "sum" norm (11), Sec. 20.3.

By taking our best possible (our smallest) $c = \|\mathbf{A}\|$ we have from (8)

$$\|\mathbf{Ax}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|. \tag{11}$$

This is the formula we shall need. Formula (9) also implies for two $n \times n$ matrices (see Ref. [E17], p. 98)

$$\|\mathbf{AB}\| \leqq \|\mathbf{A}\|\,\|\mathbf{B}\|, \qquad \text{thus} \qquad \|\mathbf{A}^n\| \leqq \|\mathbf{A}\|^n. \tag{12}$$

See Refs. [E9] and [E17] for other useful formulas on norms.

Before we go on, let us do a simple illustrative computation.

EXAMPLE 4 Matrix Norms

Compute the matrix norms of the coefficient matrix $\mathbf{A}$ in Example 1 and of its inverse $\mathbf{A}^{-1}$, assuming that we use (a) the l_1-vector norm, (b) the l_∞-vector norm.

Solution. We use (4*), Sec. 7.8, for the inverse and then (10) and (11) in Sec. 20.3. Thus

$$\mathbf{A} = \begin{bmatrix} 0.9999 & -1.0001 \\ 1.0000 & -1.0000 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} -5000.0 & 5000.5 \\ -5000.0 & 4999.5 \end{bmatrix}.$$

(a) The l_1-vector norm gives the column "sum" norm (10), Sec. 20.3; from Column 2 we thus obtain $\|\mathbf{A}\| = |-1.0001| + |-1.0000| = 2.0001$. Similarly, $\|\mathbf{A}^{-1}\| = 10{,}000$.

(b) The l_∞-vector norm gives the row "sum" norm (11), Sec. 20.3; thus $\|\mathbf{A}\| = 2$, $\|\mathbf{A}^{-1}\| = 10000.5$ from Row 1. We notice that $\|\mathbf{A}^{-1}\|$ is surprisingly large, which makes the product $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ large (20,001). We shall see below that this is typical of an ill-conditioned system.

Condition Number of a Matrix

We are now ready to introduce the key concept in our discussion of ill-conditioning, the **condition number** $\kappa(\mathbf{A})$ of a (nonsingular) square matrix $\mathbf{A}$, defined by

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|. \qquad \textbf{(13)}$$

The role of the condition number is seen from the following theorem.

THEOREM 1

Condition Number

A linear system of equations $\mathbf{Ax} = \mathbf{b}$ and its matrix $\mathbf{A}$ whose condition number (13) *is small are well-conditioned. A large condition number indicates ill-conditioning.*

PROOF $\mathbf{b} = \mathbf{Ax}$ and (11) give $\|\mathbf{b}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|$. Let $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. Then division by $\|\mathbf{b}\|\,\|\mathbf{x}\|$ gives

$$\frac{1}{\|\mathbf{x}\|} \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}. \qquad (14)$$

Multiplying (2) $\mathbf{r} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})$ by $\mathbf{A}^{-1}$ from the left and interchanging sides, we have $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{r}$. Now (11) with $\mathbf{A}^{-1}$ and $\mathbf{r}$ instead of $\mathbf{A}$ and $\mathbf{x}$ yields

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{A}^{-1}\mathbf{r}\| \leqq \|\mathbf{A}^{-1}\|\,\|\mathbf{r}\|.$$

Division by $\|\mathbf{x}\|$ [note that $\|\mathbf{x}\| \neq 0$ by (3b)] and use of (14) finally gives

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leqq \frac{1}{\|\mathbf{x}\|}\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| = \kappa(\mathbf{A})\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \qquad \textbf{(15)}$$

Hence if $\kappa(\mathbf{A})$ is small, a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ implies a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$, so that the system is well-conditioned. However, this does not hold if $\kappa(\mathbf{A})$ is large; then a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ does not necessarily imply a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$. ■

EXAMPLE 5 **Condition Numbers. Gauss–Seidel Iteration**

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 4 \end{bmatrix} \quad \text{has the inverse} \quad \mathbf{A}^{-1} = \frac{1}{56}\begin{bmatrix} 12 & -2 & -2 \\ -2 & 19 & -9 \\ -2 & -9 & 19 \end{bmatrix}.$$

Since $\mathbf{A}$ is symmetric, (10) and (11) in Sec. 20.3 give the same condition number

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| = 7 \cdot \tfrac{1}{56} \cdot 30 = 3.75.$$

We see that a linear system $\mathbf{Ax} = \mathbf{b}$ with this $\mathbf{A}$ is well-conditioned.

For instance, if $\mathbf{b} = [14 \quad 0 \quad 28]^\mathsf{T}$, the Gauss algorithm gives the solution $\mathbf{x} = [2 \quad -5 \quad 9]^\mathsf{T}$, (confirm this). Since the main diagonal entries of $\mathbf{A}$ are relatively large, we can expect reasonably good convergence of the Gauss–Seidel iteration. Indeed, starting from, say, $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$, we obtain the first 8 steps (3D values)

x_1	x_2	x_3
1.000	1.000	1.000
2.400	−1.100	6.950
1.630	−3.882	8.534
1.870	−4.734	8.900
1.967	−4.942	8.979
1.993	−4.988	8.996
1.998	−4.997	8.999
2.000	−5.000	9.000
2.000	−5.000	9.000

■

EXAMPLE 6 Ill-Conditioned Linear System

Example 4 gives by (10) or (11), Sec. 20.3, for the matrix in Example 1 the very large condition number $\kappa(\mathbf{A}) = 2.0001 \cdot 10000 = 2 \cdot 10000.5 = 200001$. This confirms that the system is very ill-conditioned.

Similarly in Example 2, where by (4*), Sec. 7.8 and 6D-computation,

$$\mathbf{A}^{-1} = \frac{1}{0.0002}\begin{bmatrix} 1.0001 & -1.0000 \\ -1.0000 & 1.0001 \end{bmatrix} = \begin{bmatrix} 5000.5 & -5.000.0 \\ -5000.0 & 5000.5 \end{bmatrix}$$

so that (10), Sec. 20.3, gives a very large $\kappa(\mathbf{A})$, explaining the surprising result in Example 2,

$$\kappa(\mathbf{A}) = (1.0001 + 1.0000)(5000.5 + 5000.0) \approx 20{,}002.$$

■

In practice, $\mathbf{A}^{-1}$ will not be known, so that in computing the condition number $\kappa(\mathbf{A})$, one must estimate $\|\mathbf{A}^{-1}\|$. A method for this (proposed in 1979) is explained in Ref. [E9] listed in App. 1.

Inaccurate Matrix Entries. $\kappa(\mathbf{A})$ can be used for estimating the effect $\delta\mathbf{x}$ of an inaccuracy $\delta\mathbf{A}$ of $\mathbf{A}$ (errors of measurements of the a_{jk}, for instance). Instead of $\mathbf{Ax} = \mathbf{b}$ we then have

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}.$$

Multiplying out and subtracting $\mathbf{Ax} = \mathbf{b}$ on both sides, we obtain

$$\mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{0}.$$

Multiplication by $\mathbf{A}^{-1}$ from the left and taking the second term to the right gives

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}).$$

Applying (11) with $\mathbf{A}^{-1}$ and vector $\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})$ instead of $\mathbf{A}$ and $\mathbf{x}$, we get

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\|.$$

Applying (11) on the right, with $\delta\mathbf{A}$ and $\mathbf{x} - \delta\mathbf{x}$ instead of $\mathbf{A}$ and $\mathbf{x}$, we obtain

$$\|\delta\mathbf{x}\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}\|\,\|\mathbf{x} + \delta\mathbf{x}\|.$$

Now $\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})/\|\mathbf{A}\|$ by the definition of $\kappa(\mathbf{A})$, so that division by $\|\mathbf{x} + \delta\mathbf{x}\|$ shows that the relative inaccuracy of $\mathbf{x}$ is related to that of $\mathbf{A}$ via the condition number by the inequality

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \approx \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x} + \delta\mathbf{x}\|} \leqq \|\mathbf{A}^{-1}\| \, \|\delta\mathbf{A}\| = \kappa(\mathbf{A}) \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}. \tag{16}$$

Conclusion. If the system is well-conditioned, small inaccuracies $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ can have only a small effect on the solution. However, in the case of ill-conditioning, if $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ is small, $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ ***may*** be large.

Inaccurate Right Side. You may show that, similarly, when $\mathbf{A}$ is accurate, an inaccuracy $\delta\mathbf{b}$ of $\mathbf{b}$ causes an inaccuracy $\delta\mathbf{x}$ satisfying

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \kappa(\mathbf{A}) \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{17}$$

Hence $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ must remain relatively small whenever $\kappa(\mathbf{A})$ is small.

EXAMPLE 7 Inaccuracies. Bounds (16) and (17)

If each of the nine entries of $\mathbf{A}$ in Example 5 is measured with an inaccuracy of 0.1, then $\|\delta\mathbf{A}\| = 9 \cdot 0.1$ and (16) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{3 \cdot 0.1}{7} = 0.321 \qquad \text{thus} \qquad \|\delta\mathbf{x}\| \leqq 0.321\,\|\mathbf{x}\| = 0.321 \cdot 16 = 5.14.$$

By experimentation you will find that the actual inaccuracy $\|\delta\mathbf{x}\|$ is only about 30% of the bound 5.14. This is typical.

Similarly, if $\delta\mathbf{b} = [0.1 \quad 0.1 \quad 0.1]^{\mathsf{T}}$, then $\|\delta\mathbf{b}\| = 0.3$ and $\|\mathbf{b}\| = 42$ in Example 5, so that (17) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{0.3}{42} = 0.0536, \qquad \text{hence} \qquad \|\delta\mathbf{x}\| \leqq 0.0536 \cdot 16 = 0.857$$

but this bound is again much greater than the actual inaccuracy, which is about 0.15. ■

Further Comments on Condition Numbers. The following additional explanations may be helpful.

1. There is no sharp dividing line between "well-conditioned" and "ill-conditioned," but generally the situation will get worse as we go from systems with small $\kappa(\mathbf{A})$ to systems with larger $\kappa(\mathbf{A})$. Now always $\kappa(\mathbf{A}) \geqq 1$, so that values of 10 or 20 or so give no reason for concern, whereas $\kappa(\mathbf{A}) = 100$, say, calls for caution, and systems such as those in Examples 1 and 2 are extremely ill-conditioned.

2. If $\kappa(\mathbf{A})$ is large (or small) in one norm, it will be large (or small, respectively) in any other norm. See Example 5.

3. The literature on ill-conditioning is extensive. For an introduction to it, see [E9].

This is the end of our discussion of numerics for solving linear systems. In the next section we consider curve fitting, an important area in which solutions are obtained from linear systems.

PROBLEM SET 20.4

1–6 VECTOR NORMS

Compute the norms (5), (6), (7). Compute a corresponding **unit vector** (vector of norm 1) with respect to the l_∞-norm.

1. $[1 \quad -3 \quad 8 \quad 0 \quad -6 \quad 0]$

2. $[4 \quad -1 \quad 8]$

3. $[0.2 \quad 0.6 \quad -2.1 \quad 3.0]$

4. $[k^2, \quad 4k, \quad k^3], \quad k > 4$

5. $[1 \quad 1 \quad 1 \quad 1 \quad 1]$

6. $[0 \quad 0 \quad 0 \quad 1 \quad 0]$

7. For what $\mathbf{x} = [a \quad b \quad c]$ will $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_2$?

8. Show that $\|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1$.

9–16 MATRIX NORMS, CONDITION NUMBERS

Compute the matrix norm and the condition number corresponding to the l_1-vector norm.

9. $\begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix}$

10. $\begin{bmatrix} 2.1 & 4.5 \\ 0.5 & 1.8 \end{bmatrix}$

11. $\begin{bmatrix} \sqrt{5} & 5 \\ 0 & -\sqrt{5} \end{bmatrix}$

12. $\begin{bmatrix} 7 & 6 \\ 6 & 5 \end{bmatrix}$

13. $\begin{bmatrix} -2 & 4 & -1 \\ -2 & 3 & 0 \\ 7 & -12 & 2 \end{bmatrix}$

14. $\begin{bmatrix} 1 & 0.01 & 0 \\ 0.01 & 1 & 0.01 \\ 0 & 0.01 & 1 \end{bmatrix}$

15. $\begin{bmatrix} -20 & 0 & 0 \\ 0 & 0.05 & 0 \\ 0 & 0 & 20 \end{bmatrix}$

16. $\begin{bmatrix} 21 & 10.5 & 7 & 5.25 \\ 10.5 & 7 & 5.25 & 4.2 \\ 7 & 5.25 & 4.2 & 3.5 \\ 5.25 & 4.2 & 3.5 & 3 \end{bmatrix}$

17. Verify (11) for $\mathbf{x} = [3 \quad 15 \quad -4]^\mathsf{T}$ taken with the l_∞-norm and the matrix in Prob. 13.

18. Verify (12) for the matrices in Probs. 9 and 10.

19–20 ILL-CONDITIONED SYSTEMS

Solve $\mathbf{Ax} = \mathbf{b}_1$, $\mathbf{Ax} = \mathbf{b}_2$. Compare the solutions and comment. Compute the condition number of $\mathbf{A}$.

19. $\mathbf{A} = \begin{bmatrix} 4.50 & 3.55 \\ 3.55 & 2.80 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 5.2 \\ 4.1 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 5.2 \\ 4.0 \end{bmatrix}$

20. $\mathbf{A} = \begin{bmatrix} 3.0 & 1.7 \\ 1.7 & 1.0 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 4.7 \\ 2.7 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 4.7 \\ 2.71 \end{bmatrix}$

21. Residual. For $\mathbf{Ax} = \mathbf{b}_1$ in Prob. 19 guess what the residual of $\tilde{\mathbf{x}} = [-10.0 \quad 14.1]^\mathsf{T}$, very poorly approximating $[-2 \quad 4]^\mathsf{T}$, might be. Then calculate and comment.

22. Show that $\kappa(\mathbf{A}) \geqq 1$ for the matrix norms (10), (11), Sec. 20.3, and $\kappa(\mathbf{A}) \geqq \sqrt{n}$ for the Frobenius norm (9), Sec. 20.3.

23. CAS EXPERIMENT. Hilbert Matrices. The 3×3 Hilbert matrix is

$$\mathbf{H}_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

The $n \times n$ Hilbert matrix is $\mathbf{H}_n = [h_{jk}]$, where $h_{jk} = 1/(j + k - 1)$. (Similar matrices occur in curve fitting by least squares.) Compute the condition number $\kappa(\mathbf{H}_n)$ for the matrix norm corresponding to the l_∞- (or l_1-) vector norm, for $n = 2, 3, \cdots, 6$ (or further if you wish). Try to find a formula that gives reasonable approximate values of these rapidly growing numbers.

Solve a few linear systems of your choice, involving an $\mathbf{H}_n$.

24. TEAM PROJECT. Norms. (a) Vector norms in our text are **equivalent**, that is, they are related by double inequalities; for instance,

$$\text{(18)} \qquad \begin{aligned} &\text{(a)} \quad \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1 \leqq n\|\mathbf{x}\|_\infty \\ &\text{(b)} \quad \frac{1}{n}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1. \end{aligned}$$

Hence if for some $\mathbf{x}$, one norm is large (or small), the other norm must also be large (or small). Thus in many investigations the particular choice of a norm is not essential. Prove (18).

(b) The Cauchy–Schwarz inequality is

$$|\mathbf{x}^\mathsf{T}\mathbf{y}| \leqq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

It is very important. (Proof in Ref. [GenRef7] listed in App. 1.) Use it to prove

(19a) $$\|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1 \leqq \sqrt{n}\,\|\mathbf{x}\|_2$$

(19b) $$\frac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1.$$

(c) Formula (10) is often more practical than (9). Derive (10) from (9).

(d) Matrix norms. Illustrate (11) with examples. Give examples of (12) with equality as well as with strict inequality. Prove that the matrix norms (10), (11) in Sec. 20.3 satisfy the *axioms of a norm*

$$\|\mathbf{A}\| \geqq \mathbf{0}.$$

$$\|\mathbf{A}\| = \mathbf{0} \text{ if and only if } \mathbf{A} = \mathbf{0},$$

$$\|k\mathbf{A}\| = |k|\,\|\mathbf{A}\|,$$

$$\|\mathbf{A} + \mathbf{B}\| \leqq \|\mathbf{A}\| + \|\mathbf{B}\|.$$

25. WRITING PROJECT. Norms and Their Use in This Section. Make a list of the most important of the many ideas covered in this section and write a two-page report on them.

20.5 Least Squares Method

Having discussed numerics for linear systems, we now turn to an important application, curve fitting, in which the solutions are obtained from linear systems.

In **curve fitting** we are given n points (pairs of numbers) $(x_1, y_1), \cdots, (x_n, y_n)$ and we want to determine a function $f(x)$ such that

$$f(x_1) \approx y_1, \cdots, f(x_n) \approx y_n,$$

approximately. The type of function (for example, polynomials, exponential functions, sine and cosine functions) may be suggested by the nature of the problem (the underlying physical law, for instance), and in many cases a polynomial of a certain degree will be appropriate.

Let us begin with a motivation.

If we require strict equality $f(x_1) = y_1, \cdots, f(x_n) = y_n$ and use polynomials of sufficiently high degree, we may apply one of the methods discussed in Sec. 19.3 in connection with interpolation. However, in certain situations this would not be the appropriate solution of the actual problem. For instance, to the four points

(1) $$(-1.3, 0.103), \quad (-0.1, 1.099), \quad (0.2, 0.808), \quad (1.3, 1.897)$$

there corresponds the interpolation polynomial $f(x) = x^3 - x + 1$ (Fig. 446), but if we graph the points, we see that they lie nearly on a straight line. Hence if these values are obtained in an experiment and thus involve an experimental error, and if the nature of the experiment suggests a linear relation, we better fit a straight line through the points (Fig. 446). Such a line may be useful for predicting values to be expected for other values of x. A widely used principle for fitting straight lines is the **method**

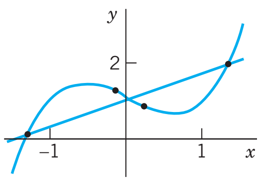


Fig. 446. Approximate fitting of a straight line

of least squares by Gauss and Legendre. In the present situation it may be formulated as follows.

Method of Least Squares. *The straight line*

$$y = a + bx \tag{2}$$

should be fitted through the given points $(x_1, y_1), \cdots, (x_n, y_n)$ *so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction).*

The point on the line with abscissa x_j has the ordinate $a + bx_j$. Hence its distance from (x_j, y_j) is $|y_j - a - bx_j|$ (Fig. 447) and that sum of squares is

$$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

q depends on a and b. A necessary condition for q to be minimum is

$$\begin{aligned} \frac{\partial q}{\partial a} &= -2 \sum (y_j - a - bx_j) = 0 \\ \frac{\partial q}{\partial b} &= -2 \sum x_j (y_j - a - bx_j) = 0 \end{aligned} \tag{3}$$

(where we sum over j from 1 to n). Dividing by 2, writing each sum as three sums, and taking one of them to the right, we obtain the result

$$\begin{aligned} an \quad &+ b \sum x_j = \sum y_j \\ a \sum x_j &+ b \sum x_j^2 = \sum x_j y_j. \end{aligned} \tag{4}$$

These equations are called the **normal equations** of our problem.

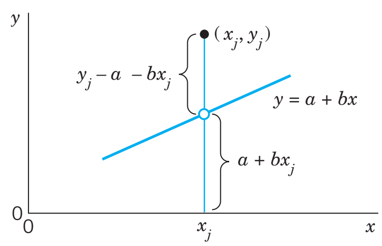


Fig. 447. Vetrical distance of a point (x_j, y_j) from a straight line $y = a + bx$

EXAMPLE 1 Straight Line

Using the method of least squares, fit a straight line to the four points given in formula (1).

Solution. We obtain

$$n = 4, \qquad \sum x_j = 0.1, \qquad \sum x_j^2 = 3.43, \qquad \sum y_j = 3.907, \qquad \sum x_j y_j = 2.3839.$$

Hence the normal equations are

$$4a + 0.10b = 3.9070$$
$$0.1a + 3.43b = 2.3839.$$

The solution (rounded to 4D) is $a = 0.9601$, $b = 0.6670$, and we obtain the straight line (Fig. 446)

$$y = 0.9601 + 0.6670x.$$

Curve Fitting by Polynomials of Degree m

Our method of curve fitting can be generalized from a polynomial $y = a + bx$ to a polynomial of degree m

$$p(x) = b_0 + b_1 x + \cdots + b_m x^m \tag{5}$$

where $m \leqq n - 1$. Then q takes the form

$$q = \sum_{j=1}^{n} (y_j - p(x_j))^2$$

and depends on $m + 1$ parameters $b_0, \cdots, b_m$. Instead of (3) we then have $m + 1$ conditions

$$\frac{\partial q}{\partial b_0} = 0, \qquad \cdots, \qquad \frac{\partial q}{\partial b_m} = 0 \tag{6}$$

which give a system of $m + 1$ normal equations.

In the case of a quadratic polynomial

$$p(x) = b_0 + b_1 x + b_2 x^2 \tag{7}$$

the normal equations are (summation from 1 to n)

$$\begin{aligned} b_0 n \quad &+ b_1 \sum x_j + b_2 \sum x_j^2 = \sum y_j \\ b_0 \sum x_j &+ b_1 \sum x_j^2 + b_2 \sum x_j^3 = \sum x_j y_j \\ b_0 \sum x_j^2 &+ b_1 \sum x_j^3 + b_2 \sum x_j^4 = \sum x_j^2 y_j. \end{aligned} \tag{8}$$

The derivation of (8) is left to the reader.

EXAMPLE 2 Quadratic Parabola by Least Squares

Fit a parabola through the data (0, 5), (2, 4), (4, 1), (6, 6), (8, 7).

Solution. For the normal equations we need $n = 5$, $\Sigma x_j = 20$, $\Sigma x_j^2 = 120$, $\Sigma x_j^3 = 800$, $\Sigma x_j^4 = 5664$, $\Sigma y_j = 23$, $\Sigma x_j y_j = 104$, $\Sigma x_j^2 y_j = 696$. Hence these equations are

$$\begin{aligned} 5b_0 + 20b_1 + 120b_2 &= 23 \\ 20b_0 + 120b_1 + 800b_2 &= 104 \\ 120b_0 + 800b_1 + 5664b_2 &= 696. \end{aligned}$$

Solving them we obtain the quadratic least squares parabola (Fig. 448)

$$y = 5.11429 - 1.41429x + 0.21429x^2.$$

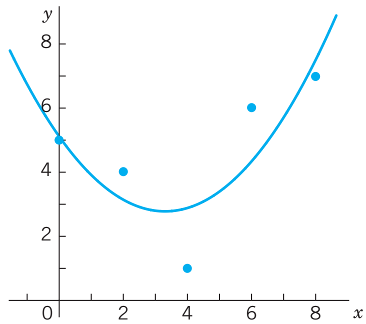


Fig. 448. Least squares parabola in Example 2

For a general polynomial (5) the normal equations form a linear system of equations in the unknowns $b_0, \cdots, b_m$. When its matrix $\mathbf{M}$ is nonsingular, we can solve the system by Cholesky's method (Sec. 20.2) because then $\mathbf{M}$ is positive definite (and symmetric). When the equations are nearly linearly dependent, the normal equations may become ill-conditioned and should be replaced by other methods; see [E5], Sec. 5.7, listed in App. 1.

The least squares method also plays a role in statistics (see Sec. 25.9).

PROBLEM SET 20.5

1–6 FITTING A STRAIGHT LINE

Fit a straight line to the given points (x, y) by least squares. Show the details. Check your result by sketching the points and the line. Judge the goodness of fit.

1. $(0, 2)$, $(2, 0)$, $(3, -2)$, $(5, -3)$
2. How does the line in Prob. 1 change if you add a point far above it, say, $(1, 3)$? Guess first.
3. $(0, 1.8)$, $(1, 1.6)$, $(2, 1.1)$, $(3, 1.5)$, $(4, 2.3)$
4. **Hooke's law $F = ks$.** Estimate the spring modulus k from the force F [lb] and the elongation s [cm], where $(F, s) = (1, 0.3)$, $(2, 0.7)$, $(4, 1.3)$, $(6, 1.9)$, $(10, 3.2)$, $(20, 6.3)$.
5. **Average speed.** Estimate the average speed v_{av} of a car traveling according to $s = v \cdot t$ [km] (s = distance traveled, t [hr] = time) from $(t, s) = (9, 140)$, $(10, 220)$, $(11, 310)$, $(12, 410)$.
6. **Ohm's law $U = Ri$.** Estimate R from $(i, U) = (2, 104)$, $(4, 206)$, $(6, 314)$, $(10, 530)$.
7. Derive the normal equations (8).

8–11 FITTING A QUADRATIC PARABOLA

Fit a parabola (7) to the points (x, y). Check by sketching.

8. $(-1, 5)$, $(1, 3)$, $(2, 4)$, $(3, 8)$
9. $(2, -3)$, $(3, 0)$, $(5, 1)$, $(6, 0)$ $(7, -2)$
10. t [hr] = Worker's time on duty, y [sec] = His/her reaction time, $(t, y) = (1, 2.0)$, $(2, 1.78)$, $(3, 1.90)$, $(4, 2.35)$, $(5, 2.70)$
11. The data in Prob. 3. Plot the points, the line, and the parabola jointly. Compare and comment.
12. **Cubic parabola.** Derive the formula for the normal equations of a cubic least squares parabola.
13. Fit curves (2) and (7) and a cubic parabola by least squares to $(x, y) = (-2, -30)$, $(-1, -4)$, $(0, 4)$, $(1, 4)$, $(2, 22)$, $(3, 68)$. Graph these curves and the points on common axes. Comment on the goodness of fit.
14. **TEAM PROJECT.** The **least squares approximation of a function** $f(x)$ on an interval $a \leqq x \leqq b$ by a function

$$F_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots + a_m y_m(x)$$

where $y_0(x), \cdots, y_m(x)$ are given functions, requires the determination of the coefficients $a_0, \cdots, a_m$ such that

$$\int_a^b [f(x) - F_m(x)]^2 \, dx \tag{9}$$

becomes minimum. This integral is denoted by $\|f - F_m\|^2$, and $\|f - F_m\|$ is called the **L_2-norm** of $f - F_m$ (L suggesting Lebesgue[5]). A necessary condition for that minimum is given by $\partial\|f - F_m\|^2/\partial a_j = 0$, $j = 0, \cdots, m$ [the analog of (6)]. **(a)** Show that this leads to $m + 1$ normal equations ($j = 0, \cdots, m$)

$$\sum_{k=0}^{m} h_{jk} a_k = b_j \qquad \text{where}$$

$$h_{jk} = \int_a^b y_j(x) y_k(x) \, dx, \tag{10}$$

$$b_j = \int_a^b f(x) y_j(x) \, dx.$$

(b) Polynomial. What form does (10) take if $F_m(x) = a_0 + a_1 x + \cdots + a_m x^m$? What is the coefficient matrix of (10) in this case when the interval is $0 \leqq x \leqq 1$?

(c) Orthogonal functions. What are the solutions of (10) if $y_0(x), \cdots, y_m(x)$ are orthogonal on the interval $a \leqq x \leqq b$? (For the definition, see Sec. 11.5. See also Sec. 11.6.)

15. CAS EXPERIMENT. Least Squares versus Interpolation. For the given data and for data of your choice find the interpolation polynomial and the least squares approximations (linear, quadratic, etc.). Compare and comment.

(a) $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$

(b) $(-4, 0)$, $(-3, 0)$, $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$, $(3, 0)$, $(4, 0)$

(c) Choose five points on a straight line, e.g., $(0, 0)$, $(1, 1), \cdots, (4, 4)$. Move one point 1 unit upward and find the quadratic least squares polynomial. Do this for each point. Graph the five polynomials on common axes. Which of the five motions has the greatest effect?

20.6 Matrix Eigenvalue Problems: Introduction

We now come to the second part of our chapter on numeric linear algebra. In the *first part of this chapter* we discussed methods of solving systems of linear equations, which included Gauss elimination with backward substitution. This method is known as a direct method since it gives solutions after a prescribed amount of computation. The Gauss method was modified by Doolittle's method, Crout's method, and Cholesky's method, each requiring fewer arithmetic operations than Gauss. Finally we presented indirect methods of solving systems of linear equations, that is, the Gauss–Seidel method and the Jacobi iteration. The indirect methods require an undetermined number of iterations. That number depends on how far we start from the true solution and what degree of accuracy we require. Moreover, depending on the problem, convergence may be fast or slow or our computation cycle might not even converge. This led to the concepts of ill-conditioned problems and condition numbers that help us gain some control over difficulties inherent in numerics.

The second part of this chapter deals with some of the most important ideas and numeric methods for matrix eigenvalue problems. This very extensive part of numeric linear algebra is of great practical importance, with much research going on, and hundreds, if not thousands, of papers published in various mathematical journals (see the references in [E8], [E9], [E11], [E29]). We begin with the concepts and general results we shall need in explaining and applying numeric methods for eigenvalue problems. (For typical models of eigenvalue problems see Chap. 8.)

[5] HENRI LEBESGUE (1875–1941), great French mathematician, creator of a modern theory of measure and integration in his famous doctoral thesis of 1902.

An **eigenvalue** or **characteristic value** (or *latent root*) of a given $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is a real or complex number λ such that the vector equation

$$\mathbf{(1)} \qquad \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

has a nontrivial solution, that is, a solution $\mathbf{x} \neq \mathbf{0}$, which is then called an **eigenvector** or **characteristic vector** of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. Equation (1) can be written

$$(2) \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix. This homogeneous system has a nontrivial solution if and only if the **characteristic determinant** det $(\mathbf{A} - \lambda\mathbf{I})$ is 0 (see Theorem 2 in Sec. 7.5). This gives (see Sec. 8.1)

THEOREM 1

Eigenvalues

The eigenvalues of $\mathbf{A}$ *are the solutions* λ *of the* **characteristic equation**

$$(3) \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Developing the characteristic determinant, we obtain the **characteristic polynomial** of $\mathbf{A}$, which is of degree n in λ. Hence $\mathbf{A}$ has at least one and at most n numerically different eigenvalues. If $\mathbf{A}$ is real, so are the coefficients of the characteristic polynomial. By familiar algebra it follows that then the roots (the eigenvalues of $\mathbf{A}$) are ***real or complex conjugates*** in pairs.

To give you some orientation of the underlying approaches of numerics for eigenvalue problems, note the following. For large or very large matrices it may be very difficult to determine the eigenvalues, since, in general, it is difficult to find the roots of characteristic polynomials of higher degrees. We will discuss different numeric methods for finding eigenvalues that achieve different results. Some methods, such as in Sec. 20.7, will give us only regions in which complex eigenvalues lie (Geschgorin's method) or the intervals in which the largest and smallest real eigenvalue lie (Collatz method). Other methods compute all eigenvalues, such as the Householder tridiagonalization method and the QR-method in Sec. 20.9.

To continue our discussion, we shall usually denote the eigenvalues of $\mathbf{A}$ by

$$\lambda_1, \lambda_2, \cdots, \lambda_n$$

with the understanding that some (or all) of them may be equal.

The sum of these n eigenvalues equals the sum of the entries on the main diagonal of $\mathbf{A}$, called the trace of $\mathbf{A}$; thus

$$(4) \qquad \text{trace } \mathbf{A} = \sum_{j=1}^{n} a_{jj} = \sum_{k=1}^{n} \lambda_k.$$

Also, the product of the eigenvalues equals the determinant of $\mathbf{A}$,

$$\det \mathbf{A} = \lambda_1 \lambda_2 \cdots \lambda_n. \tag{5}$$

Both formulas follow from the product representation of the characteristic polynomial, which we denote by $f(\lambda)$,

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n).$$

If we take equal factors together and denote the *numerically distinct* eigenvalues of $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_r$ $(r \leqq n)$, then the product becomes

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_r)^{m_r}. \tag{6}$$

The exponent m_j is called the **algebraic multiplicity** of λ_j. The maximum number of linearly independent eigenvectors corresponding to λ_j is called the **geometric multiplicity** of λ_j. It is equal to or smaller than m_j.

A subspace S of R^n or C^n (if $\mathbf{A}$ is complex) is called an **invariant subspace** of $\mathbf{A}$ if for every $\mathbf{v}$ in S the vector $\mathbf{Av}$ is also in S. **Eigenspaces** of $\mathbf{A}$ (spaces of eigenvectors; Sec. 8.1) are important invariant subspaces of $\mathbf{A}$.

An $n \times n$ matrix $\mathbf{B}$ is called **similar** to $\mathbf{A}$ if there is a nonsingular $n \times n$ matrix $\mathbf{T}$ such that

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{AT}. \tag{7}$$

Similarity is important for the following reason.

THEOREM 2

Similar Matrices

Similar matrices have the same eigenvalues. If $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$, *then* $\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$ *is an eigenvector of* $\mathbf{B}$ *in* (7) *corresponding to the same eigenvalue.* (Proof in Sec. 8.4.)

Another theorem that has various applications in numerics is as follows.

THEOREM 3

Spectral Shift

If $\mathbf{A}$ *has the eigenvalues* $\lambda_1, \cdots, \lambda_n$, *then* $\mathbf{A} - k\mathbf{I}$ *with arbitrary k has the eigenvalues* $\lambda_1 - k, \cdots, \lambda_n - k$.

This theorem is a special case of the following **spectral mapping theorem**.

THEOREM 4

Polynomial Matrices

If λ *is an eigenvalue of* $\mathbf{A}$, *then*

$$q(\lambda) = \alpha_s \lambda^s + \alpha_{s-1}\lambda^{s-1} + \cdots + \alpha_1 \lambda + \alpha_0$$

is an eigenvalue of the **polynomial matrix**

$$q(\mathbf{A}) = \alpha_s \mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots + \alpha_1 \mathbf{A} + a_0 \mathbf{I}.$$

PROOF $\mathbf{Ax} = \lambda\mathbf{x}$ implies $\mathbf{A}^2\mathbf{x} = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, $\mathbf{A}^3\mathbf{x} = \lambda^3\mathbf{x}$, etc. Thus

$$\begin{aligned} q(\mathbf{A})\mathbf{x} &= (\alpha_s\mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots)\,\mathbf{x} \\ &= \alpha_s\mathbf{A}^s\mathbf{x} + \alpha_{s-1}A^{s-1}\mathbf{x} + \cdots \\ &= \alpha_s\lambda^s\mathbf{x} + \alpha_{s-1}\lambda^{s-1}\mathbf{x} + \cdots = q(\lambda)\,\mathbf{x}. \end{aligned}$$

The eigenvalues of important special matrices can be characterized as follows.

THEOREM 5

Special Matrices

The eigenvalues of Hermitian matrices (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$), *hence of real symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}$), *are real. The eigenvalues of skew-Hermitian matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$), *hence of real skew-symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = -\mathbf{A}$), *are pure imaginary or* 0. *The eigenvalues of unitary matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$), *hence of orthogonal matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$), *have absolute value* 1. (Proofs in Secs. 8.3 and 8.5.)

The **choice of a numeric method** for matrix eigenvalue problems depends essentially on two circumstances, on the kind of matrix (real symmetric, real general, complex, sparse, or full) and on the kind of information to be obtained, that is, whether one wants to know all eigenvalues or merely specific ones, for instance, the largest eigenvalue, whether eigenvalues *and* eigenvectors are wanted, and so on. It is clear that we cannot enter into a systematic discussion of all these and further possibilities that arise in practice, but we shall concentrate on some basic aspects and methods that will give us a general understanding of this fascinating field.

20.7 Inclusion of Matrix Eigenvalues

The whole of numerics for matrix eigenvalues is motivated by the fact that, except for a few trivial cases, we cannot determine eigenvalues *exactly* by a finite process because these values are the roots of a polynomial of nth degree. Hence we must mainly use iteration.

In this section we state a few general theorems that give approximations and error bounds for eigenvalues. Our matrices will continue to be real (except in formula (5) below), but since (nonsymmetric) matrices may have complex eigenvalues, complex numbers will play a (very modest) role in this section.

The important theorem by Gerschgorin gives a region consisting of closed circular disks in the complex plane and including all the eigenvalues of a given matrix. Indeed, for each $j = 1, \cdots, n$ the inequality (1) in the theorem determines a closed circular disk in the complex λ-plane with center a_{jj} and radius given by the right side of (1); and Theorem 1 states that each of the eigenvalues of $\mathbf{A}$ lies in one of these n disks.

THEOREM 1

Gerschgorin's Theorem[6]

Let λ be an eigenvalue of an arbitrary $n \times n$ matrix $\mathbf{A} = [a_{jk}]$. Then for some integer j ($1 \leqq j \leqq n$) we have

$$(1) \qquad |a_{jj} - \lambda| \leqq |a_{j1}| + |a_{j2}| + \cdots + |a_{j,j-1}| + |a_{j,j+1}| + \cdots + |a_{jn}|.$$

[6]SEMYON ARANOVICH GERSCHGORIN (1901–1933), Russian mathematician.

PROOF Let $\mathbf{x}$ be an eigenvector corresponding to an eigenvalue λ of $\mathbf{A}$. Then

$$\text{(2)} \qquad \mathbf{Ax} = \lambda\mathbf{x} \qquad \text{or} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Let x_j be a component of $\mathbf{x}$ that is largest in absolute value. Then we have $|x_m/x_j| \leqq 1$ for $m = 1, \cdots, n$. The vector equation (2) is equivalent to a system of n equations for the n components of the vectors on both sides. The jth of these n equations with j as just indicated is

$$a_{j1}x_1 + \cdots + a_{j,j-1}x_{j-1} + (a_{jj} - \lambda)x_j + a_{j,j+1}x_{j+1} + \cdots + a_{jn}x_n = 0.$$

Division by x_j (which cannot be zero; why?) and reshuffling terms gives

$$a_{jj} - \lambda = -a_{j1}\frac{x_1}{x_j} - \cdots - a_{j,j-1}\frac{x_{j-1}}{x_j} - a_{j,j+1}\frac{x_{j+1}}{x_j} - \cdots - a_{jn}\frac{x_n}{x_j}.$$

By taking absolute values on both sides of this equation, applying the triangle inequality $|a + b| \leqq |a| + |b|$ (where a and b are any complex numbers), and observing that because of the choice of j (which is crucial!), $|x_1/x_j| \leqq 1, \cdots, |x_n/x_j| \leqq 1$, we obtain (1), and the theorem is proved. ■

EXAMPLE 1 **Gerschgorin's Theorem**

For the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 5 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix}$$

we get the Gerschgorin disks (Fig. 449)

D_1: Center 0, radius 1, D_2: Center 5, radius 1.5, D_3: Center 1, radius 1.5.

The centers are the main diagonal entries of $\mathbf{A}$. These would be the eigenvalues of $\mathbf{A}$ if $\mathbf{A}$ were diagonal. We can take these values as crude approximations of the unknown eigenvalues (3D-values) $\lambda_1 = -0.209$, $\lambda_2 = 5.305$, $\lambda_3 = 0.904$ (verify this); then the radii of the disks are corresponding error bounds.

Since $\mathbf{A}$ is symmetric, it follows from Theorem 5, Sec. 20.6, that the spectrum of $\mathbf{A}$ must actually lie in the intervals $[-1, 2.5]$ and $[3.5, 6.5]$.

It is interesting that here the Gerschgorin disks form two disjoint sets, namely, $D_1 \cup D_3$, which contains two eigenvalues, and D_2, which contains one eigenvalue. This is typical, as the following theorem shows. ■

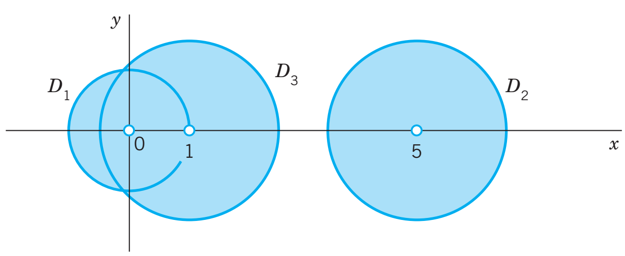


Fig. 449. Gerschgorin disks in Example 1

THEOREM 2

Extension of Gerschgorin's Theorem

If p Gerschgorin disks form a set S that is disjoint from the $n - p$ other disks of a given matrix $\mathbf{A}$, *then S contains precisely p eigenvalues of* $\mathbf{A}$ (*each counted with its algebraic multiplicity, as defined in Sec.* 20.6).

Idea of Proof. Set $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{B}$ is the diagonal matrix with entries a_{jj}, and apply Theorem 1 to $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$ with real t growing from 0 to 1. ■

EXAMPLE 2

Another Application of Gerschgorin's Theorem. Similarity

Suppose that we have diagonalized a matrix by some numeric method that left us with some off-diagonal entries of size 10^{-5}, say,

$$\mathbf{A} = \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix}.$$

What can we conclude about deviations of the eigenvalues from the main diagonal entries?

Solution. By Theorem 2, one eigenvalue must lie in the disk of radius $2 \cdot 10^{-5}$ centered at 4 and two eigenvalues (or an eigenvalue of algebraic multiplicity 2) in the disk of radius $2 \cdot 10^{-5}$ centered at 2. Actually, since the matrix is symmetric, these eigenvalues must lie in the intersections of these disks and the real axis, by Theorem 5 in Sec. 20.6.

We show how an isolated disk can always be reduced in size by a similarity transformation. The matrix

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix} \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{5} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 10^{-5} & 1 \\ 10^{-5} & 2 & 1 \\ 10^{-10} & 10^{-10} & 4 \end{bmatrix}$$

is similar to $\mathbf{A}$. Hence by Theorem 2, Sec. 20.6, it has the same eigenvalues as $\mathbf{A}$. From Row 3 we get the smaller disk of radius $2 \cdot 10^{-10}$. Note that the other disks got bigger, approximately by a factor of 10^5. And in choosing $\mathbf{T}$ we have to watch that the new disks do not overlap with the disk whose size we want to decrease.

For further interesting facts, see the book [E28]. ■

By definition, a **diagonally dominant** matrix $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix such that

$$|a_{jj}| \geqq \sum_{k \neq j} |a_{jk}| \qquad j = 1, \cdots, n \tag{3}$$

where we sum over all off-diagonal entries in Row j. The matrix is said to be **strictly diagonally dominant** if $>$ in (3) for all j. Use Theorem 1 to prove the following basic property.

THEOREM 3

Strict Diagonal Dominance

Strictly diagonally dominant matrices are nonsingular.

Further Inclusion Theorems

An **inclusion theorem** is a theorem that specifies a set which contains at least one eigenvalue of a given matrix. Thus, Theorems 1 and 2 are inclusion theorems; they even include the whole spectrum. We now discuss some famous theorems that yield further inclusions of eigenvalues. We state the first two of them without proofs (which would exceed the level of this book).

THEOREM 4

Schur's Theorem[7]

Let $\mathbf{A} = [a_{jk}]$ *be a* $n \times n$ *matrix. Then for each of its eigenvalues* $\lambda_1, \cdots, \lambda_n$,

$$|\lambda_m|^2 \leqq \sum_{i=1}^{n} |\lambda_i|^2 \leqq \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \quad \textbf{(Schur's inequality)}. \tag{4}$$

In (4) *the second equality sign holds if and only if* $\mathbf{A}$ *is such that*

$$\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A} = \mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}}. \tag{5}$$

Matrices that satisfy (5) are called **normal matrices**. It is not difficult to see that Hermitian, skew-Hermitian, and unitary matrices are normal, and so are real symmetric, skew-symmetric, and orthogonal matrices.

EXAMPLE 3 **Bounds for Eigenvalues Obtained from Schur's Inequality**

For the matrix

$$\mathbf{A} = \begin{bmatrix} 26 & -2 & 2 \\ 2 & 21 & 4 \\ 4 & 2 & 28 \end{bmatrix}$$

we obtain from Schur's inequality $|\lambda| \leqq \sqrt{1949} = 44.1475$. You may verify that the eigenvalues are 30, 25, and 20. Thus $30^2 + 25^2 + 20^2 = 1925 < 1949$; in fact, $\mathbf{A}$ is not normal. ■

The preceding theorems are valid for every real or complex square matrix. Other theorems hold for special classes of matrices only. Famous is the following one, which has various applications, for instance, in economics.

THEOREM 5

Perron's Theorem[8]

Let $\mathbf{A}$ *be a real* $n \times n$ *matrix whose entries are all positive. Then* $\mathbf{A}$ *has a positive real eigenvalue* $\lambda = \rho$ *of multiplicity* 1. *The corresponding eigenvector can be chosen with all components positive.* (*The other eigenvalues are less than* ρ *in absolute value.*)

[7]ISSAI SCHUR (1875–1941), German mathematician, also known by his important work in group theory.

[8]OSKAR PERRON (1880–1975) and GEORG FROBENIUS (1849–1917), German mathematicians, known for their work in potential theory, ODEs (Sec. 5.4), and group theory.

For a proof see Ref. [B3], vol. II, pp. 53–62. The theorem also holds for matrices with *nonnegative* real entries (“**Perron–Frobenius Theorem**”[8]) provided **A** is **irreducible**, that is, it cannot be brought to the following form by interchanging rows and columns; here **B** and **F** are square and **0** is a zero matrix.

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}$$

Perron's theorem has various applications, for instance, in economics. It is interesting that one can obtain from it a theorem that gives a numeric algorithm:

THEOREM 6

Collatz Inclusion Theorem[9]

Let $\mathbf{A} = [a_{jk}]$ *be a real* $n \times n$ *matrix whose elements are all positive. Let* $\mathbf{x}$ *be any real vector whose components* $x_1, \cdots, x_n$ *are positive, and let* $y_1, \cdots, y_n$ *be the components of the vector* $\mathbf{y} = \mathbf{Ax}$. *Then the closed interval on the real axis bounded by the smallest and the largest of the n quotients* $q_j = y_j/x_j$ *contains at least one eigenvalue of* $\mathbf{A}$.

PROOF We have $\mathbf{Ax} = \mathbf{y}$ or

$$\mathbf{y} - \mathbf{Ax} = \mathbf{0}. \tag{6}$$

The transpose $\mathbf{A}^\mathsf{T}$ satisfies the conditions of Theorem 5. Hence $\mathbf{A}^\mathsf{T}$ has a positive eigenvalue λ and, corresponding to this eigenvalue, an eigenvector $\mathbf{u}$ whose components u_j are all positive. Thus $\mathbf{A}^\mathsf{T}\mathbf{u} = \lambda\mathbf{u}$ and by taking the transpose we obtain $\mathbf{u}^\mathsf{T}\mathbf{A} = \lambda\mathbf{u}^\mathsf{T}$. From this and (6) we have

$$\mathbf{u}^\mathsf{T}(\mathbf{y} - \mathbf{Ax}) = \mathbf{u}^\mathsf{T}\mathbf{y} - \mathbf{u}^\mathsf{T}\mathbf{Ax} = \mathbf{u}^\mathsf{T}\mathbf{y} - \lambda\mathbf{u}^\mathsf{T}\mathbf{x} = \mathbf{u}^\mathsf{T}(\mathbf{y} - \lambda\mathbf{x}) = 0$$

or written out

$$\sum_{j=1}^{n} u_j(y_j - \lambda x_j) = 0.$$

Since all the components u_j are positive, it follows that

$$\begin{aligned} &y_j - \lambda x_j \geqq 0, \quad \text{that is,} \quad q_j \geqq \lambda \quad \text{for at least one } j, \\ &y_j - \lambda x_j \leqq 0, \quad \text{that is,} \quad q_j \leqq \lambda \quad \text{for at least one } j. \end{aligned} \tag{7}$$ and

Since $\mathbf{A}$ and $\mathbf{A}^\mathsf{T}$ have the same eigenvalues, λ is an eigenvalue of $\mathbf{A}$, and from (7) the statement of the theorem follows. ■

[9]LOTHAR COLLATZ (1910–1990), German mathematician known for his work in numerics.

EXAMPLE 4 Bounds for Eigenvalues from Collatz's Theorem. Iteration

For a given matrix $\mathbf{A}$ with positive entries we choose an $\mathbf{x} = \mathbf{x}_0$ and **iterate**, that is, we compute $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$, $\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \cdots, \mathbf{x}_{20} = \mathbf{A}\mathbf{x}_{19}$. In each step, taking $\mathbf{x} = \mathbf{x}_j$ and $\mathbf{y} = \mathbf{A}\mathbf{x}_j = \mathbf{x}_{j+1}$ we compute an inclusion interval by Collatz's theorem. This gives (6S)

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.73 \\ 0.50 \\ 0.82 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.5481 \\ 0.3186 \\ 0.5886 \end{bmatrix},$$

$$\cdots, \mathbf{x}_{19} = \begin{bmatrix} 0.00216309 \\ 0.00108155 \\ 0.00216309 \end{bmatrix}, \quad \mathbf{x}_{20} = \begin{bmatrix} 0.00155743 \\ 0.000778713 \\ 0.00155743 \end{bmatrix}$$

and the intervals $0.5 \leqq \lambda \leqq 0.82$, $0.3186/0.50 = 0.6372 \leqq \lambda \leqq 0.5481/0.73 = 0.750822$, etc. These intervals have length

j	1	2	3	10	15	20
Length	0.32	0.113622	0.0539835	0.0004217	0.0000132	0.0000004

Using the characteristic polynomial, you may verify that the eigenvalues of $\mathbf{A}$ are 0.72, 0.36, 0.09, so that those intervals include the largest eigenvalue, 0.72. Their lengths decreased with j, so that the iteration was worthwhile. The reason will appear in the next section, where we discuss an iteration method for eigenvalues. ■

PROBLEM SET 20.7

1–6 GERSCHGORIN DISKS

Find and sketch disks or intervals that contain the eigenvalues. If you have a CAS, find the spectrum and compare.

1. $\begin{bmatrix} 5 & 2 & 4 \\ -2 & 0 & 2 \\ 2 & 4 & 7 \end{bmatrix}$

2. $\begin{bmatrix} 5 & 10^{-2} & 10^{-2} \\ 10^{-2} & 8 & 10^{-2} \\ 10^{-2} & 10^{-2} & 9 \end{bmatrix}$

3. $\begin{bmatrix} 0 & 0.4 & -0.1 \\ -0.4 & 0 & 0.3 \\ 0.1 & -0.3 & 0 \end{bmatrix}$

4. $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 3 \\ 1 & 3 & 12 \end{bmatrix}$

5. $\begin{bmatrix} 2 & i & 1+i \\ -i & 3 & 0 \\ 1-i & 0 & 8 \end{bmatrix}$

6. $\begin{bmatrix} 10 & 0.1 & -0.2 \\ 0.1 & 6 & 0 \\ -0.2 & 0 & 3 \end{bmatrix}$

7. Similarity. In Prob. 2, find $\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ such that the radius of the Gerschgorin circle with center 5 is reduced by a factor 1/100.

8. By what integer factor can you at most reduce the Gerschgorin circle with center 3 in Prob. 6?

9. If a symmetric $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has been diagonalized except for small off-diagonal entries of size 10^{-5}, what can you say about the eigenvalues?

10. Optimality of Gerschgorin disks. Illustrate with a 2×2 matrix that an eigenvalue may very well lie on a Gerschgorin circle, so that Gerschgorin disks can generally not be replaced with smaller disks without losing the inclusion property.

11. Spectral radius $\rho(\mathbf{A})$. Using Theorem 1, show that $\rho(\mathbf{A})$ cannot be greater than the row sum norm of $\mathbf{A}$.

12–16 SPECTRAL RADIUS

Use (4) to obtain an upper bound for the spectral radius:

12. In Prob. 4

13. In Prob. 1

14. In Prob. 6

15. In Prob. 3

16. In Prob. 5

17. Verify that the matrix in Prob. 5 is normal.

18. Normal matrices. Show that Hermitian, skew-Hermitian, and unitary matrices (hence real symmetric, skew-symmetric, and orthogonal matrices) are normal. Why is this of practical interest?

19. Prove Theorem 3 by using Theorem 1.

20. Extended Gerschgorin theorem. Prove Theorem 2. *Hint.* Let $\mathbf{A} = \mathbf{B} + \mathbf{C}$, $\mathbf{B} = \operatorname{diag}(a_{jj})$, $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$, and let t increase continuously from 0 to 1.

20.8 Power Method for Eigenvalues

A simple standard procedure for computing approximate values of the eigenvalues of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the **power method**. In this method we start from any vector $\mathbf{x}_0$ $(\neq \mathbf{0})$ with n components and compute successively

$$\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0, \qquad \mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \qquad \cdots, \qquad \mathbf{x}_s = \mathbf{A}\mathbf{x}_{s-1}.$$

For simplifying notation, we denote $\mathbf{x}_{s-1}$ by $\mathbf{x}$ and $\mathbf{x}_s$ by $\mathbf{y}$, so that $\mathbf{y} = \mathbf{A}\mathbf{x}$.

The method applies to any $n \times n$ matrix $\mathbf{A}$ that has a **dominant eigenvalue** (a λ such that $|\lambda|$ is greater than the absolute values of the other eigenvalues). If $\mathbf{A}$ is *symmetric*, it also gives the error bound (2), in addition to the approximation (1).

THEOREM 1

Power Method, Error Bounds

Let $\mathbf{A}$ be an $n \times n$ real symmetric matrix. Let $\mathbf{x}$ $(\neq \mathbf{0})$ be any real vector with n components. Furthermore, let

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad m_0 = \mathbf{x}^\mathsf{T}\mathbf{x}, \qquad m_1 = \mathbf{x}^\mathsf{T}\mathbf{y}, \qquad m_2 = \mathbf{y}^\mathsf{T}\mathbf{y}.$$

Then the quotient

(1)
$$q = \frac{m_1}{m_0}$$
(Rayleigh[10] quotient)

is an approximation for an eigenvalue λ of $\mathbf{A}$ (usually that which is greatest in absolute value, but no general statements are possible).

Furthermore, if we set $q = \lambda - \epsilon$, so that ϵ is the error of q, then

(2)
$$|\epsilon| \leqq \delta = \sqrt{\frac{m_2}{m_0} - q^2}.$$

PROOF δ^2 denotes the radicand in (2). Since $m_1 = qm_0$ by (1), we have

(3)
$$(\mathbf{y} - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = m_2 - 2qm_1 + q^2m_0 = m_2 - q^2m_0 = \delta^2 m_0.$$

Since $\mathbf{A}$ is real symmetric, it has an orthogonal set of n real unit eigenvectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively (some of which may be equal). (Proof in Ref. [B3], vol. 1, pp. 270–272, listed in App. 1.) Then $\mathbf{x}$ has a representation of the form

$$\mathbf{x} = a_1\mathbf{z}_1 + \cdots + a_n\mathbf{z}_n.$$

[10] LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), great English physicist and mathematician, professor at Cambridge and London, known for his important contributions to various branches of applied mathematics and theoretical physics, in particular, the theory of waves, elasticity, and hydrodynamics. In 1904 he received a Nobel Prize in physics.

Now $\mathbf{A}\mathbf{z}_1 = \lambda_1\mathbf{z}_1$, etc., and we obtain

$$\mathbf{y} = \mathbf{A}\mathbf{x} = a_1\lambda_1\mathbf{z}_1 + \cdots + a_n\lambda_n\mathbf{z}_n$$

and, since the $\mathbf{z}_j$ are orthogonal unit vectors,

$$(4) \qquad m_0 = \mathbf{x}^\mathsf{T}\mathbf{x} = a_1^2 + \cdots + a_n^2.$$

It follows that in (3),

$$\mathbf{y} - q\mathbf{x} = a_1(\lambda_1 - q)\mathbf{z}_1 + \cdots + a_n(\lambda_n - q)\mathbf{z}_n.$$

Since the $\mathbf{z}_j$ are orthogonal unit vectors, we thus obtain from (3)

$$(5) \qquad \delta^2 m_0 = (y - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = a_1^2(\lambda_1 - q)^2 + \cdots + a_n^2(\lambda_n - q)^2.$$

Now let λ_c be an eigenvalue of $\mathbf{A}$ to which q is closest, where c suggests "closest." Then $(\lambda_c - q)^2 \leqq (\lambda_j - q)^2$ for $j = 1, \cdots, n$. From this and (5) we obtain the inequality

$$\delta^2 m_0 \geqq (\lambda_c - q)^2(a_1^2 + \cdots + a_n^2) = (\lambda_c - q)^2 m_0.$$

Dividing by m_0, taking square roots, and recalling the meaning of δ^2 gives

$$\delta = \sqrt{\frac{m_2}{m_0} - q^2} \geqq |\lambda_c - q|.$$

This shows that δ is a bound for the error ϵ of the approximation q of an eigenvalue of $\mathbf{A}$ and completes the proof. ∎

The main advantage of the method is its simplicity. And it can handle *sparse matrices* too large to store as a full square array. Its disadvantage is its possibly slow convergence. From the proof of Theorem 1 we see that the speed of convergence depends on the ratio of the dominant eigenvalue to the next in absolute value (2:1 in Example 1, below).

If we want a convergent sequence of **eigenvectors,** then at the beginning of each step we **scale** the vector, say, by dividing its components by an absolutely largest one, as in Example 1, as follows.

EXAMPLE 1 Application of Theorem 1. Scaling

For the symmetric matrix $\mathbf{A}$ in Example 4, Sec. 20.7, and $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ we obtain from (1) and (2) and the indicated scaling

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.890244 \\ 0.609756 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.931193 \\ 0.541284 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_5 = \begin{bmatrix} 0.990663 \\ 0.504682 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{10} = \begin{bmatrix} 0.999707 \\ 0.500146 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{15} = \begin{bmatrix} 0.999991 \\ 0.500005 \\ 1 \end{bmatrix}.$$

Here $\mathbf{Ax}_0 = [0.73 \quad 0.5 \quad 0.82]^\mathsf{T}$, scaled to $\mathbf{x}_1 = [0.73/0.82 \quad 0.5/0.82 \quad 1]^\mathsf{T}$, etc. The dominant eigenvalue is 0.72, an eigenvector $[1 \quad 0.5 \quad 1]^\mathsf{T}$. The corresponding q and δ are computed each time before the next scaling. Thus in the first step,

$$q = \frac{m_1}{m_0} = \frac{\mathbf{x}_0^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} = \frac{2.05}{3} = 0.683333$$

$$\delta = \left(\frac{m_2}{m_0} - q^2\right)^{1/2} = \left(\frac{(\mathbf{Ax}_0)^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} - q^2\right)^{1/2} = \left(\frac{1.4553}{3} - q^2\right)^{1/2} = 0.134743.$$

This gives the following values of q, δ, and the error $\epsilon = 0.72 - q$ (calculations with 10D, rounded to 6D):

j	1	2	5	10
q	0.683333	0.716048	0.719944	0.720000
δ	0.134743	0.038887	0.004499	0.000141
ϵ	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$

The error bounds are much larger than the actual errors. This is typical, although the bounds cannot be improved; that is, for special symmetric matrices they agree with the errors.

Our present results are somewhat better than those of Collatz's method in Example 4 of Sec. 20.7, at the expense of more operations. ■

Spectral shift, the transition from $\mathbf{A}$ to $\mathbf{A} - k\mathbf{I}$, shifts every eigenvalue by $-k$. Although finding a good k can hardly be made automatic, it may be helped by some other method or small preliminary computational experiments. In Example 1, Gerschgorin's theorem gives $-0.02 \leqq \lambda \leqq 0.82$ for the whole spectrum (verify!). Shifting by -0.4 might be too much (then $-0.42 \leqq \lambda \leqq 0.42$), so let us try -0.2.

EXAMPLE 2 Power Method with Spectral Shift

For $\mathbf{A} - 0.2\mathbf{I}$ with $\mathbf{A}$ as in Example 1 we obtain the following substantial improvements (where the index 1 refers to Example 1 and the index 2 to the present example).

j	1	2	5	10
δ_1	0.134743	0.038887	0.004499	0.000141
δ_2	0.134743	0.034474	0.000693	$1.8 \cdot 10^{-6}$
ϵ_1	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$
ϵ_2	0.036667	0.002477	$1.3 \cdot 10^{-6}$	$9 \cdot 10^{-12}$

■

PROBLEM SET 20.8

1–4 POWER METHOD WITHOUT SCALING

Apply the power method without scaling (3 steps), using $\mathbf{x}_0 = [1, \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1]^\mathsf{T}$. Give Rayleigh quotients and error bounds. Show the details of your work.

1. $\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$

2. $\begin{bmatrix} 7 & -3 \\ -3 & -1 \end{bmatrix}$

3. $\begin{bmatrix} 2 & -1 & 1 \\ -1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

4. $\begin{bmatrix} 3.6 & -1.8 & 1.8 \\ -1.8 & 2.8 & -2.6 \\ 1.8 & -2.6 & 2.8 \end{bmatrix}$

5–8 POWER METHOD WITH SCALING

Apply the power method (3 steps) with scaling, using $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1 \quad 1]^\mathsf{T}$, as applicable. Give

Rayleigh quotients and error bounds. Show the details of your work.

5. The matrix in Prob. 3

6. $\begin{bmatrix} 4 & 2 & 3 \\ 2 & 7 & 6 \\ 3 & 6 & 4 \end{bmatrix}$

7. $\begin{bmatrix} 5 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 5 \end{bmatrix}$

8. $\begin{bmatrix} 2 & 4 & 0 & 1 \\ 4 & 1 & 2 & 8 \\ 0 & 2 & 5 & 2 \\ 1 & 8 & 2 & 0 \end{bmatrix}$

9. Prove that if $\mathbf{x}$ is an eigenvector, then $\delta = 0$ in (2). Give two examples.

10. **Rayleigh quotient.** Why does q generally approximate the eigenvalue of greatest absolute value? When will q be a good approximation?

11. **Spectral shift, smallest eigenvalue.** In Prob. 3 set $\mathbf{B} = \mathbf{A} - 3\mathbf{I}$ (as perhaps suggested by the diagonal entries) and see whether you may get a sequence of q's converging to an eigenvalue of $\mathbf{A}$ that *is smallest* (not largest) in absolute value. Use $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$. Do 8 steps. Verify that $\mathbf{A}$ has the spectrum $\{0, 3, 5\}$.

12. **CAS EXPERIMENT. Power Method with Scaling. Shifting. (a)** Write a program for $n \times n$ matrices that prints every step. Apply it to the (nonsymmetric!) matrix (20 steps), starting from $[1 \quad 1 \quad 1]^\mathsf{T}$.

$$\mathbf{A} = \begin{bmatrix} 15 & 12 & 3 \\ 18 & 44 & 18 \\ -19 & -36 & -7 \end{bmatrix}.$$

(b) Experiment in (a) with shifting. Which shift do you find optimal?

(c) Write a program as in (a) but for symmetric matrices that prints vectors, scaled vectors, q, and δ. Apply it to the matrix in Prob. 8.

(d). Optimality of δ. Consider $\mathbf{A} = \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$ and take $\mathbf{x}_0 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$. Show that $q = 0, \delta = 1$ for all steps and the eigenvalues are ± 1, so that the interval $[q - \delta, q + \delta]$ cannot be shortened (by omitting ± 1) without losing the inclusion property. Experiment with other $\mathbf{x}_0$'s.

(e) Find a (nonsymmetric) matrix for which δ in (2) is no longer an error bound.

(f) Experiment systematically with speed of convergence by choosing matrices with the second greatest eigenvalue (i) almost equal to the greatest, (ii) somewhat different, (iii) much different.

20.9 Tridiagonalization and QR-Factorization

We consider the problem of computing *all* the eigenvalues of a ***real symmetric*** matrix $\mathbf{A} = [a_{jk}]$, discussing a method widely used in practice. In the ***first stage*** we reduce the given matrix stepwise to a **tridiagonal matrix**, that is, a matrix having all its nonzero entries on the main diagonal and in the positions immediately adjacent to the main diagonal (such as $\mathbf{A}_3$ in Fig. 450, Third Step). This reduction was invented by A. S. Householder[11] (*J. Assn. Comput. Machinery* **5** (1958), 335–342). See also Ref. [E29] in App. 1.

This Householder tridiagonalization will simplify the matrix without changing its eigenvalues. The latter will then be determined (approximately) by factoring the tridiagonalized matrix, as discussed later in this section.

[11]ALSTON SCOTT HOUSEHOLDER (1904–1993), American mathematician, known for his work in numerical analysis and mathematical biology. He was head of the mathematics division at Oakridge National Laboratory and later professor at the University of Tennessee. He was both president of ACM (Association for Computing Machinery) 1954–1956 and SIAM (Society for Industrial and Applied Mathematics) 1963–1964.

Householder's Tridiagonalization Method[11]

An $n \times n$ real symmetric matrix $\mathbf{A} = [a_{jk}]$ being given, we reduce it by $n - 2$ successive similarity transformations (see Sec. 20.6) involving matrices $\mathbf{P}_1, \cdots, \mathbf{P}_{n-2}$ to tridiagonal form. These matrices are orthogonal and symmetric. Thus $\mathbf{P}_1^{-1} = \mathbf{P}_1^{\mathsf{T}} = \mathbf{P}_1$ and similarly for the others. These transformations produce, from the given $\mathbf{A}_0 = \mathbf{A} = [a_{jk}]$, the matrices $\mathbf{A}_1 = [a_{jk}^{(1)}]$, $\mathbf{A}_2 = [a_{jk}^{(2)}], \cdots, \mathbf{A}_{n-2} = [a_{jk}^{(n-2)}]$ in the form

$$
\begin{aligned}
\mathbf{A}_1 &= \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 \\
\mathbf{A}_2 &= \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 \\
&\cdots\cdots\cdots\cdots \\
\mathbf{B} = \mathbf{A}_{n-2} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2}.
\end{aligned}
\tag{1}
$$

The transformations (1) create the necessary zeros, in the first step in Row 1 and Column 1, in the second step in Row 2 and Column 2, etc., as Fig. 450 illustrates for a 5×5 matrix. $\mathbf{B}$ is tridiagonal.

$$
\begin{bmatrix} * & * & & & \\ * & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \end{bmatrix}
\qquad
\begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & * \\ & & * & * & * \\ & & * & * & * \end{bmatrix}
\qquad
\begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & \\ & & * & * & * \\ & & & * & * \end{bmatrix}
$$

First Step $\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}\mathbf{P}_1$ | Second Step $\mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2$ | Third Step $\mathbf{A}_3 = \mathbf{P}_3\mathbf{A}_2\mathbf{P}_3$

Fig. 450. Householder's method for a 5×5 matrix. Positions left blank are zeros created by the method.

How do we determine $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{n-2}$? Now, all these $\mathbf{P}_r$ are of the form

$$
\mathbf{P}_r = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} \qquad (r = 1, \cdots, n-2) \tag{2}
$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{v}_r = [v_{jr}]$ is a unit vector with its first r components 0; thus

$$
\mathbf{v}_1 = \begin{bmatrix} 0 \\ * \\ * \\ \vdots \\ * \end{bmatrix}, \qquad
\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ * \\ \vdots \\ * \end{bmatrix}, \qquad \cdots, \qquad
\mathbf{v}_{n-2} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ * \\ * \end{bmatrix}
\tag{3}
$$

where the asterisks denote the other components (which will be nonzero in general).

Step 1. $\mathbf{v}_1$ has the components

$$
\begin{aligned}
&v_{11} = 0 \\
\textbf{(4)} \quad \textbf{(a)} \quad & v_{21} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{21}|}{S_1}\right)} \\
\textbf{(b)} \quad & v_{j1} = \frac{a_{j1}\,\mathrm{sgn}\, a_{21}}{2v_{21}S_1} \qquad j = 3, 4, \cdots, n \\
& \text{where} \\
\textbf{(c)} \quad & S_1 = \sqrt{a_{21}^2 + a_{31}^2 + \cdots + a_{n1}^2}
\end{aligned}
$$

where $S_1 > 0$, and $\mathrm{sgn}\, a_{21} = +1$ if $a_{21} \geqq 0$ and $\mathrm{sgn}\, a_{21} = -1$ if $a_{21} < 0$. With this we compute $\mathbf{P}_1$ by (2) and then $\mathbf{A}_1$ by (1). This was the first step.

Step 2. We compute $\mathbf{v}_2$ by (4) with all subscripts increased by 1 and the a_{jk} replaced by $a_{jk}^{(1)}$, the entries of $\mathbf{A}_1$ just computed. Thus [see also (3)]

$$
\begin{aligned}
& v_{12} = v_{22} = 0 \\
(4^*) \quad & v_{32} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{32}^{(1)}|}{S_2}\right)} \\
& v_{j2} = \frac{a_{j2}^{(1)}\,\mathrm{sgn}\, a_{32}^{(1)}}{2v_{32}S_2} \qquad j = 4, 5, \cdots, n
\end{aligned}
$$

where

$$S_2 = \sqrt{a_{32}^{(1)^2} + a_{42}^{(1)^2} + \cdots + a_{n2}^{(1)^2}}.$$

With this we compute $\mathbf{P}_2$ by (2) and then $\mathbf{A}_2$ by (1).

Step 3. We compute $\mathbf{v}_3$ by (4*) with all subscripts increased by 1 and the $a_{jk}^{(1)}$ replaced by the entries $a_{jk}^{(2)}$ of $\mathbf{A}_2$, and so on.

EXAMPLE 1 Householder Tridiagonalization

Tridiagonalize the real symmetric matrix

$$\mathbf{A} = \mathbf{A}_0 = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. ***Step 1.*** We compute $S_1^2 = 4^2 + 1^2 + 1^2 = 18$ from (4c). Since $a_{21} = 4 > 0$, we have $\mathrm{sgn}\, a_{21} = +1$ in (4b) and get from (4) by straightforward computation

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ v_{21} \\ v_{31} \\ v_{41} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.98559856 \\ 0.11957316 \\ 0.11957316 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.94280904 & -0.23570227 & -0.23570227 \\ 0 & -0.23570227 & 0.97140452 & -0.02859548 \\ 0 & -0.23570227 & -0.02859548 & 0.97140452 \end{bmatrix}.$$

From the first line in (1) we now get

$$\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & -1 & -1 \\ 0 & -1 & \frac{9}{2} & \frac{3}{2} \\ 0 & -1 & \frac{3}{2} & \frac{9}{2} \end{bmatrix}.$$

Step 2. From (4*) we compute $S_2^2 = 2$ and

$$\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ v_{32} \\ v_{42} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.92387953 \\ 0.38268343 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

The second line in (1) now gives

$$\mathbf{B}_2 = \mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

This matrix **B** is tridiagonal. Since our given matrix has order $n = 4$, we needed $n - 2 = 2$ steps to accomplish this reduction, as claimed. (Do you see that we got more zeros than we can expect in general?)

B is similar to **A**, as we now show in general. This is essential because **B** thus has the same spectrum as **A**, by Theorem 2 in Sec. 20.6. ■

B *Similar to* A. We assert that **B** in (1) is similar to $\mathbf{A} = \mathbf{A}_0$. The matrix $\mathbf{P}_r$ is symmetric; indeed,

$$\mathbf{P}_r^{\mathsf{T}} = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I}^{\mathsf{T}} - 2(\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} = \mathbf{P}_r$$

Also, $\mathbf{P}_r$ is orthogonal because $\mathbf{v}_r$ is a unit vector, so that $\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r = 1$ and thus

$$\mathbf{P}_r\mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r^2 = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^2 = \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}}$$
$$= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r(\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r)\mathbf{v}_r^{\mathsf{T}} = \mathbf{I}.$$

Hence $\mathbf{P}_r^{-1} = \mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r$ and from (1) we now obtain

$$\begin{aligned} \mathbf{B} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2} = \cdots \\ \cdots &= \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_1\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}_{n-2}^{-1}\mathbf{P}_{n-3}^{-1}\cdots\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \end{aligned}$$

where $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-2}$. This proves our assertion. ■

QR-Factorization Method

In 1958 H. Rutishauser[12] of Switzerland proposed the idea of using the LU-factorization (Sec. 20.2; he called it LR-factorization) in solving eigenvalue problems. An improved version of Rutishauser's method (avoiding breakdown if certain submatrices become singular, etc.; see Ref. [E29]) is the QR-method, independently proposed by the American J. G. F. Francis (*Computer J.* **4** (1961–62), 265–271, 332–345) and the Russian V. N. Kublanovskaya (*Zhurnal Vych. Mat. i Mat. Fiz.* **1** (1961), 555–570). The QR-method uses the factorization **QR** with orthogonal **Q** and upper triangular **R.** We discuss the **QR**-method for a real ***symmetric*** matrix. (For extensions to general matrices see Ref. [E29] in App. 1.)

In this method we first transform a given real symmetric $n \times n$ matrix $\mathbf{A}$ into a tridiagonal matrix $\mathbf{B}_0 = \mathbf{B}$ by Householder's method. This creates many zeros and thus reduces the amount of further work. Then we compute $\mathbf{B}_1, \mathbf{B}_2, \cdots$ stepwise according to the following iteration method.

Step 1. Factor $\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0$ with orthogonal $\mathbf{Q}_0$ and upper triangular $\mathbf{R}_0$. Then compute $\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0$.

Step 2. Factor $\mathbf{B}_1 = \mathbf{Q}_1\mathbf{R}_1$. Then compute $\mathbf{B}_2 = \mathbf{R}_1\mathbf{Q}_1$.

General Step s + 1.

$$\text{(5)} \qquad \begin{array}{ll} \textbf{(a)} & \text{Factor } \mathbf{B}_s = \mathbf{Q}_s\mathbf{R}_s. \\ \textbf{(b)} & \text{Compute } \mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s. \end{array}$$

Here $\mathbf{Q}_s$ is orthogonal and $\mathbf{R}_s$ upper triangular. The factorization (5a) will be explained below.

$\mathbf{B}_{s+1}$ Similar to B. Convergence to a Diagonal Matrix. From (5a) we have $\mathbf{R}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s$. Substitution into (5b) gives

$$\text{(6)} \qquad \mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s\mathbf{Q}_s.$$

[12] HEINZ RUTISHAUSER (1918–1970). Swiss mathematician, professor at ETH Zurich. Known for his pioneering work in numerics and computer science.

Thus $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_s$. Hence $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_0 = \mathbf{B}$ for all s. By Theorem 2, Sec. 20.6, this implies that $\mathbf{B}_{s+1}$ has the same eigenvalues as $\mathbf{B}$.

Also, $\mathbf{B}_{s+1}$ is symmetric. This follows by induction. Indeed, $\mathbf{B}_0 = \mathbf{B}$ is symmetric. Assuming $\mathbf{B}_s$ to be symmetric, that is, $\mathbf{B}_s{}^\mathsf{T} = \mathbf{B}_s$, and using $\mathbf{Q}_s^{-1} = \mathbf{Q}_s{}^\mathsf{T}$ (since $\mathbf{Q}_s$ is orthogonal), we get from (6) the symmetry,

$$\mathbf{B}_{s+1}{}^\mathsf{T} = (\mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s)^\mathsf{T} = Q_s{}^\mathsf{T}\mathbf{B}_s{}^\mathsf{T}\mathbf{Q}_s = \mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s = \mathbf{B}_{s+1}.$$

If the eigenvalues of $\mathbf{B}$ are different in absolute value, say, $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, then

$$\lim_{s\to\infty} \mathbf{B}_s = \mathbf{D}$$

where $\mathbf{D}$ is diagonal, with main diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_n$. (Proof in Ref. [E29] listed in App. 1.)

How to Get the QR-Factorization, say, $\mathbf{B} = \mathbf{B}_0 = [b_{jk}] = \mathbf{Q}_0\mathbf{R}_0$. The tridiagonal matrix $\mathbf{B}$ has $n - 1$ generally nonzero entries below the main diagonal. These are $b_{21}, b_{32}, \cdots, b_{n,n-1}$. We multiply $\mathbf{B}$ from the left by a matrix $\mathbf{C}_2$ such that $\mathbf{C}_2\mathbf{B} = [b_{jk}^{(2)}]$ has $b_{21}^{(2)} = 0$. We multiply this by a matrix $\mathbf{C}_3$ such that $\mathbf{C}_3\mathbf{C}_2\mathbf{B} = [b_{jk}^{(3)}]$ has $b_{32}^{(3)} = 0$, etc. After $n - 1$ such multiplications we are left with an upper triangular matrix $\mathbf{R}_0$, namely,

$$\mathbf{C}_n\mathbf{C}_{n-1}\cdots\mathbf{C}_3\mathbf{C}_2 B_0 = \mathbf{R}_0. \tag{7}$$

These $n \times n$ matrices $\mathbf{C}_j$ are very simple. $\mathbf{C}_j$ has the 2×2 submatrix

$$\begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix} \qquad (\theta_j \text{ suitable})$$

in Rows $j - 1$ and j and Columns $j - 1$ and j; everywhere else on the main diagonal the matrix $\mathbf{C}_j$ has entries 1; and all its other entries are 0. (This submatrix is the matrix of a plane rotation through the angle θ_j; see Team Project 30, Sec. 7.2.) For instance, if $n = 4$, writing $c_j = \cos\theta_j$, $s_j = \sin\theta_j$, we have

$$\mathbf{C}_2 = \begin{bmatrix} c_2 & s_2 & 0 & 0 \\ -s_2 & c_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -s_4 & c_4 \end{bmatrix}.$$

These $\mathbf{C}_j$ are orthogonal. Hence their product in (7) is orthogonal, and so is the inverse of this product. We call this inverse $\mathbf{Q}_0$. Then from (7),

$$\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0 \tag{8}$$

where, with $\mathbf{C}_j^{-1} = \mathbf{C}_j{}^\mathsf{T}$,

$$\mathbf{Q}_0 = (\mathbf{C}_n\mathbf{C}_{n-1}\cdots\mathbf{C}_3\mathbf{C}_2)^{-1} = \mathbf{C}_2{}^\mathsf{T}\mathbf{C}_3{}^\mathsf{T}\cdots\mathbf{C}_{n-1}{}^\mathsf{T}\mathbf{C}_n{}^\mathsf{T}. \tag{9}$$

This is our QR-factorization of $\mathbf{B}_0$. From it we have by (5b) with $s = 0$

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}}\cdots\mathbf{C}_{n-1}{}^{\mathsf{T}}\mathbf{C}_n{}^{\mathsf{T}}. \tag{10}$$

We do not need $\mathbf{Q}_0$ explicitly, but to get $\mathbf{B}_1$ from (10), we first compute $\mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}$, then $(\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}})\mathbf{C}_3^{\mathsf{T}}$, etc. Similarly in the further steps that produce $\mathbf{B}_2, \mathbf{B}_3, \cdots$.

Determination of cos $\boldsymbol{\theta_j}$ and sin $\boldsymbol{\theta_j}$. We finally show how to find the angles of rotation. $\cos\theta_2$ and $\sin\theta_2$ in $\mathbf{C}_2$ must be such that $b_{21}^{(2)} = 0$ in the product

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} c_2 & s_2 & 0 & \cdots \\ -s_2 & c_2 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} & \cdots \\ b_{21} & b_{22} & b_{23} & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Now $b_{21}^{(2)}$ is obtained by multiplying the second row of $\mathbf{C}_2$ by the first column of $\mathbf{B}$,

$$b_{21}^{(2)} = -s_2 b_{11} + c_2 b_{21} = -(\sin\theta_2)b_{11} + (\cos\theta_2)b_{21} = 0.$$

Hence $\tan\theta_2 = s_2/c_2 = b_{21}/b_{11}$, and

$$\begin{aligned} \cos\theta_2 &= \frac{1}{\sqrt{1+\tan^2\theta_2}} = \frac{1}{\sqrt{1+(b_{21}/b_{11})^2}} \\ \sin\theta_2 &= \frac{\tan\theta_2}{\sqrt{1+\tan^2\theta_2}} = \frac{b_{21}/b_{11}}{\sqrt{1+(b_{21}/b_{11})^2}}. \end{aligned} \tag{11}$$

Similarly for $\theta_3, \theta_4, \cdots$. The next example illustrates all this.

EXAMPLE 2 QR-Factorization Method

Compute all the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. We first reduce $\mathbf{A}$ to tridiagonal form. Applying Householder's method, we obtain (see Example 1)

$$\mathbf{A}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

From the characteristic determinant we see that $\mathbf{A}_2$, hence $\mathbf{A}$, has the eigenvalue 3. (Can you see this directly from $\mathbf{A}_2$?) Hence it suffices to apply the QR-method to the tridiagonal 3×3 matrix

$$\mathbf{B}_0 = \mathbf{B} = \begin{bmatrix} 6 & -\sqrt{18} & 0 \\ -\sqrt{18} & 7 & \sqrt{2} \\ 0 & \sqrt{2} & 6 \end{bmatrix}.$$

Step 1. We multiply $\mathbf{B}$ from the left by

$$\mathbf{C}_2 = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and then } \mathbf{C}_2\mathbf{B} \text{ by} \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & \sin\theta_3 \\ 0 & -\sin\theta_3 & \cos\theta_3 \end{bmatrix}.$$

Here $(-\sin\theta_2)\cdot 6 + (\cos\theta_2)(-\sqrt{18}) = 0$ gives (11) $\cos\theta_2 = 0.81649658$ and $\sin\theta_2 = -0.57735027$. With these values we compute

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.26598632 & 1.15470054 \\ 0 & 1.41421356 & 6.00000000 \end{bmatrix}.$$

In $\mathbf{C}_3$ we get from $(-\sin\theta_3)\cdot 3.26598632 + (\cos\theta_3)\cdot 1.41421356 = 0$ the values $\cos\theta_3 = 0.91766294$ and $\sin\theta_3 = 0.39735971$. This gives

$$\mathbf{R}_0 = \mathbf{C}_3\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.55902608 & 3.44378413 \\ 0 & 0 & 5.04714615 \end{bmatrix}.$$

From this we compute

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}} = \begin{bmatrix} 10.33333333 & -2.05480467 & 0 \\ -2.05480467 & 4.03508772 & 2.00553251 \\ 0 & 2.00553251 & 4.63157895 \end{bmatrix}$$

which is symmetric and tridiagonal. The off-diagonal entries in $\mathbf{B}_1$ are still large in absolute value. Hence we have to go on.

Step 2. We do the same computations as in the first step, with $\mathbf{B}_0 = \mathbf{B}$ replaced by $\mathbf{B}_1$ and $\mathbf{C}_2$ and $\mathbf{C}_3$ changed accordingly, the new angles being $\theta_2 = -0.196291533$ and $\theta_3 = 0.513415589$. We obtain

$$\mathbf{R}_1 = \begin{bmatrix} 10.53565375 & -2.80232241 & -0.39114588 \\ 0 & 4.08329584 & 3.98824028 \\ 0 & 0 & 3.06832668 \end{bmatrix}$$

and from this

$$\mathbf{B}_2 = \begin{bmatrix} 10.87987988 & -0.79637918 & 0 \\ -0.79637918 & 5.44738664 & 1.50702500 \\ 0 & 1.50702500 & 2.67273348 \end{bmatrix}.$$

We see that the off-diagonal entries are somewhat smaller in absolute value than those of $\mathbf{B}_1$, but still much too large for the diagonal entries to be good approximations of the eigenvalues of $\mathbf{B}$.

Further Steps. We list the main diagonal entries and the absolutely largest off-diagonal entry, which is $|b_{12}^{(j)}| = |b_{21}^{(j)}|$ in all steps. You may show that the given matrix **A** has the spectrum 11, 6, 3, 2.

Step j	$b_{11}^{(j)}$	$b_{22}^{(j)}$	$b_{33}^{(j)}$	$\max_{j \neq k} \lvert b_{jk}^{(J)} \rvert$
3	10.9668929	5.94589856	2.08720851	0.58523582
5	10.9970872	6.00181541	2.00109738	0.12065334
7	10.9997421	6.00024439	2.00001355	0.03591107
9	10.9999772	6.00002267	2.00000017	0.01068477

Looking back at our discussion, we recognize that the purpose of applying Householder's tridiagonalization before the QR-factorization method is a substantial reduction of cost in each QR-factorization, in particular if **A** is large.

Convergence acceleration and thus further reduction of cost can be achieved by a **spectral shift,** that is, by taking $\mathbf{B}_s - k_s\mathbf{I}$ instead of $\mathbf{B}_s$ with a suitable k_s. Possible choices of k_s are discussed in Ref. [E29], p. 510.

PROBLEM SET 20.9

1–5 HOUSEHOLDER TRIDIAGONALIZATION

Tridiagonalize. Show the details.

1. $\begin{bmatrix} 0.98 & 0.04 & 0.44 \\ 0.04 & 0.56 & 0.40 \\ 0.44 & 0.40 & 0.80 \end{bmatrix}$

2. $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 7 & 2 & 3 \\ 2 & 10 & 6 \\ 3 & 6 & 7 \end{bmatrix}$

4. $\begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}$

5. $\begin{bmatrix} 3 & 52 & 10 & 42 \\ 52 & 59 & 44 & 80 \\ 10 & 44 & 39 & 42 \\ 42 & 80 & 42 & 35 \end{bmatrix}$

6–9 QR-FACTORIZATION

Do three QR-steps to find approximations of the eigenvalues of:

6. The matrix in the answer to Prob. 1

7. The matrix in the answer to Prob. 3

8. $\begin{bmatrix} 14.2 & -0.1 & 0 \\ -0.1 & -6.3 & 0.2 \\ 0 & 0.2 & 2.1 \end{bmatrix}$

9. $\begin{bmatrix} 140 & 10 & 0 \\ 10 & 70 & 2 \\ 0 & 2 & -30 \end{bmatrix}$

10. CAS EXPERIMENT. QR-Method. Try to find out experimentally on what properties of a matrix the speed of decrease of off-diagonal entries in the QR-method depends. For this purpose write a program that first tridiagonalizes and then does QR-steps. Try the program out on the matrices in Probs. 1, 3, and 4. Summarize your findings in a short report.

CHAPTER 20 REVIEW QUESTIONS AND PROBLEMS

1. What are the main problem areas in numeric linear algebra?

2. When would you apply Gauss elimination and when Gauss–Seidel iteration?

3. What is pivoting? Why and how is it done?

4. What happens if you apply Gauss elimination to a system that has no solutions?

5. What is Cholesky's method? When would you apply it?

6. What do you know about the convergence of the Gauss–Seidel iteration?

7. What is ill-conditioning? What is the condition number and its significance?

8. Explain the idea of least squares approximation.

9. What are eigenvalues of a matrix? Why are they important? Give typical examples.

10. How did we use similarity transformations of matrices in designing numeric methods?

11. What is the power method for eigenvalues? What are its advantages and disadvantages?

12. State Gerschgorin's theorem from memory. Give typical applications.

13. What is tridiagonalization and QR? When would you apply it?

14–17 GAUSS ELIMINATION

Solve

14.
$$\begin{aligned} 3x_2 - 6x_3 &= 0 \\ 4x_1 - x_2 + 2x_3 &= 16 \\ -5x_1 + 2x_2 - 4x_3 &= -20 \end{aligned}$$

15.
$$\begin{aligned} 8x_2 - 6x_3 &= 23.6 \\ 10x_1 + 6x_2 + 2x_3 &= 68.4 \\ 12x_1 - 14x_2 + 4x_3 &= -6.2 \end{aligned}$$

16.
$$\begin{aligned} 5x_1 + x_2 - 3x_3 &= 17 \\ -5x_2 + 15x_3 &= -10 \\ 2x_1 - 3x_2 + 9x_3 &= 0 \end{aligned}$$

17.
$$\begin{aligned} 42x_1 + 74x_2 + 36x_3 &= 96 \\ -46x_1 - 12x_2 - 2x_3 &= 82 \\ 3x_1 + 25x_2 + 5x_3 &= 19 \end{aligned}$$

18–20 INVERSE MATRIX

Compute the inverse of:

18. $\begin{bmatrix} 2.0 & 0.1 & 3.3 \\ 1.6 & 4.4 & 0.5 \\ 0.3 & -4.3 & 2.8 \end{bmatrix}$

19. $\begin{bmatrix} 15 & 20 & 10 \\ 20 & 35 & 15 \\ 10 & 15 & 90 \end{bmatrix}$

20. $\begin{bmatrix} 5 & 1 & 1 \\ 1 & 6 & 0 \\ 1 & 0 & 8 \end{bmatrix}$

21–23 GAUSS–SEIDEL ITERATION

Do 3 steps without scaling, starting from $[1 \quad 1 \quad 1]^T$.

21.
$$\begin{aligned} 4x_1 - x_2 \qquad &= 22.0 \\ 4x_2 - x_3 &= 13.4 \\ -x_1 \qquad + 4x_3 &= -2.4 \end{aligned}$$

22.
$$\begin{aligned} 0.2x_1 + 4.0x_2 - 0.4x_3 &= 32.0 \\ 0.5x_1 - 0.2x_2 + 2.5x_3 &= -5.1 \\ 7.5x_1 + 0.1x_2 - 1.5x_3 &= -12.7 \end{aligned}$$

23.
$$\begin{aligned} 10x_1 + x_2 - x_3 &= 17 \\ 2x_1 + 20x_2 + x_3 &= 28 \\ 3x_1 - x_2 + 25x_3 &= 105 \end{aligned}$$

24–26 VECTOR NORMS

Compute the ℓ_1-, ℓ_2-, and ℓ_∞-norms of the vectors.

24. $[0.2 \quad -8.1 \quad 0.4 \quad 0 \quad 0 \quad -1.3 \quad 2]^T$

25. $[8 \quad -21 \quad 13 \quad 0]^T$

26. $[0 \quad 0 \quad 0 \quad -1 \quad 0]^T$

27–30 MATRIX NORM

Compute the matrix norm corresponding to the ℓ_∞-vector norm for the coefficient matrix:

27. In Prob. 15

28. In Prob. 17

29. In Prob. 21

30. In Prob. 22

31–33 CONDITION NUMBER

Compute the condition number (corresponding to the ℓ_∞-vector norm) of the coefficient matrix:

31. In Prob. 19

32. In Prob. 18

33. In Prob. 21

34–35 FITTING BY LEAST SQUARES

Fit and graph:

34. A straight line to $(-1, 0)$, $(0, 2)$, $(1, 2)$, $(2, 3)$, $(3, 3)$

35. A quadratic parabola to the data in Prob. 34.

36–39 EIGENVALUES

Find and graph three circular disks that must contain all the eigenvalues of the matrix:

36. In Prob. 18

37. In Prob. 19

38. In Prob. 20

39. Of the coefficients in Prob. 14

40. Power method. Do 4 steps with scaling for the matrix in Prob. 19, starting for [1 1 1] and computing the Rayliegh quotients and error bounds.

SUMMARY OF CHAPTER 20
Numeric Linear Algebra

Main tasks are the numeric solution of linear systems (Secs. 20.1–20.4), curve fitting (Sec. 20.5), and eigenvalue problems (Secs. 20.6–20.9).

Linear systems $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = [a_{jk}]$, written out

$$
\begin{aligned}
\mathrm{E}_1\colon &\quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\mathrm{E}_2\colon &\quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\
&\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathrm{E}_n\colon &\quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n
\end{aligned}
\tag{1}
$$

can be solved by a **direct method** (one in which the number of numeric operations can be specified in advance, e.g., Gauss's elimination) or by an **indirect** or **iterative method** (in which an initial approximation is improved stepwise).

The **Gauss elimination** (Sec. 20.1) is direct, namely, a systematic elimination process that reduces (1) stepwise to triangular form. In Step 1 we eliminate x_1 from equations E_2 to E_n by subtracting $(a_{21}/a_{11})\,\mathrm{E}_1$ from E_2, then $(a_{31}/a_{11})\,\mathrm{E}_1$ from E_3, etc. Equation E_1 is called the **pivot equation** in this step and a_{11} the **pivot.** In Step 2 we take the new second equation as pivot equation and eliminate x_2, etc. If the triangular form is reached, we get x_n from the last equation, then x_{n-1} from the second last, etc. **Partial pivoting** (= interchange of equations) is *necessary* if candidates for pivots are zero, and *advisable* if they are small in absolute value.

Doolittle's, Crout's, and **Cholesky's methods** in Sec. 20.2 are variants of the Gauss elimination. They factor $\mathbf{A} = \mathbf{LU}$ ($\mathbf{L}$ lower triangular, $\mathbf{U}$ upper triangular) and solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ by solving $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$ and then $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$.

In the **Gauss–Seidel iteration** (Sec. 20.3) we make $a_{11} = a_{22} = \cdots = a_{nn} = 1$ (by division) and write $\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}$; thus $\mathbf{x} = \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}$, which suggests the iteration formula

$$
\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - \mathbf{U}\mathbf{x}^{(m)}
\tag{2}
$$

in which we always take the most recent approximate x_j's on the right. If $\|\mathbf{C}\| < 1$, where $\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}$, then this process converges. Here, $\|\mathbf{C}\|$ denotes any matrix norm (Sec. 20.3).

If the **condition number** $k(\mathbf{A}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$ of $\mathbf{A}$ is large, then the system $\mathbf{Ax} = \mathbf{b}$ is **ill-conditioned** (Sec. 20.4), and a small **residual** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ does ***not*** imply that $\tilde{\mathbf{x}}$ is close to the exact solution.

The fitting of a polynomial $p(x) = b_0 + b_1x + \cdots + b_mx^m$ through given data (points in the xy-plane) $(x_1, y_1), \cdots, (x_n, y_n)$ by the method of **least squares** is discussed in Sec. 20.5 (and in statistics in Sec. 25.9). If $m = n$, the least squares polynomial will be the same as an interpolating polynomial (uniqueness).

Eigenvalues λ (values λ for which $\mathbf{Ax} = \lambda\mathbf{x}$ has a solution $\mathbf{x} \neq \mathbf{0}$, called an **eigenvector**) can be characterized by inequalities (Sec. 20.7), e.g. in **Gerschgorin's theorem,** which gives n circular disks which contain the whole spectrum (all eigenvalues) of $\mathbf{A}$, of centers a_{jj} and radii $\Sigma|a_{jk}|$ (sum over k from 1 to n, $k \neq j$).

Approximations of eigenvalues can be obtained by iteration, starting from an $\mathbf{x}_0 \neq \mathbf{0}$ and computing $\mathbf{x}_1 = \mathbf{Ax}_0$, $\mathbf{x}_2 = \mathbf{Ax}_1, \cdots, \mathbf{x}_n = \mathbf{Ax}_{n-1}$. In this **power method** (Sec. 20.8) the **Rayleigh quotient**

$$q = \frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \qquad (\mathbf{x} = \mathbf{x}_n) \tag{3}$$

gives an approximation of an eigenvalue (usually that of the greatest absolute value) and, if $\mathbf{A}$ is symmetric, an error bound is

$$|\epsilon| \leqq \sqrt{\frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} - q^2}. \tag{4}$$

Convergence may be slow but can be improved by a *spectral shift.*

For determining all the eigenvalues of a symmetric matrix $\mathbf{A}$ it is best to first tridiagonalize $\mathbf{A}$ and then to apply the QR-method (Sec. 20.9), which is based on a factorization $\mathbf{A} = \mathbf{QR}$ with orthogonal $\mathbf{Q}$ and upper triangular $\mathbf{R}$ and uses similarity transformations.

CHAPTER 21

Numerics for ODEs and PDEs

Ordinary differential equations (ODEs) and partial differential equations (PDEs) play a central role in modeling problems of engineering, mathematics, physics, aeronautics, astronomy, dynamics, elasticity, biology, medicine, chemistry, environmental science, economics, and many other areas. Chapters 1–6 and 12 explained the major approaches to solving ODEs and PDEs analytically. However, in your career as an engineer, applied mathematicians, or physicist you will encounter ODEs and PDEs that *cannot* be solved by those analytic methods or whose solutions are so difficult that other approaches are needed. It is precisely in these real-world projects that numeric methods for ODEs and PDEs are used, often as part of a software package. Indeed, numeric software has become an indispensable tool for the engineer.

This chapter is evenly divided between numerics for ODEs and numerics for PDEs. We start with ODEs and discuss, in Sec. 21.1, methods for first-order ODEs. The main initial idea is that we can obtain approximations to the solution of such an ODE at points that are a distance h apart by using the first two terms of Taylor's formula from calculus. We use these approximations to construct the iteration formula for a method known as Euler's method. While this method is rather unstable and of little practical use, it serves as a pedagogical tool and a starting point toward understanding more sophisticated methods such as the Runge–Kutta method and its variant the Runga–Kutta–Fehlberg (RKF) method, which are popular and useful in practice. As is usual in mathematics, one tends to generalize mathematical ideas. The methods of Sec. 21.1 are one-step methods, that is, the current approximation uses only the approximation from the previous step. Multistep methods, such as the Adams–Bashforth methods and Adams–Moulton methods, use values computed from several previous steps. We conclude numerics for ODEs with applying Runge–Kutta–Nyström methods and other methods to higher order ODEs and systems of ODEs.

Numerics for PDEs are perhaps even more exciting and ingenious than those for ODEs. We first consider PDEs of the elliptic type (Laplace, Poisson). Again, Taylor's formula serves as a starting point and lets us replace partial derivatives by difference quotients. The end result leads to a mesh and an evaluation scheme that uses the Gauss–Seidel method (here also know as Liebmann's method). We continue with methods that use grids to solve Neuman and mixed problems (Sec. 21.5) and conclude with the important Crank–Nicholson method for parabolic PDEs in Sec. 21.6.

***Sections* 21.1 *and* 21.2 *may be studied immediately after Chap.* 1 *and Sec.* 21.3 *immediately after Chaps.* 2–4,** because these sections are independent of Chaps. 19 and 20.

***Sections* 21.4–21.7 *on PDEs may be studied immediately after Chap.* 12** if students have some knowledge of linear systems of algebraic equations.

Prerequisite: Secs. 1.1–1.5 for ODEs, Secs. 12.1–12.3, 12.5, 12.10 for PDEs.

References and Answers to Problems: App. 1 Part E (see also Parts A and C), App. 2.

21.1 Methods for First-Order ODEs

Take a look at Sec. 1.2, where we briefly introduced Euler's method with an example. *We shall develop* ***Euler's method*** *more rigorously.* Pay close attention to the derivation that uses Taylor's formula from calculus to approximate the solution to a first-order ODE at points that are a distance h apart. If you understand this approach, which is typical for numerics for ODEs, then you will understand other methods more easily.

From Chap. 1 we know that an ODE of the first order is of the form $F(x, y, y') = 0$ and can often be written in the explicit form $y' = f(x, y)$. An **initial value problem** for this equation is of the form

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

where x_0 and y_0 are given and we assume that the problem has a unique solution on some open interval $a < x < b$ containing x_0.

In this section we shall discuss methods of computing approximate numeric values of the solution $y(x)$ of (1) at the equidistant points on the x-axis

$$x_1 = x_0 + h, \qquad x_2 = x_0 + 2h, \qquad x_3 = x_0 + 3h, \qquad \cdots$$

where the **step size** h is a fixed number, for instance, 0.2 or 0.1 or 0.01, whose choice we discuss later in this section. Those methods are **step-by-step methods**, using the same formula in each step. Such formulas are suggested by the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots. \tag{2}$$

Formula (2) is the key idea that lets us develop Euler's method and its variant called—you guessed it—*improved Euler method*, also known as *Heun's method*. Let us start by deriving Euler's method.

For small h the higher powers $h^2, h^3, \cdots$ in (2) are very small. Dropping all of them gives the crude approximation

$$\begin{aligned} y(x + h) &\approx y(x) + hy'(x) \\ &= y(x) + hf(x, y) \end{aligned}$$

and the corresponding **Euler method** (or **Euler–Cauchy method**)

$$\boxed{y_{n+1} = y_n + hf(x_n, y_n)} \qquad (n = 0, 1, \cdots) \tag{3}$$

discussed in Sec. 1.2. Geometrically, this is an approximation of the curve of $y(x)$ by a polygon whose first side is tangent to this curve at x_0 (see Fig. 8 in Sec. 1.2).

Error of the Euler Method. Recall from calculus that Taylor's formula with remainder has the form

$$y(x + h) = y(x) + hy'(x) + \tfrac{1}{2} h^2 y''(\xi)$$

(where $x \leqq \xi \leqq x + h$). It shows that, in the Euler method, the *truncation error in each step* or **local truncation error** is proportional to h^2, written $O(h^2)$, where O suggests *order* (see also Sec. 20.1). Now, over a fixed x-interval in which we want to solve an ODE, the number of steps is proportional to $1/h$. Hence the *total error* or **global error** is proportional to $h^2(1/h) = h^1$. For this reason, the Euler method is called a **first-order method**. In addition, there are **roundoff errors** in this and other methods, which may affect the accuracy of the values $y_1, y_2, \cdots$ more and more as n increases.

Automatic Variable Step Size Selection in Modern Software. The idea of adaptive integration, as motivated and explained in Sec. 19.5, applies equally well to the numeric solution of ODEs. It now concerns automatically changing the step size h depending on the variability of $y' = f$ determined by

$$y'' = f' = f_x + f_y y' = f_x + f_y f. \tag{4*}$$

Accordingly, modern software automatically selects variable step sizes h_n so that the error of the solution will not exceed a given maximum size TOL (suggesting *tolerance*). Now for the Euler method, when the step size is $h = h_n$, the local error at x_n is about $\frac{1}{2}h_n^2|y''(\xi_n)|$. We require that this be equal to a given tolerance TOL,

$$\text{(a)}\quad \tfrac{1}{2}h_n^2|y''(\xi_n)| = \text{TOL}, \qquad \text{thus} \qquad \text{(b)}\quad h_n = \sqrt{\frac{2\ \text{TOL}}{|y''(\xi_n)|}}. \tag{4}$$

$y''(x)$ must not be zero on the interval J: $x_0 \leqq x = x_N$ on which the solution is wanted. Let K be the minimum of $|y''(x)|$ on J and assume that $K > 0$. Minimum $|y''(x)|$ corresponds to maximum $h = H = \sqrt{2\ \text{TOL}/K}$ by (4). Thus, $\sqrt{2\ \text{TOL}} = H\sqrt{K}$. We can insert this into (4b), obtaining by straightforward algebra

$$h_n = \varphi(x_n)H \qquad \text{where} \qquad \varphi(x_n) = \sqrt{\frac{K}{|y''(\xi_n)|}}. \tag{5}$$

For other methods, automatic step size selection is based on the same principle.

Improved Euler Method. Predictor, Corrector. Euler's method is generally much too inaccurate. For a large h (0.2) this is illustrated in Sec. 1.2 by the computation for

$$y' = y + x, \qquad y(0) = 0. \tag{6}$$

And for small h the computation becomes prohibitive; also, roundoff in so many steps may result in meaningless results. Clearly, methods of higher order and precision are obtained by taking more terms in (2) into account. But this involves an important practical problem. Namely, if we substitute $y' = f(x, y(x))$ into (2), we have

$$y(x + h) = y(x) + hf + \tfrac{1}{2}h^2 f' + \tfrac{1}{6}h^3 f'' + \cdots. \tag{2*}$$

Now y in f depends on x, so that we have f' as shown in (4*) and f'', f''' even much more cumbersome. The ***general strategy*** now is to avoid the computation of these derivatives and to replace it by computing f for one or several suitably chosen auxiliary values of (x, y). "Suitably" means that these values are chosen to make the order of the method as high as possible (to have high accuracy). Let us discuss two such methods that are of practical importance, namely, the improved Euler method and the (classical) Runge–Kutta method.

In each step of the **improved Euler method** we compute *two* values, first the ***predictor***

$$y^*_{n+1} = y_n + hf(x_n, y_n), \tag{7a}$$

which is an auxiliary value, and then the new y-value, the **corrector**

$$y_{n+1} = y_n + \tfrac{1}{2}h\left[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})\right]. \tag{7b}$$

Hence the improved Euler method is a predictor–corrector method: In each step we predict a value (7a) and then we correct it by (7b).

In algorithmic form, using the notations $k_1 = hf(x_n, y_n)$ in (7a) and $k_2 = hf(x_{n+1}, y^*_{n+1})$ in (7b), we can write this method as shown in Table 21.1.

Table 21.1 Improved Euler Method (Heun's Method)

ALGORITHM EULER (f, x_0, y_0, h, N)

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points $x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh$; here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.6).

INPUT: Initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$x_{n+1} = x_n + h$

$k_1 = hf(x_n, y_n)$

$k_2 = hf(x_{n+1}, y_n + k_1)$

$y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$

OUTPUT x_{n+1}, y_{n+1}

End

Stop

End EULER

EXAMPLE 1 Improved Euler Method. Comparison with Euler Method.

Apply the improved Euler method to the initial value problem (6), choosing $h = 0.2$ as in Sec. 1.2.

Solution. For the present problem we have in Table 21.1

$$k_1 = 0.2(x_n + y_n)$$

$$k_2 = 0.2(x_n + 0.2 + y_n + 0.2(x_n + y_n))$$

$$y_{n+1} = y_n + \frac{0.2}{2}(2.2x_n + 2.2y_n + 0.2) = y_n + 0.22(x_n + y_n) + 0.02.$$

Table 21.2 shows that our present results are much more accurate than those for Euler's method in Table 21.1 but at the cost of more computations. ■

Table 21.2 Improved Euler Method for (6). Errors

n	x_n	y_n	Exact Values (4D)	Error of Improved Euler	Error of Euler
0	0.0	0.0000	0.0000	0.0000	0.000
1	0.2	0.0200	0.0214	0.0014	0.021
2	0.4	0.0884	0.0918	0.0034	0.052
3	0.6	0.2158	0.2221	0.0063	0.094
4	0.8	0.4153	0.4255	0.0102	0.152
5	1.0	0.7027	0.7183	0.0156	0.230

Error of the Improved Euler Method. *The local error is of order h^3 and the global error of order h^2, so that the method is a* **second-order method**.

PROOF Setting $\tilde{f}_n = f(x_n, y(x_n))$ and using (2*) (after (6)), we have

$$y(x_n + h) - y(x_n) = h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}'_n + \tfrac{1}{6}h^3\tilde{f}''_n + \cdots. \tag{8a}$$

Approximating the expression in the brackets in (7b) by $\tilde{f}_n + \tilde{f}_{n+1}$ and again using the Taylor expansion, we obtain from (7b)

$$\begin{aligned} y_{n+1} - y_n &\approx \tfrac{1}{2}h\,[\tilde{f}_n + \tilde{f}_{n+1}] \\ &= \tfrac{1}{2}h\,[\tilde{f}_n + (\tilde{f}_n + h\tilde{f}'_n + \tfrac{1}{2}h^2\tilde{f}''_n + \cdots)] \\ &= h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}'_n + \tfrac{1}{4}h^3\tilde{f}''_n + \cdots \end{aligned} \tag{8b}$$

(where $' = d/dx_n$, etc.). Subtraction of (8b) from (8a) gives the local error

$$\frac{h^3}{6}\tilde{f}''_n - \frac{h^3}{4}\tilde{f}''_n + \cdots = -\frac{h^3}{12}\tilde{f}''_n + \cdots.$$

Since the number of steps over a fixed x-interval is proportional to $1/h$, the global error is of order $h^3/h = h^2$, so that the method is of second order. ■

Since the Euler method was an attractive pedagogical tool to teach the beginning of solving first-order ODEs numerically but had its drawbacks in terms of accuracy and could even produce wrong answers, we studied the improved Euler method and thereby introduced the idea of a predictor–corrector method. Although improved Euler is better than Euler, there are better methods that are used in industrial settings. Thus the practicing engineer has to know about the Runga–Kutta methods and its variants.

Runge–Kutta Methods (RK Methods)

A method of great practical importance and much greater accuracy than that of the improved Euler method is the *classical Runge–Kutta method of fourth order*, which we call briefly the **Runge–Kutta method.**[1] It is shown in Table 21.3. We see that in each step we first compute four auxiliary quantities k_1, k_2, k_3, k_4 and then the new value y_{n+1}. The method is well suited to the computer because it needs no special starting procedure, makes light demand on storage, and repeatedly uses the same straightforward computational procedure. It is numerically stable.

Note that, if f depends only on x, this method reduces to Simpson's rule of integration (Sec. 19.5). Note further that $k_1, \cdots, k_4$ depend on n and generally change from step to step.

Table 21.3 Classical Runge–Kutta Method of Fourth Order

ALGORITHM RUNGE–KUTTA (f, x_0, y_0, h, N).

This algorithm computes the solution of the initial value problem $y' = f(x, y), y(x_0) = y_0$ at equidistant points

(9) $$x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh;$$

here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.7).

INPUT: Function f, initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, 1, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$k_1 = hf(x_n, y_n)$

$k_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1)$

$k_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2)$

$k_4 = hf(x_n + h, y_n + k_3)$

$x_{n+1} = x_n + h$

$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$

OUTPUT x_{n+1}, y_{n+1}

End

Stop

End RUNGE–KUTTA

[1]Named after the German mathematicians KARL RUNGE (Sec. 19.4) and WILHELM KUTTA (1867–1944). Runge [*Math. Annalen* **46** (1895), 167–178], the German mathematician KARL HEUN (1859–1929) [*Zeitschr. Math. Phys.* **45** (1900), 23–38], and Kutta [*Zeitschr. Math. Phys.* **46** (1901), 435–453] developed various similar methods. Theoretically, there are infinitely many fourth-order methods using four function values per step. The method in Table 21.3 is most popular from a practical viewpoint because of its "symmetrical" form and its simple coefficients. It was given by Kutta.

EXAMPLE 2 Classical Runge–Kutta Method

Apply the Runge–Kutta method to the initial value problem in Example 1, choosing $h = 0.2$, as before, and computing five steps.

Solution. For the present problem we have $f(x, y) = x + y$. Hence

$$k_1 = 0.2(x_n + y_n), \qquad k_2 = 0.2(x_n + 0.1 + y_n + 0.5k_1),$$
$$k_3 = 0.2(x_n + 0.1 + y_n + 0.5k_2), \qquad k_4 = 0.2(x_n + 0.2 + y_n + k_3).$$

Table 21.4 shows the results and their errors, which are smaller by factors 10^3 and 10^4 than those for the two Euler methods. See also Table 21.5. We mention in passing that since the present $k_1, \cdots, k_4$ are simple, operations were saved by substituting k_1 into k_2, then k_2 into k_3, etc.; the resulting formula is shown in Column 4 of Table 21.4. Keep in mind that we have four function evaluations at each step. ■

Table 21.4 Runge–Kutta Method Applied to (4)

n	x_n	y_n	$0.2214(x_n + y_n) + 0.0214$	Exact Values (6D) $y = e^x - x - 1$	$10^6 \times$ Error of y_n
0	0.0	0	0.021400	0.000000	0
1	0.2	0.021400	0.070418	0.021403	3
2	0.4	0.091818	0.130289	0.091825	7
3	0.6	0.222107	0.203414	0.222119	12
4	0.8	0.425521	0.292730	0.425541	20
5	1.0	0.718251		0.718282	31

Table 21.5 Comparison of the Accuracy of the Three Methods under Consideration in the Case of the Initial Value Problem (4), with $h = 0.2$

x	$y = e^x - x - 1$	Error		
		Euler (Table 21.1)	Improved Euler (Table 21.3)	Runge–Kutta (Table 21.5)
0.2	0.021403	0.021	0.0014	0.000003
0.4	0.091825	0.052	0.0034	0.000007
0.6	0.222119	0.094	0.0063	0.000011
0.8	0.425541	0.152	0.0102	0.000020
1.0	0.718282	0.230	0.0156	0.000031

Error and Step Size Control. RKF (Runge–Kutta–Fehlberg)

The idea of adaptive integration (Sec. 19.5) has analogs for Runge–Kutta (and other) methods. In Table 21.3 for RK (Runge–Kutta), if we compute in each step approximations $\tilde{y}$ and $\tilde{\tilde{y}}$ with step sizes h and $2h$, respectively, the latter has error per step equal to $2^5 = 32$ times that of the former; however, since we have only half as many steps for $2h$, the actual factor is $2^5/2 = 16$, so that, say,

$$\epsilon^{(2h)} \approx 16\epsilon^{(h)} \qquad \text{and thus} \qquad y^{(h)} - y^{(2h)} = \epsilon^{(2h)} - \epsilon^{(h)} \approx (16 - 1)\epsilon^{(h)}.$$

Hence the error $\epsilon = \epsilon^{(h)}$ for step size h is about

$$\epsilon = \tfrac{1}{15}(\tilde{y} - \tilde{\tilde{y}}) \tag{10}$$

where $\tilde{y} - \tilde{\tilde{y}} = y^{(h)} - y^{(2h)}$, as said before. Table 21.6 illustrates (10) for the initial value problem

$$y' = (y - x - 1)^2 + 2, \qquad y(0) = 1, \tag{11}$$

the step size $h = 0.1$ and $0 \leqq x \leqq 0.4$. We see that the estimate is close to the actual error. This method of error estimation is simple but may be unstable.

Table 21.6 Runge–Kutta Method Applied to the Initial Value Problem (11) and Error Estimate (10). Exact Solution $y = \tan x + x + 1$

x	$\tilde{y}$ (Step size h)	$\tilde{\tilde{y}}$ (Step size $2h$)	Error Estimate (10)	Actual Error	Exact Solution (9D)
0.0	1.000000000	1.000000000	0.000000000	0.000000000	1.000000000
0.1	1.200334589			0.000000083	1.200334672
0.2	1.402709878	1.402707408	0.000000165	0.000000157	1.402710036
0.3	1.609336039			0.000000210	1.609336250
0.4	1.822792993	1.822788993	0.000000267	0.000000226	1.822793219

RKF. E. Fehlberg [*Computing* **6** (1970), 61–71] proposed and developed error control by using two RK methods of different orders to go from (x_n, y_n) to (x_{n+1}, y_{n+1}). The difference of the computed y-values at x_{n+1} gives an error estimate to be used for step size control. Fehlberg discovered two RK formulas that together need only six function evaluations per step. We present these formulas here because RKF has become quite popular. For instance, Maple uses it (also for systems of ODEs).

Fehlberg's fifth-order RK method is

$$y_{n+1} = y_n + \gamma_1 k_1 + \cdots + \gamma_6 k_6 \tag{12a}$$

with coefficient vector $\gamma = [\gamma_1 \cdots \gamma_6]$,

$$\gamma = \left[\tfrac{16}{135} \quad 0 \quad \tfrac{6656}{12,825} \quad \tfrac{28,561}{56,430} \quad -\tfrac{9}{50} \quad \tfrac{2}{55}\right]. \tag{12b}$$

His **fourth-order RK method** is

$$y^*_{n+1} = y_n + \gamma^*_1 k_1 + \cdots + \gamma^*_5 k_5 \tag{13a}$$

with coefficient vector

$$\gamma^* = \left[\tfrac{25}{216} \quad 0 \quad \tfrac{1408}{2565} \quad \tfrac{2197}{4104} \quad -\tfrac{1}{5}\right]. \tag{13b}$$

In both formulas we use only six different function evaluations altogether, namely,

$$
\begin{aligned}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf(x_n + \tfrac{1}{4}h, \quad y_n + \tfrac{1}{4}k_1) \\
k_3 &= hf(x_n + \tfrac{3}{8}h, \quad y_n + \tfrac{3}{32}k_1 + \tfrac{9}{32}k_2) \\
k_4 &= hf(x_n + \tfrac{12}{13}h, \quad y_n + \tfrac{1932}{2197}k_1 - \tfrac{7200}{2197}k_2 + \tfrac{7296}{2197}k_3) \\
k_5 &= hf(x_n + h, \quad y_n + \tfrac{439}{216}k_1 - 8k_2 + \tfrac{3680}{513}k_3 - \tfrac{845}{4104}k_4) \\
k_6 &= hf(x_n + \tfrac{1}{2}h, \quad y_n - \tfrac{8}{27}k_1 + 2k_2 - \tfrac{3544}{2565}k_3 + \tfrac{1859}{4104}k_4 - \tfrac{11}{40}k_5).
\end{aligned}
\tag{14}
$$

The difference of (12) and (13) gives the **error estimate**

$$\epsilon_{n+1} \approx y_{n+1} - y^*_{n+1} = \tfrac{1}{360}k_1 - \tfrac{128}{4275}k_3 - \tfrac{2197}{75{,}240}k_4 + \tfrac{1}{50}k_5 + \tfrac{2}{55}k_6. \tag{15}$$

EXAMPLE 3 Runge–Kutta–Fehlberg

For the initial value problem (11) we obtain from (12)–(14) with $h = 0.1$ in the first step the 12S-values

$$
\begin{aligned}
k_1 &= 0.200000000000 & k_2 &= 0.200062500000 \\
k_3 &= 0.200140756867 & k_4 &= 0.200856926154 \\
k_5 &= 0.201006676700 & k_6 &= 0.200250418651
\end{aligned}
$$

$$
\begin{aligned}
y_1^* &= 1.20033466949 \\
y_1 &= 1.20033467253
\end{aligned}
$$

and the error estimate

$$\epsilon_1 \approx y_1 - y_1^* = 0.00000000304.$$

The exact 12S-value is $y(0.1) = 1.20033467209$. Hence the actual error of y_1 is $-4.4 \cdot 10^{-10}$, smaller than that in Table 21.6 by a factor of 200. ■

Table 21.7 summarizes essential features of the methods in this section. It can be shown that these methods are *numerically stable* (definition in Sec. 19.1). They are **one-step methods** because in each step we use the data of just *one* preceding step, in contrast to **multistep methods** where in each step we use data from *several* preceding steps, as we shall see in the next section.

Table 21.7 Methods Considered and Their Order (= Their Global Error)

Method	Function Evaluation per Step	Global Error	Local Error
Euler	1	$O(h)$	$O(h^2)$
Improved Euler	2	$O(h^2)$	$O(h^3)$
RK (fourth order)	4	$O(h^4)$	$O(h^5)$
RKF	6	$O(h^5)$	$O(h^6)$

Backward Euler Method. Stiff ODEs

The **backward Euler formula** for numerically solving (1) is

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \qquad (n = 0, 1, \cdots). \tag{16}$$

This formula is obtained by evaluating the right side at the *new* location (x_{n+1}, y_{n+1}); this is called the **backward Euler scheme**. For known y_n it gives y_{n+1} *implicitly,* so it defines an **implicit method**, in contrast to the Euler method (3), which gives y_{n+1} explicitly. Hence (16) must be solved for y_{n+1}. How difficult this is depends on f in (1). For a linear ODE this provides no problem, as Example 4 (below) illustrates. The method is particularly useful for "stiff" ODEs, as they occur quite frequently in the study of vibrations, electric circuits, chemical reactions, etc. The situation of stiffness is roughly as follows; for details, see, for example, [E5], [E25], [E26] in App. 1.

Error terms of the methods considered so far involve a higher derivative. And we ask what happens if we let h *increase.* Now if the error (the derivative) grows fast but the desired solution also grows fast, nothing will happen. However, if that solution does not grow fast, then with growing h the error term can take over to an extent that the numeric result becomes completely nonsensical, as in Fig. 451. Such an ODE for which h must thus be restricted to small values, and the physical system the ODE models, are called **stiff**. This term is suggested by a mass–spring system with a stiff spring (spring with a large k; see Sec. 2.4). Example 4 illustrates that implicit methods remove the difficulty of increasing h in the case of stiffness: It can be shown that in the application of an implicit method the solution remains stable under any increase of h, although the accuracy decreases with increasing h.

EXAMPLE 4 Backward Euler Method. Stiff ODE

The initial value problem

$$y' = f(x, y) = -20hy + 20x^2 + 2x, \quad y(0) = 1$$

has the solution (verify!)

$$y = e^{-20x} + x^2.$$

The backward Euler formula (16) is

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) = y_n + h(-20y_{n+1} + 20x_{n+1}^2 + 2x_{n+1}).$$

Noting that $x_{n+1} = x_n + h$, taking the term $-20y_{n+1}$ to the left, and dividing, we obtain

$$y_{n+1} = \frac{y_n + h[20(x_n + h)^2 + 2(x_n + h)]}{1 + 20h}. \tag{16*}$$

The numeric results in Table 21.8 show the following.

Stability of the backward Euler method for $h = 0.05$ and also for $h = 0.2$ with an error increase by about a factor 4 for $h = 0.2$,

Stability of the Euler method for $h = 0.05$ but instability for $h = 0.1$ (Fig. 451),

Stability of RK for $h = 0.1$ but instability for $h = 0.2$.

This illustrates that the ODE is stiff. Note that even in the case of stability the approximation of the solution near $x = 0$ is poor. ■

Stiffness will be considered further in Sec. 21.3 in connection with systems of ODEs.

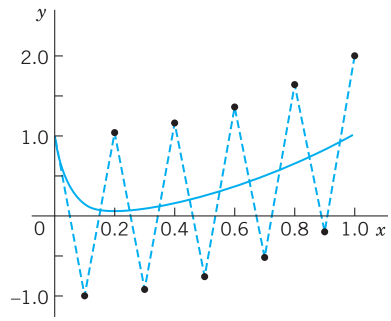


Fig. 451. Euler method with $h = 0.1$ for the stiff ODE in Example 4 and exact solution

Table 21.8 Backward Euler Method (BEM) for Example 6. Comparison with Euler and RK

x	BEM $h = 0.05$	BEM $h = 0.2$	Euler $h = 0.05$	Euler $h = 0.1$	RK $h = 0.1$	RK $h = 0.2$	Exact
0.0	1.00000	1.00000	1.00000	1.00000	1.00000	1.000	1.00000
0.1	0.26188		0.00750	-1.00000	0.34500		0.14534
0.2	0.10484	0.24800	0.03750	1.04000	0.15333	5.093	0.05832
0.3	0.10809		0.08750	-0.92000	0.12944		0.09248
0.4	0.16640	0.20960	0.15750	1.16000	0.17482	25.48	0.16034
0.5	0.25347		0.24750	-0.76000	0.25660		0.25004
0.6	0.36274	0.37792	0.35750	1.36000	0.36387	127.0	0.36001
0.7	0.49256		0.48750	-0.52000	0.49296		0.49001
0.8	0.64252	0.65158	0.63750	1.64000	0.64265	634.0	0.64000
0.9	0.81250		0.80750	-0.20000	0.81255		0.81000
1.0	1.00250	1.01032	0.99750	2.00000	1.00252	3168	1.00000

PROBLEM SET 21.1

1–4 EULER METHOD

Do 10 steps. Solve exactly. Compute the error. Show details.

1. $y' + 0.2y = 0, \quad y(0) = 5, \quad h = 0.2$

2. $y' = \frac{1}{2}\pi\sqrt{1 - y^2}, \quad y(0) = 0, \quad h = 0.1$

3. $y' = (y - x)^2, \quad y(0) = 0, \quad h = 0.1$

4. $y' = (y + x)^2, \quad y(0) = 0, \quad h = 0.1$

5–10 IMPROVED EULER METHOD

Do 10 steps. Solve exactly. Compute the error. Show details.

5. $y' = y, \quad y(0) = 1, \quad h = 0.1$

6. $y' = 2(1 + y^2), \quad y(0) = 0, \quad h = 0.05$

7. $y' - xy^2 = 0, \quad y(0) = 1, \quad h = 0.1$

8. Logistic population model. $y' = y - y^2, \quad y(0) = 0.2, \quad h = 0.1$

9. Do Prob. 7 using Euler's method with $h = 0.1$ and compare the accuracy.

10. Do Prob. 7 using the improved Euler method, 20 steps with $h = 0.05$. Compare.

11–17 CLASSICAL RUNGE–KUTTA METHOD OF FOURTH ORDER

Do 10 steps. Compare as indicated. Show details.

11. $y' - xy^2 = 0, \quad y(0) = 1, \quad h = 0.1$. Compare with Prob. 7. Apply the error estimate (10) to y_{10}.

12. $y' = y - y^2, \quad y(0) = 0.2, \quad h = 0.1$. Compare with Prob. 8.

13. $y' = 1 + y^2, \quad y(0) = 0, \quad h = 0.1$

14. $y' = (1 - x^{-1})y, \quad y(1) = 1, \quad h = 0.1$

15. $y' + y \tan x = \sin 2x, \quad y(0) = 1, \quad h = 0.1$

16. Do Prob. 15 with $h = 0.2$, 5 steps, and compare the errors with those in Prob. 15.

17. $y' = 4x^3y^2, \quad y(0) = 0.5, \quad h = 0.1$

18. **Kutta's third-order method** is defined by $y_{n+1} = y_n + \frac{1}{6}(k_1 + 4k_2 + k_3^*)$ with k_1 and k_2 as in RK (Table 21.3) and $k_3^* = hf(x_{n+1}, y_n - k_1 + 2k_2)$. Apply this method to (4) in (6). Choose $h = 0.2$ and do 5 steps. Compare with Table 21.5.

19. **CAS EXPERIMENT. Euler–Cauchy vs. RK.** Consider the initial value problem

$$y' = (y - 0.01x^2)^2 \sin(x^2) + 0.02x, \qquad y(0) = 0.4 \tag{17}$$

(solution: $y = 1/[2.5 - S(x)] + 0.01x^2$ where $S(x)$ is the Fresnel integral (38) in App. 3.1).

(a) Solve (17) by Euler, improved Euler, and RK methods for $0 \leqq x \leqq 5$ with step $h = 0.2$. Compare the errors for $x = 1, 3, 5$ and comment.

(b) Graph solution curves of the ODE in (17) for various positive and negative initial values.

(c) Do a similar experiment as in (a) for an initial value problem that has a monotone increasing or monotone decreasing solution. Compare the behavior of the error with that in (a). Comment.

20. **CAS EXPERIMENT. RKF. (a)** Write a program for RKF that gives x_n, y_n, the estimate (10), and, if the solution is known, the actual error ϵ_n.

(b) Apply the program to Example 3 in the text (10 steps, $h = 0.1$).

(c) ϵ_n in (b) gives a relatively good idea of the size of the actual error. Is this typical or accidental? Find out, by experimentation with other problems, on what properties of the ODE or solution this might depend.

21.2 Multistep Methods

In a **one-step method** we compute y_{n+1} using only a single step, namely, the previous value y_n. *One-step methods are* "**self-starting**," they need no help to get going because they obtain y_1 from the initial value y_0, etc. All methods in Sec. 21.1 are one-step.

In contrast, a **multistep method** uses, in each step, values from two or more previous steps. These methods are motivated by the expectation that the additional information will increase accuracy and stability. But to get started, one needs values, say, y_0, y_1, y_2, y_3 in a 4-step method, obtained by Runge–Kutta or another accurate method. Thus, multistep methods are not self-starting. Such methods are obtained as follows.

Adams–Bashforth Methods

We consider an initial value problem

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

as before, with f such that the problem has a unique solution on some open interval containing x_0. We integrate $y' = f(x, y)$ from x_n to $x_{n+1} = x_n + h$. This gives

$$\int_{x_n}^{x_{n+1}} y'(x)\, dx = y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx.$$

Now comes the main idea. We replace $f(x, y(x))$ by an interpolation polynomial $p(x)$ (see Sec. 19.3), so that we can later integrate. This gives approximations y_{n+1} of $y(x_{n+1})$ and y_n of $y(x_n)$,

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(x)\, dx. \tag{2}$$

Different choices of $p(x)$ will now produce different methods. We explain the principle by taking a cubic polynomial, namely, the polynomial $p_3(x)$ that at (equidistant)

$$x_n, \qquad x_{n-1}, \qquad x_{n-2}, \qquad x_{n-3}$$

has the respective values

$$\begin{aligned} f_n &= f(x_n, y_n) \\ f_{n-1} &= f(x_{n-1}, y_{n-1}) \\ f_{n-2} &= f(x_{n-2}, y_{n-2}) \\ f_{n-3} &= f(x_{n-3}, y_{n-3}). \end{aligned} \tag{3}$$

This will lead to a practically useful formula. We can obtain $p_3(x)$ from Newton's backward difference formula (18), Sec. 19.3:

$$p_3(x) = f_n + r\nabla f_n + \tfrac{1}{2}r(r+1)\nabla^2 f_n + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_n$$

where

$$r = \frac{x - x_n}{h}.$$

We integrate $p_3(x)$ over x from x_n to $x_{n+1} = x_n + h$, thus over r from 0 to 1. Since

$$x = x_n + hr, \qquad \text{we have} \qquad dx = h\,dr.$$

The integral of $\frac{1}{2}r(r+1)$ is $\frac{5}{12}$ and that of $\frac{1}{6}r(r+1)(r+2)$ is $\frac{3}{8}$. We thus obtain

$$\int_{x_n}^{x_{n+1}} p_3\,dx = h\int_0^1 p_3\,dr = h\left(f_n + \frac{1}{2}\nabla f_n + \frac{5}{12}\nabla^2 f_n + \frac{3}{8}\nabla^3 f_n\right). \tag{4}$$

It is practical to replace these differences by their expressions in terms of f:

$$\begin{aligned} \nabla f_n &= f_n - f_{n-1} \\ \nabla^2 f_n &= f_n - 2f_{n-1} + f_{n-2} \\ \nabla^3 f_n &= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}. \end{aligned}$$

We substitute this into (4) and collect terms. This gives the multistep formula of the **Adams–Bashforth method**[2] *of fourth order*

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \tag{5}$$

[2]Named after JOHN COUCH ADAMS (1819–1892), English astronomer and mathematician, one of the predictors of the existence of the planet Neptune (using mathematical calculations), director of the Cambridge Observatory; and FRANCIS BASHFORTH (1819–1912), English mathematician.

It expresses the new value y_{n+1} [approximation of the solution y of (1) at x_{n+1}] in terms of 4 values of f computed from the y-values obtained in the preceding 4 steps. The local truncation error is of order h^5, as can be shown, so that the global error is of order h^4; hence (5) does define a fourth-order method.

Adams–Moulton Methods

Adams–Moulton methods are obtained if for $p(x)$ in (2) we choose a polynomial that interpolates $f(x, y(x))$ at $x_{n+1}, x_n, x_{n-1}, \cdots$ (as opposed to $x_n, x_{n-1}, \cdots$ used before; this is the main point). We explain the principle for the cubic polynomial $\widetilde{p}_3(x)$ that interpolates at $x_{n+1}, x_n, x_{n-1}, x_{n-2}$. (Before we had $x_n, x_{n-1}, x_{n-2}, x_{n-3}$.) Again using (18) in Sec. 19.3 but now setting $r = (x - x_{n+1})/h$, we have

$$\widetilde{p}_3(x) = f_{n+1} + r\nabla f_{n+1} + \tfrac{1}{2}r(r+1)\nabla^2 f_{n+1} + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_{n+1}.$$

We now integrate over x from x_n to x_{n+1} as before. This corresponds to integrating over r from -1 to 0. We obtain

$$\int_{x_n}^{x_{n+1}} \widetilde{p}_3(x)\, dx = h\left(f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1} - \frac{1}{24}\nabla^3 f_{n+1}\right).$$

Replacing the differences as before gives

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \widetilde{p}_3(x)\, dx = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}). \tag{6}$$

This is usually called an **Adams–Moulton formula.**[3] It is an **implicit formula** because $f_{n+1} = f(x_{n+1}, y_{n+1})$ appears on the right, so that it defines y_{n+1} only *implicitly,* in contrast to (5), which is an **explicit formula**, not involving y_{n+1} on the right. To use (6) we must ***predict*** a value y^*_{n+1}, for instance, by using (5), that is,

$$y^*_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \tag{7a}$$

The ***corrected*** new value y_{n+1} is then obtained from (6) with f_{n+1} replaced by $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$ and the other f's as in (6); thus,

$$y_{n+1} = y_n + \frac{h}{24}(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}). \tag{7b}$$

This **predictor–corrector method** (7a), (7b) is usually called the **Adams–Moulton method** *of fourth order.* It has the advantage over RK that (7) gives the error estimate

$$\epsilon_{n+1} \approx \tfrac{1}{15}(y_{n+1} - y^*_{n+1}),$$

as can be shown. This is the analog of (10) in Sec. 21.1.

---

[3]FOREST RAY MOULTON (1872–1952), American astronomer at the University of Chicago. For ADAMS see footnote 2.

Sometimes the name Adams–Moulton method is reserved for the method with *several* corrections per step by (7b) until a specific accuracy is reached. Popular codes exist for both versions of the method.

Getting Started. In (5) we need f_0, f_1, f_2, f_3. Hence from (3) we see that we must first compute y_1, y_2, y_3 by some other method of comparable accuracy, for instance, by RK or by RKF. For other choices see Ref. [E26] listed in App. 1.

EXAMPLE 1 Adams–Bashforth Prediction (7a), Adams–Moulton Correction (7b)

Solve the initial value problem

$$y' = x + y, \qquad y(0) = 0 \tag{8}$$

by (7a), (7b) on the interval $0 \leqq x \leqq 2$, choosing $h = 0.2$.

Solution. The problem is the same as in Examples 1 and 2, Sec. 21.1, so that we can compare the results. We compute starting values y_1, y_2, y_3 by the classical Runge–Kutta method. Then in each step we predict by (7a) and make one correction by (7b) before we execute the next step. The results are shown and compared with the exact values in Table 21.9. We see that the corrections improve the accuracy considerably. This is typical. ■

Table 21.9 Adams–Moulton Method Applied to the Initial Value Problem (8); Predicted Values Computed by (7a) and Corrected Values by (7b)

n	x_n	Starting y_n	Predicted y_n^*	Corrected y_n	Exact Values	$10^6 \cdot$ Error of y_n
0	0.0	0.000000			0.000000	0
1	0.2	0.021400			0.021403	3
2	0.4	0.091818			0.091825	7
3	0.6	0.222107			0.222119	12
4	0.8		0.425361	0.425529	0.425541	12
5	1.0		0.718066	0.718270	0.718282	12
6	1.2		1.119855	1.120106	1.120117	11
7	1.4		1.654885	1.655191	1.655200	9
8	1.6		2.352653	2.353026	2.353032	6
9	1.8		3.249190	3.249646	3.249647	1
10	2.0		4.388505	4.389062	4.389056	−6

Comments on Comparison of Methods. An Adams–Moulton formula is generally much more accurate than an Adams–Bashforth formula of the same order. This justifies the greater complication and expense in using the former. The method (7a), (7b) is ***numerically stable***, whereas the exclusive use of (7a) might cause instability. Step size control is relatively simple. If $|\text{Corrector} - \text{Predictor}| > \text{TOL}$, use interpolation to generate "old" results at half the current step size and then try $h/2$ as the new step.

Whereas the Adams–Moulton formula (7a), (7b) needs only 2 evaluations per step, Runge–Kutta needs 4; however, with Runge–Kutta one may be able to take a step size more than twice as large, so that a comparison of this kind (widespread in the literature) is meaningless.

For more details, see Refs. [E25], [E26] listed in App. 1.

PROBLEM SET 21.2

1–10 ADAMS–MOULTON METHOD

Solve the initial value problem by Adams–Moulton (7a), (7b), 10 steps with 1 correction per step. Solve exactly and compute the error. Use RK where no starting values are given.

1. $y' = y, \quad y(0) = 1, \quad h = 0.1, \quad (1.105171, 1.221403, 1.349858)$
2. $y' = 2xy, \quad y(0) = 1, \quad h = 0.1$
3. $y' = 1 + y^2, \quad y(0) = 0, \quad h = 0.1, \quad (0.100335, 0.202710, 0.309336)$
4. Do Prob. 2 by RK, 5 steps, $h = 0.2$. Compare the errors.
5. Do Prob. 3 by RK, 5 steps, $h = 0.2$. Compare the errors.
6. $y' = (y - x - 1)^2 + 2, \quad y(0) = 1, \quad h = 0.1,$ 10 steps
7. $y' = 3y - 12y^2, \quad y(0) = 0.2, \quad h = 0.1$
8. $y' = 1 - 4y^2, \quad y(0) = 0, \quad h = 0.1$
9. $y' = 3x^2(1 + y), \quad y(0) = 0, \quad h = 0.05$
10. $y' = x/y, \quad y(1) = 3, \quad h = 0.2$
11. Do and show the calculations leading to (4)–(7) in the text.
12. **Quadratic polynomial.** Apply the method in the text to a polynomial of second degree. Show that this leads to the predictor and corrector formulas

$$y^*_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}),$$
$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}).$$

13. Using Prob. 12, solve $y' = 2xy, \quad y(0) = 1$ (10 steps, $h = 0.1$, RK starting values). Compare with the exact solution and comment.
14. How much can you reduce the error in Prob. 13 by halfing h (20 steps, $h = 0.05$)? First guess, then compute.
15. **CAS PROJECT. Adams–Moulton. (a) Accurate starting** is important in (7a), (7b). Illustrate this in Example 1 of the text by using starting values from the improved Euler–Cauchy method and compare the results with those in Table 21.8.

(b) How much does the error in Prob. 11 decrease if you use exact starting values (instead of RK values)?

(c) Experiment to find out for what ODEs poor starting is very damaging and for what ODEs it is not.

(d) The classical **RK method** often gives the same accuracy with step $2h$ as Adams–Moulton with step h, so that the total number of function evaluations is the same in both cases. Illustrate this with Prob. 8. (Hence corresponding comparisons in the literature in favor of Adams–Moulton are not valid. See also Probs. 6 and 7.)

21.3 Methods for Systems and Higher Order ODEs

Initial value problems for first-order systems of ODEs are of the form

$$\textbf{(1)} \qquad \mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \mathbf{y}(x_0) = \mathbf{y}_0,$$

in components

$$\begin{aligned}
y_1' &= f_1(x, y_1, \cdots, y_m), & y_1(x_0) &= y_{10}\\
y_2' &= f_2(x, y_1, \cdots, y_m), & y_2(x_0) &= y_{20}\\
&\cdots\cdots\cdots\cdots\cdots\cdots & &\cdots\cdots\cdots\\
y_m' &= f_m(x, y_1, \cdots, y_m). & y_m(x_0) &= y_{m0}.
\end{aligned}$$

Here, $\mathbf{f}$ is assumed to be such that the problem has a unique solution $\mathbf{y}(x)$ on some open x-interval containing x_0. Our discussion will be independent of Chap. 4 on systems.

Before explaining solution methods it is important to note that (1) includes initial value problems for single mth-order ODEs,

$$y^{(m)} = f(x, y, y', y'', \cdots, y^{(m-1)}) \tag{2}$$

and initial conditions $y(x_0) = K_1, y'(x_0) = K_2, \cdots, y^{(m-1)}(x_0) = K_m$ as special cases.

Indeed, the connection is achieved by setting

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \quad \cdots, \quad y_m = y^{(m-1)}. \tag{3}$$

Then we obtain the system

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{m-1}' &= y_m \\ y_m' &= f(x, y_1, \cdots, y_m) \end{aligned} \tag{4}$$

and the initial conditions $y_1(x_0) = K_1, \quad y_2(x_0) = K_2, \quad \cdots, \quad y_m(x_0) = K_m$.

Euler Method for Systems

Methods for single first-order ODEs can be extended to systems (1) simply by writing vector functions $\mathbf{y}$ and $\mathbf{f}$ instead of scalar functions y and f, whereas x remains a scalar variable.

We begin with the Euler method. Just as for a single ODE, this method will not be accurate enough for practical purposes, but it nicely illustrates the extension principle.

EXAMPLE 1 Euler Method for a Second-Order ODE. Mass–Spring System

Solve the initial value problem for a damped mass–spring system

$$y'' + 2y' + 0.75y = 0, \qquad y(0) = 3, \qquad y'(0) = -2.5$$

by the Euler method for systems with step $h = 0.2$ for x from 0 to 1 (where x is time).

Solution. The **Euler method** (3), Sec. 21.1, generalizes to systems in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n), \tag{5}$$

in components

$$\begin{aligned} y_{1,n+1} &= y_{1,n} + hf_1(x_n, y_{1,n}, y_{2,n}) \\ y_{2,n+1} &= y_{2,n} + hf_2(x_n, y_{1,n}, y_{2,n}) \end{aligned}$$

and similarly for systems of more than two equations. By (4) the given ODE converts to the system

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2) = y_2 \\ y_2' &= f_2(x, y_1, y_2) = -2y_2 - 0.75y_1. \end{aligned}$$

Hence (5) becomes

$$y_{1,n+1} = y_{1,n} + 0.2y_{2,n}$$
$$y_{2,n+1} = y_{2,n} + 0.2(-2y_{2,n} - 0.75y_{1,n}).$$

The initial conditions are $y(0) = y_1(0) = 3$, $y'(0) = y_2(0) = -2.5$. The calculations are shown in Table 21.10. As for single ODEs, the results would not be accurate enough for practical purposes. The example merely serves to illustrate the method because the problem can be readily solved exactly,

$$y = y_1 = 2e^{-0.5x} + e^{-1.5x}, \qquad \text{thus} \qquad y' = y_2 = -e^{-0.5x} - 1.5e^{-1.5x}.$$ ■

Table 21.10 Euler Method for Systems in Example 1 (Mass–Spring System)

n	x_n	$y_{1,n}$	y_1 Exact (5D)	Error $\epsilon_1 = y_1 - y_{1,n}$	$y_{2,n}$	y_2 Exact (5D)	Error $\epsilon_2 = y_2 - y_{2,n}$
0	0.0	3.00000	3.00000	0.00000	−2.50000	−2.50000	0.00000
1	0.2	2.50000	2.55049	0.05049	−1.95000	−2.01606	−0.06606
2	0.4	2.11000	2.18627	0.76270	−1.54500	−1.64195	−0.09695
3	0.6	1.80100	1.88821	0.08721	−1.24350	−1.35067	−0.10717
4	0.8	1.55230	1.64183	0.08953	−1.01625	−1.12211	−0.10586
5	1.0	1.34905	1.43619	0.08714	−0.84260	−0.94123	−0.09863

Runge–Kutta Methods for Systems

As for Euler methods, we obtain RK methods for an initial value problem (1) simply by writing vector formulas for vectors with m components, which, for $m = 1$, reduce to the previous scalar formulas.

Thus, for the *classical* **RK method** *of fourth order* in Table 21.3, we obtain

(6a)
$$\mathbf{y}(x_0) = \mathbf{y}_0 \qquad \text{(Initial values)}$$

and for each step $n = 0, 1, \cdots, N - 1$ we obtain the 4 auxiliary quantities

(6b)
$$\begin{aligned}
\mathbf{k}_1 &= h\mathbf{f}(x_n, & &\mathbf{y}_n) \\
\mathbf{k}_2 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, & &\mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_1) \\
\mathbf{k}_3 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, & &\mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_2) \\
\mathbf{k}_4 &= h\mathbf{f}(x_n + h, & &\mathbf{y}_n + \mathbf{k}_3)
\end{aligned}$$

and the new value [approximation of the solution $\mathbf{y}(x)$ at $x_{n+1} = x_0 + (n + 1)h$]

(6c)
$$\mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4).$$

EXAMPLE 2 RK Method for Systems. Airy's Equation. Airy Function Ai(x)

Solve the initial value problem

$$y'' = xy, \qquad y(0) = 1/(3^{2/3} \cdot \Gamma(\tfrac{2}{3})) = 0.35502805, \qquad y'(0) = -1/(3^{1/3} \cdot \Gamma(\tfrac{1}{3})) = -0.25881940$$

by the Runge–Kutta method for systems with $h = 0.2$; do 5 steps. This is **Airy's equation**,[4] which arose in optics (see Ref. [A13], p. 188, listed in App. 1). Γ is the gamma function (see App. A3.1). The initial conditions are such that we obtain a standard solution, the **Airy function** $\mathrm{Ai}(x)$, a special function that has been thoroughly investigated; for numeric values, see Ref. [GenRef1], pp. 446, 475.

Solution. For $y'' = xy$, setting $y_1 = y$, $y_2 = y_1' = y'$ we obtain the system (4)

$$y_1' = y_2$$
$$y_2' = xy_1.$$

Hence $\mathbf{f} = [f_1 \quad f_2]^{\mathsf{T}}$ in (1) has the components $f_1(x, y) = y_2$, $f_2(x, y) = xy_1$. We now write (6) in components. The initial conditions (6a) are $y_{1,0} = 0.35502805$, $y_{2,0} = -0.25881940$. In (6b) we have fewer subscripts by simply writing $\mathbf{k}_1 = \mathbf{a}$, $\mathbf{k}_2 = \mathbf{b}$, $\mathbf{k}_3 = \mathbf{c}$, $\mathbf{k}_4 = \mathbf{d}$, so that $\mathbf{a} = [a_1 \quad a_2]^{\mathsf{T}}$, etc. Then (6b) takes the form

$$\text{(6b*)}\qquad \begin{aligned} \mathbf{a} &= h\begin{bmatrix} y_{2,n} \\ x_n y_{1,n} \end{bmatrix} \\ \mathbf{b} &= h\begin{bmatrix} y_{2,n} + \frac{1}{2}a_2 \\ (x_n + \frac{1}{2}h)(y_{1,n} + \frac{1}{2}a_1) \end{bmatrix} \\ \mathbf{c} &= h\begin{bmatrix} y_{2,n} + \frac{1}{2}b_2 \\ (x_n + \frac{1}{2}h)(y_{1,n} + \frac{1}{2}b_1) \end{bmatrix} \\ \mathbf{d} &= h\begin{bmatrix} y_{2,n} + c_2 \\ (x_n + h)(y_{1,n} + c_1) \end{bmatrix}. \end{aligned}$$

For example, the second component of $\mathbf{b}$ is obtained as follows. $\mathbf{f}(x, y)$ has the second component $f_2(x, y) = xy_1$. Now in $\mathbf{b}$ $(= \mathbf{k}_2)$ the first argument is

$$x = x_n + \tfrac{1}{2}h.$$

The second argument in $\mathbf{b}$ is

$$\mathbf{y} = \mathbf{y}_n + \tfrac{1}{2}\mathbf{a},$$

and the first component of this is

$$y_1 = y_{1,n} + \tfrac{1}{2}a_1.$$

Together,

$$xy_1 = (x_n + \tfrac{1}{2}h)(y_{1,n} + \tfrac{1}{2}a_1).$$

Similarly for the other components in (6b*). Finally,

$$\text{(6c*)}\qquad \mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{a} + 2\mathbf{b} + 2\mathbf{c} + \mathbf{d}).$$

Table 21.11 shows the values $y(x) = y_1(x)$ of the Airy function $\mathrm{Ai}(x)$ and of its derivative $y'(x) = y_2(x)$ as well as of the (rather small!) error of $y(x)$. ■

[4]Named after Sir GEORGE BIDELL AIRY (1801–1892), English mathematician, who is known for his work in elasticity and in PDEs.

Table 21.11 RK Method for Systems: Values $y_{1,n}(x_n)$ of the Airy Function Ai(x) in Example 2

n	x_n	$y_{1,n}(x_n)$	$y_1(x_n)$ Exact (8D)	$10^8 \cdot$ Error of y_1	$y_{2,n}(x_n)$
0	0.0	0.35502805	0.35502805	0	−0.25881940
1	0.2	0.30370303	0.30370315	12	−0.25240464
2	0.4	0.25474211	0.25474235	24	−0.23583073
3	0.6	0.20979973	0.20980006	33	−0.21279185
4	0.8	0.16984596	0.16984632	36	−0.18641171
5	1.0	0.13529207	0.13529242	35	−0.15914687

Runge–Kutta–Nyström Methods (RKN Methods)

RKN methods are direct extensions of RK methods (Runge–Kutta methods) to second-order ODEs $y'' = f(x, y, y')$, as given by the Finnish mathematician E. J. Nyström [*Acta Soc. Sci. fenn.,* 1925, L, No. 13]. The best known of these uses the following formulas, where $n = 0, 1, \cdots, N-1$ (N the number of steps):

$$\text{(7a)}\qquad \begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n, y_n') \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y_n' + k_1) && \text{where } K = \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1) \\ k_3 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y_n' + k_2) \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + L, y_n' + 2k_3) && \text{where } L = h(y_n' + k_3). \end{aligned}$$

From this we compute the approximation y_{n+1} of $y(x_{n+1})$ at $x_{n+1} = x_0 + (n+1)h$,

$$\text{(7b)}\qquad y_{n+1} = y_n + h(y_n' + \tfrac{1}{3}(k_1 + k_2 + k_3)),$$

and the approximation y_{n+1}' of the derivative $y'(x_{n+1})$ needed in the next step,

$$\text{(7c)}\qquad y_{n+1}' = y_n' + \tfrac{1}{3}(k_1 + 2k_2 + 2k_3 + k_4).$$

RKN for ODEs $y'' = f(x, y)$ Not Containing y'. Then $k_2 = k_3$ in (7), which makes the method particularly advantageous and reduces (7a)–(7c) to

$$\textbf{(7*)}\qquad \begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n) \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1)) = k_3 \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + h(y_n' + k_2)) \\ y_{n+1} &= y_n + h(y_n' + \tfrac{1}{3}(k_1 + 2k_2)) \\ y_{n+1}' &= y_n' + \tfrac{1}{3}(k_1 + 4k_2 + k_4). \end{aligned}$$

EXAMPLE 3 RKN Method. Airy's Equation. Airy Function Ai(x)

For the problem in Example 2 and $h = 0.2$ as before we obtain from (7*) simply $k_1 = 0.1x_ny_n$ and

$$k_2 = k_3 = 0.1(x_n + 0.1)(y_n + 0.1y_n' + 0.05k_1), \qquad k_4 = 0.1(x_n + 0.2)(y_n + 0.2y_n' + 0.2k_2).$$

Table 21.12 shows the results. The accuracy is the same as in Example 2, but the work was much less. ■

Table 21.12 Runge–Kutta–Nyström Method Applied to Airy's Equation, Computation of the Airy Function $y = \mathrm{Ai}(x)$

x_n	y_n	y'_n	$y(x)$ Exact (8D)	$10^8 \cdot$ Error of y_n
0.0	0.35502805	−0.25881940	0.35502805	0
0.2	0.30370304	−0.25240464	0.30370315	11
0.4	0.25474211	−0.23583070	0.25474235	24
0.6	0.20979974	−0.21279172	0.20980006	32
0.8	0.16984599	−0.18641134	0.16984632	33
1.0	0.13529218	−0.15914609	0.13529242	24

Our work in Examples 2 and 3 also illustrates that usefulness of methods for ODEs in the computation of values of "**higher transcendental functions**."

Backward Euler Method for Systems. Stiff Systems

The backward Euler formula (16) in Sec. 21.1 generalizes to systems in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}) \qquad (n = 0, 1, \cdots). \tag{8}$$

This is again an implicit method, giving $\mathbf{y}_{n+1}$ implicitly for given $\mathbf{y}_n$. Hence (8) must be solved for $\mathbf{y}_{n+1}$. For a linear system this is shown in the next example. This example also illustrates that, similar to the case of a single ODE in Sec. 21.1, the method is very useful for **stiff systems**. These are systems of ODEs whose matrix has eigenvalues λ of very different magnitudes, having the effect that, just as in Sec. 21.1, the step in direct methods, RK for example, cannot be increased beyond a certain threshold without losing stability. ($\lambda = -1$ and -10 in Example 4, but larger differences do occur in applications.)

EXAMPLE 4 Backward Euler Method for Systems of ODEs. Stiff Systems

Compare the backward Euler method (8) with the Euler and the RK methods for numerically solving the initial value problem

$$y'' + 11y' + 10y = 10x + 11, \qquad y(0) = 2, \qquad y'(0) = -10$$

converted to a system of first-order ODEs.

Solution. The given problem can easily be solved, obtaining

$$y = e^{-x} + e^{-10x} + x$$

so that we can compute errors. Conversion to a system by setting $y = y_1, y' = y_2$ [see (4)] gives

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2 \\ y_2' &= -10y_1 - 11y_2 + 10x + 11 & y_2(0) &= -10. \end{aligned}$$

The coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -10 & -11 \end{bmatrix} \quad \text{has the characteristic determinant} \quad \begin{vmatrix} -\lambda & 1 \\ -10 & -\lambda - 11 \end{vmatrix}$$

whose value is $\lambda^2 + 11\lambda + 10 = (\lambda + 1)(\lambda + 10)$. Hence the eigenvalues are -1 and -10 as claimed above. The backward Euler formula is

$$\mathbf{y}_{n+1} = \begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \end{bmatrix} = \begin{bmatrix} y_{1,n} \\ y_{2,n} \end{bmatrix} + h \begin{bmatrix} y_{2,n+1} \\ -10y_{1,n+1} - 11y_{2,n+1} + 10x_{n+1} + 11 \end{bmatrix}.$$

Reordering terms gives the linear system in the unknowns $y_{1,n+1}$ and $y_{2,n+1}$

$$\begin{aligned} y_{1,n+1} - \qquad\qquad hy_{2,n+1} &= y_{1,n} \\ 10hy_{1,n+1} + (1 + 11h)y_{2,n+1} &= y_{2,n} + 10h(x_n + h) + 11h. \end{aligned}$$

The coefficient determinant is $D = 1 + 11h + 10h^2$, and Cramer's rule (in Sec. 7.6) gives the solution

$$\mathbf{y}_{n+1} = \frac{1}{D} \begin{bmatrix} (1 + 11h)y_{1,n} + hy_{2,n} + 10h^2 x_n + 11h^2 + 10h^3 \\ -10hy_{1,n} + \quad y_{2,n} + \quad 10hx_n + \quad 11h + 10h^2 \end{bmatrix}.$$

Table 21.13 Backward Euler Method (BEM) for Example 4. Comparison with Euler and RK

x	BEM $h = 0.2$	BEM $h = 0.4$	Euler $h = 0.1$	Euler $h = 0.2$	RK $h = 0.2$	RK $h = 0.3$	Exact
0.0	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000
0.2	1.36667		1.01000	0.00000	1.35207		1.15407
0.4	1.20556	1.31429	1.56100	2.04000	1.18144		1.08864
0.6	1.21574		1.13144	0.11200	1.18585	3.03947	1.15129
0.8	1.29460	1.35020	1.23047	2.20960	1.26168		1.24966
1.0	1.40599		1.34868	0.32768	1.37200		1.36792
1.2	1.53627	1.57243	1.48243	2.46214	1.50257	5.07569	1.50120
1.4	1.67954		1.62877	0.60972	1.64706		1.64660
1.6	1.83272	1.86191	1.78530	2.76777	1.80205		1.80190
1.8	1.99386		1.95009	0.93422	1.96535	8.72329	1.96530
2.0	2.16152	2.18625	2.12158	3.10737	2.13536		2.13534

Table 21.13 shows the following.

Stability of the backward Euler method for $h = 0.2$ and 0.4 (and in fact for any h; try $h = 5.0$) with decreasing accuracy for increasing h

Stability of the Euler method for $h = 0.1$ but instability for $h = 0.2$

Stability of RK for $h = 0.2$ but instability for $h = 0.3$

Figure 452 shows the Euler method for $h = 0.18$, an interesting case with initial jumping (for about $x > 3$) but later monotone following the solution curve of $y = y_1$. See also CAS Experiment 15. ■

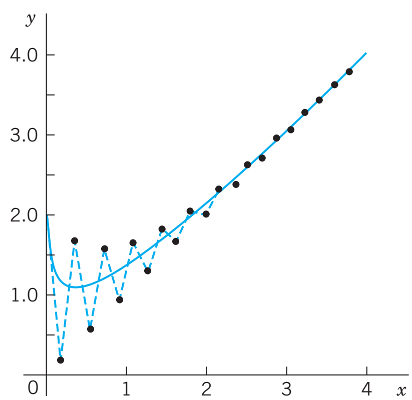


Fig. 452. Euler method with $h = 0.18$ in Example 4

PROBLEM SET 21.3

1–6 EULER FOR SYSTEMS AND SECOND-ORDER ODEs

Solve by the Euler's method. Graph the solution in the y_1y_2-plane. Calculate the errors.

1. $y_1' = 2y_1 - 4y_2, \quad y_2' = y_1 - 3y_2, \quad y_1(0) = 3,$ $y_2(0) = 0, \quad h = 0.1, \quad$ 10 steps
2. **Spiral.** $y_1' = -y_1 + y_2, \ y_2' = -y_1 - y_2, \ y_1(0) = 0,$ $y_2(0) = 4, \quad h = 0.2, \quad$ 5 steps
3. $y'' + \frac{1}{4}y = 0, \quad y(0) = 1, \quad y'(0) = 0, \quad h = 0.2,$ 5 steps
4. $y_1' = -3y_1 + y_2, \quad y_2' = y_1 - 3y_2, \quad y_1(0) = 2,$ $y_2(0) = 0, \quad h = 0.1, \quad$ 5 steps
5. $y'' - y = x, \quad y(0) = 1, \quad y'(0) = -2, \quad h = 0.1,$ 5 steps
6. $y_1' = y_1, \quad y_2' = -y_2, \quad y_1(0) = 2, \quad y_2(0) = 2,$ $h = 0.1, \quad$ 10 steps

7–10 RK FOR SYSTEMS

Solve by the classical RK.

7. The ODE in Prob. 5. By what factor did the error decrease?
8. The system in Prob. 2
9. The system in Prob. 1
10. The system in Prob. 4
11. **Pendulum equation** $y'' + \sin y = 0, \quad y(\pi) = 0,$ $y'(\pi) = 1,$ as a system, $h = 0.2$, 20 steps. How does your result fit into Fig. 93 in Sec. 4.5?
12. **Bessel Function J_0.** $xy'' + y' + xy = 0, \quad y(1) = 0.765198, \quad y'(1) = -0.440051, \quad h = 0.5,$ 5 steps. (This gives the standard solution $J_0(x)$ in Fig. 110 in Sec. 5.4.)
13. Verify the formulas and calculations for the Airy equation in Example 2 of the text.
14. **RKN.** The classical RK for a first-order ODE extends to second-order ODEs (E. J. Nyström, *Acta fenn.* No 13, 1925). If the ODE is $y'' = f(x, y)$, not containing y', then

$$k_1 = \tfrac{1}{2}hf(x_n, y_n)$$
$$k_2 = \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h,\ y_n + \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1)) = k_3$$
$$k_4 = \tfrac{1}{2}hf(x_n + h,\ y_n + h(y_n' + k_2))$$
$$y_{n+1} = y_n + h(y_n' + \tfrac{1}{3}(k_1 + 2k_2))$$
$$y_{n+1}' = y_n' + \tfrac{1}{8}(k_1 + 4k_2 + k_4).$$

Apply this RKN (Runge–Kutta–Nyström) method to the Airy ODE in Example 2 with $h = 0.2$ as before, to obtain approximate values of $\mathrm{Ai}(x)$.

15. **CAS EXPERIMENT. Backward Euler and Stiffness.** Extend Example 3 as follows.

(a) Verify the values in Table 21.13 and show them graphically as in Fig. 452.

(b) Compute and graph Euler values for h near the "critical" $h = 0.18$ to determine more exactly when instability starts.

(c) Compute and graph RK values for values of h between 0.2 and 0.3 to find h for which the RK approximation begins to increase away from the exact solution.

(d) Compute and graph backward Euler values for large h; confirm stability and investigate the error increase for growing h.

21.4 Methods for Elliptic PDEs

We have arrived at the second half of this chapter, which is devoted to numerics for partial differential equations (PDEs). As we have seen in Chap.12, there are many applications to PDEs, such as in dynamics, elasticity, heat transfer, electromagnetic theory, quantum mechanics, and others. Selected because of their importance in applications, the PDEs covered here include the Laplace equation, the Poisson equation, the heat equation, and the wave equation. By covering these equations based on their importance in applications we also selected equations that are important for theoretical considerations. Indeed, these equations serve as models for elliptic, parabolic, and hyperbolic PDEs. For example, the Laplace equation is a representative example of an elliptic type of PDE, and so forth.

Recall, from Sec. 12.4, that a PDE is called **quasilinear** if it is linear in the highest derivatives. Hence a second-order quasilinear PDE in two independent variables x, y is of the form

$$(1) \qquad au_{xx} + 2bu_{xy} + cu_{yy} = F(x, y, u, u_x, u_y).$$

u is an unknown function of x and y (a solution sought). F is a given function of the indicated variables.

Depending on the discriminant $ac - b^2$, the PDE (1) is said to be of

elliptic type	if	$ac - b^2 > 0$	(example: *Laplace equation*)
parabolic type	if	$ac - b^2 = 0$	(example: *heat equation*)
hyperbolic type	if	$ac - b^2 < 0$	(example: *wave equation*).

Here, in the heat and wave equations, y is time t. The *coefficients* a, b, c may be functions of x, y, so that the type of (1) may be different in different regions of the xy-plane. This classification is not merely a formal matter but is of great practical importance because the general behavior of solutions differs from type to type and so do the additional conditions (boundary and initial conditions) that must be taken into account.

Applications involving ***elliptic equations*** usually lead to boundary value problems in a region R, called a *first boundary value problem* or **Dirichlet problem** if u is prescribed on the boundary curve C of R, a *second boundary value problem* or **Neumann problem** if $u_n = \partial u/\partial n$ (normal derivative of u) is prescribed on C, and a *third* or **mixed problem** if u is prescribed on a part of C and u_n on the remaining part. C usually is a closed curve (or sometimes consists of two or more such curves).

Difference Equations for the Laplace and Poisson Equations

In this section we develop numeric methods for the two most important elliptic PDEs that appear in applications. The two PDEs are the **Laplace equation**

$$\textbf{(2)} \qquad \nabla^2 u = u_{xx} + u_{yy} = 0$$

and the **Poisson equation**

$$\textbf{(3)} \qquad \nabla^2 u = u_{xx} + u_{yy} = f(x, y).$$

The starting point for developing our numeric methods is the idea that we can replace the partial derivatives of these PDEs by corresponding **difference quotients**. Details are as follows:

To develop this idea, we start with the Taylor formula and obtain

$$(4) \qquad \begin{aligned} &\text{(a)} \quad u(x + h, y) = u(x, y) + hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) + \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots \\ &\text{(b)} \quad u(x - h, y) = u(x, y) - hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) - \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots . \end{aligned}$$

We subtract (4b) from (4a), neglect terms in $h^3, h^4, \cdots$, and solve for u_x. Then

$$u_x(x, y) \approx \frac{1}{2h}\left[u(x + h, y) - u(x - h, y)\right]. \tag{5a}$$

Similarly,

$$u(x, y + k) = u(x, y) + ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots$$

and

$$u(x, y - k) = u(x, y) - ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots.$$

By subtracting, neglecting terms in $k^3, k^4, \cdots$, and solving for u_y we obtain

$$u_y(x, y) \approx \frac{1}{2k}\left[u(x, y + k) - u(x, y - k)\right]. \tag{5b}$$

We now turn to second derivatives. Adding (4a) and (4b) and neglecting terms in $h^4, h^5, \cdots$, we obtain $u(x + h, y) + u(x - h, y) \approx 2u(x, y) + h^2 u_{xx}(x, y)$. Solving for u_{xx} we have

$$u_{xx}(x, y) \approx \frac{1}{h^2}\left[u(x + h, y) - 2u(x, y) + u(x - h, y)\right]. \tag{6a}$$

Similarly,

$$u_{yy}(x, y) \approx \frac{1}{k^2}\left[u(x, y + k) - 2u(x, y) + u(x, y - k)\right]. \tag{6b}$$

We shall not need (see Prob. 1)

$$\begin{aligned} u_{xy}(x, y) \approx \frac{1}{4hk}\big[&u(x + h, y + k) - u(x - h, y + k) \\ &- u(x + h, y - k) + u(x - h, y - k)\big]. \end{aligned} \tag{6c}$$

Figure 453a shows the points $(x + h, y), (x - h, y), \cdots$ in (5) and (6).

We now substitute (6a) and (6b) into the ***Poisson equation*** (3), choosing $k = h$ to obtain a simple formula:

$$u(x + h, y) + u(x, y + h) + u(x - h, y) + u(x, y - h) - 4u(x, y) = h^2 f(x, y). \tag{7}$$

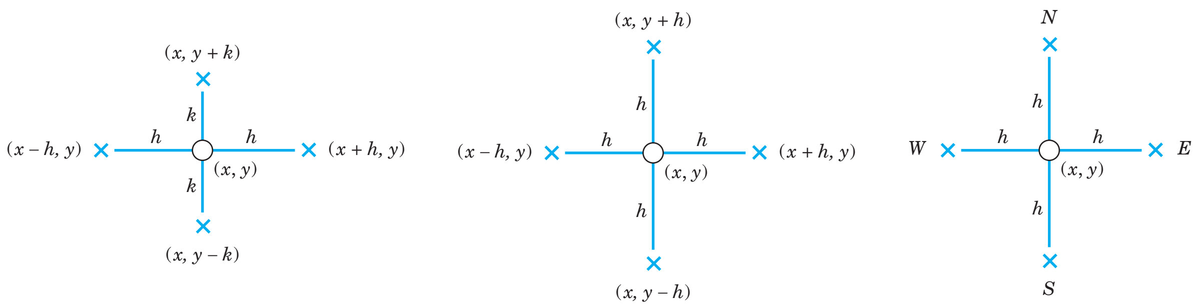
This is a **difference equation** corresponding to (3). Hence for the ***Laplace equation*** (2) the corresponding difference equation is

$$u(x + h, y) + u(x, y + h) + u(x - h, y) + u(x, y - h) - 4u(x, y) = 0. \tag{8}$$

h is called the **mesh size**. Equation (8) relates u at (x, y) to u at the four neighboring points shown in Fig. 453b. It has a remarkable interpretation: u at (x, y) equals the mean of the values of u at the four neighboring points. This is an analog of the mean value property of harmonic functions (Sec. 18.6).

Those neighbors are often called E (East), N (North), W (West), S (South). Then Fig. 453b becomes Fig. 453c and (7) is

$$u(E) + u(N) + u(W) + u(S) - 4u(x, y) = h^2 f(x, y). \tag{7*}$$



(a) Points in (5) and (6)   (b) Points in (7) and (8)   (c) Notation in (7*)

Fig. 453. Points and notation in (5)–(8) and (7*)

Our approximation of $h^2\nabla^2 u$ in (7) and (8) is a 5-point approximation with the coefficient scheme or **stencil** (also called *pattern, molecule,* or *star*)

$$\text{(9)}\quad \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix}. \quad \text{We may now write (7) as} \quad \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix} u = h^2 f(x, y).$$

Dirichlet Problem

In numerics for the Dirichlet problem in a region R we choose an h and introduce a square grid of horizontal and vertical straight lines of distance h. Their intersections are called **mesh points** (or *lattice points* or *nodes*). See Fig. 454.

Then we approximate the given PDE by a difference equation [(8) for the Laplace equation], which relates the unknown values of u at the mesh points in R to each other and to the given boundary values (details in Example 1). This gives a linear system of *algebraic* equations. By solving it we get approximations of the unknown values of u at the mesh points in R.

We shall see that the number of equations equals the number of unknowns. Now comes an important point. If the number of internal mesh points, call it p, is small, say, $p < 100$, then a direct solution method may be applied to that linear system of $p < 100$ equations in p unknowns. However, if p is large, a storage problem will arise. Now since each unknown u is related to only 4 of its neighbors, the coefficient matrix of the system is a **sparse matrix**, that is, a matrix with relatively few nonzero entries (for instance, 500 of 10,000 when $p = 100$). Hence for large p we may avoid storage difficulties by using an iteration method, notably the Gauss–Seidel method (Sec. 20.3), which in PDEs is also called **Liebmann's method** (note the strict diagonal dominance). Remember that in this method we have the storage convenience that we can overwrite any solution component (value of u) as soon as a "new" value is available.

Both cases, large p and small p, are of interest to the engineer, large p if a fine grid is used to achieve high accuracy, and small p if the boundary values are known only rather inaccurately, so that a coarse grid will do it because in this case it would be meaningless to try for great accuracy in the interior of the region R.

We illustrate this approach with an example, keeping the number of equations small, for simplicity. As convenient ***notations for mesh points and corresponding values of the solution*** (and of approximate solutions) we use (see also Fig. 454)

$$P_{ij} = (ih, jh), \qquad u_{ij} = u(ih, jh). \tag{10}$$

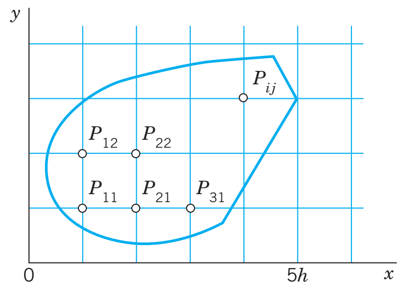


Fig. 454. Region in the xy-plane covered by a grid of mesh h, also showing mesh points $P_{11} = (h, h), \cdots, P_{ij} = (ih, jh), \cdots$

With this notation we can write (8) for any mesh point P_{ij} in the form

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0. \tag{11}$$

Remark. Our current discussion and the example that follows illustrate what we may call the ***reuseability of mathematical ideas and methods***. Recall that we applied the Gauss–Seidel method to a system of ODEs in Sec. 20.3 and that we can now apply it again to elliptic PDEs. This shows that engineering mathematics has a structure and important mathematical ideas and methods will appear again and again in different situations. The student should find this attractive in that previous knowledge can be reapplied.

EXAMPLE 1 Laplace Equation. Liebmann's Method

The four sides of a square plate of side 12 cm, made of homogeneous material, are kept at constant temperature 0°C and 100°C as shown in Fig. 455a. Using a (very wide) grid of mesh 4 cm and applying Liebmann's method (that is, Gauss–Seidel iteration), find the (steady-state) temperature at the mesh points.

Solution. In the case of independence of time, the heat equation (see Sec. 10.8)

$$u_t = c^2(u_{xx} + u_{yy})$$

reduces to the Laplace equation. Hence our problem is a Dirichlet problem for the latter. We choose the grid shown in Fig. 455b and consider the mesh points in the order $P_{11}, P_{21}, P_{12}, P_{22}$. We use (11) and, in each equation, take to the right all the terms resulting from the given boundary values. Then we obtain the system

$$
(12) \qquad
\begin{aligned}
-4u_{11} + u_{21} + u_{12} \phantom{+ u_{22}} &= -200 \\
-u_{11} - 4u_{21} \phantom{+ u_{12}} + u_{22} &= -200 \\
u_{11} \phantom{+ u_{21}} - 4u_{12} + u_{22} &= -100 \\
u_{21} + u_{12} - 4u_{22} &= -100.
\end{aligned}
$$

In practice, one would solve such a small system by the Gauss elimination, finding $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$.

More exact values (exact to 3S) of the solution of the actual problem [as opposed to its model (12)] are 88.1 and 61.9, respectively. (These were obtained by using Fourier series.) Hence the error is about 1%, which is surprisingly accurate for a grid of such a large mesh size h. If the system of equations were large, one would solve it by an indirect method, such as Liebmann's method. For (12) this is as follows. We write (12) in the form (divide by -4 and take terms to the right)

$$
\begin{aligned}
u_{11} &= \phantom{0.25u_{11} +} 0.25u_{21} + 0.25u_{12} \phantom{+ 0.25u_{22}} + 50 \\
u_{21} &= 0.25u_{11} \phantom{+ 0.25u_{21} + 0.25u_{12}} + 0.25u_{22} + 50 \\
u_{12} &= 0.25u_{11} \phantom{+ 0.25u_{21} + 0.25u_{12}} + 0.25u_{22} + 25 \\
u_{22} &= \phantom{0.25u_{11} +} 0.25u_{21} + 0.25u_{12} \phantom{+ 0.25u_{22}} + 25.
\end{aligned}
$$

These equations are now used for the Gauss–Seidel iteration. They are identical with (2) in Sec. 20.3, where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$, and the iteration is explained there, with 100, 100, 100, 100 chosen as starting values. Some work can be saved by better starting values, usually by taking the average of the boundary values that enter into the linear system. The exact solution of the system is $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$, as you may verify.

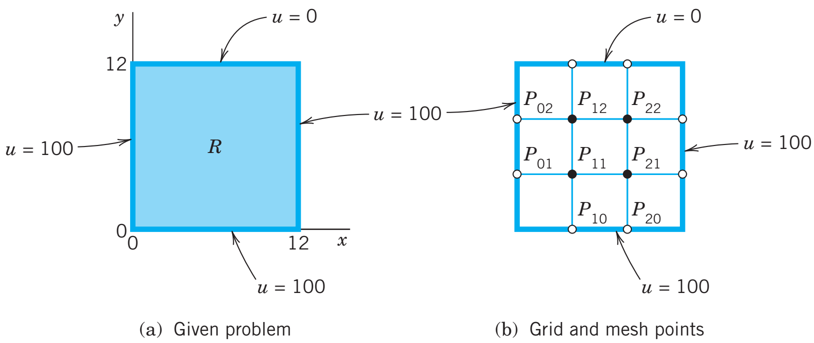


Fig. 455. Example 1

Remark. It is interesting to note that, if we choose mesh $h = L/n$ (L = side of R) and consider the $(n-1)^2$ internal mesh points (i.e., mesh points not on the boundary) row by row in the order

$$P_{11}, P_{21}, \cdots, P_{n-1,1}, P_{12}, P_{22}, \cdots, P_{n-2,2}, \cdots,$$

then the system of equations has the $(n-1)^2 \times (n-1)^2$ coefficient matrix

$$
(13) \qquad \mathbf{A} = \begin{bmatrix}
\mathbf{B} & \mathbf{I} & & & & \\
\mathbf{I} & \mathbf{B} & \mathbf{I} & & & \\
& & \cdot & & & \\
& & & \cdot & & \\
& & & & \cdot & \\
& & & \mathbf{I} & \mathbf{B} & \mathbf{I} \\
& & & & \mathbf{I} & \mathbf{B}
\end{bmatrix}. \qquad \text{Here} \qquad \mathbf{B} = \begin{bmatrix}
-4 & 1 & & & & \\
1 & -4 & 1 & & & \\
& & \cdot & & & \\
& & & \cdot & & \\
& & & & \cdot & \\
& & & 1 & -4 & 1 \\
& & & & 1 & -4
\end{bmatrix}
$$

is an $(n-1) \times (n-1)$ matrix. (In (12) we have $n = 3$, $(n-1)^2 = 4$ internal mesh points, two submatrices **B**, and two submatrices **I**.) The matrix **A** is nonsingular. This follows by noting that the off-diagonal entries in each row of **A** have the sum 3 (or 2), whereas each diagonal entry of **A** equals -4, so that nonsingularity is implied by Gerschgorin's theorem in Sec. 20.7 because no Gerschgorin disk can include 0. ■

A matrix is called a **band matrix** if it has all its nonzero entries on the main diagonal and on sloping lines parallel to it (separated by sloping lines of zeros or not). For example, **A** in (13) is a band matrix. Although the Gauss elimination does not preserve zeros between bands, it does not introduce nonzero entries outside the limits defined by the original bands. Hence a band structure is advantageous. In (13) it has been achieved by carefully ordering the mesh points.

ADI Method

A matrix is called a **tridiagonal matrix** if it has all its nonzero entries on the main diagonal and on the two sloping parallels immediately above or below the diagonal. (See also Sec. 20.9.) In this case the Gauss elimination is particularly simple.

This raises the question of whether, in the solution of the Dirichlet problem for the Laplace or Poisson equations, one could obtain a system of equations whose coefficient matrix is tridiagonal. The answer is yes, and a popular method of that kind, called the **ADI method** (*alternating direction implicit method*) was developed by Peaceman and Rachford. The idea is as follows. The stencil in (9) shows that we could obtain a tridiagonal matrix if there were only the three points in a row (or only the three points in a column). This suggests that we write (11) in the form

$$u_{i-1,j} - 4u_{ij} + u_{i+1,j} = -u_{i,j-1} - u_{i,j+1} \tag{14a}$$

so that the left side belongs to y-Row j only and the right side to x-Column i. Of course, we can also write (11) in the form

$$u_{i,j-1} - 4u_{ij} + u_{i,j+1} = -u_{i-1,j} - u_{i+1,j} \tag{14b}$$

so that the left side belongs to Column i and the right side to Row j. In the ADI method we proceed by iteration. At every mesh point we choose an arbitrary starting value $u_{ij}^{(0)}$. In each step we compute new values at all mesh points. In one step we use an iteration formula resulting from (14a) and in the next step an iteration formula resulting from (14b), and so on in alternating order.

In detail: suppose approximations $u_{ij}^{(m)}$ have been computed. Then, to obtain the next approximations $u_{ij}^{(m+1)}$, we substitute the $u_{ij}^{(m)}$ ***on the right*** side of (14a) and solve for the $u_{ij}^{(m+1)}$ on the left side; that is, we use

$$u_{i-1,j}^{(m+1)} - 4u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} - u_{i,j+1}^{(m)}. \tag{15a}$$

We use (15a) for a fixed j, that is, ***for a fixed row*** j, and for all internal mesh points in this row. This gives a linear system of N algebraic equations (N = number of internal mesh points per row) in N unknowns, the new approximations of u at these mesh points. Note that (15a) involves not only approximations computed in the previous step but also given boundary values. We solve the system (15a) (j fixed!) by Gauss elimination. Then we go to the next row, obtain another system of N equations and solve it by Gauss, and so on, until all rows are done. In the next step we ***alternate direction***, that is, we compute the next approximations $u_{ij}^{(m+2)}$ column by column from the $u_{ij}^{(m+1)}$ and the given boundary values, using a formula obtained from (14b) by substituting the $u_{ij}^{(m+1)}$ ***on the right***:

$$u_{i,j-1}^{(m+2)} - 4u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} - u_{i+1,j}^{(m+1)}. \tag{15b}$$

For each fixed i, that is, ***for each column,*** this is a system of M equations (M = number of internal mesh points per column) in M unknowns, which we solve by Gauss elimination. Then we go to the next column, and so on, until all columns are done.

Let us consider an example that merely serves to explain the entire method.

EXAMPLE 2 Dirichlet Problem. ADI Method

Explain the procedure and formulas of the ADI method in terms of the problem in Example 1, using the same grid and starting values 100, 100, 100, 100.

Solution. While working, we keep an eye on Fig. 455b and the given boundary values. We obtain first approximations $u_{11}^{(1)}, u_{21}^{(1)}, u_{12}^{(1)}, u_{22}^{(1)}$ from (15a) with $m = 0$. We write boundary values contained in (15a) without an upper index, for better identification and to indicate that these given values remain the same during the iteration. From (15a) with $m = 0$ we have for $j = 1$ (first row) the system

$$\begin{aligned}
&(i = 1) \quad u_{01} - 4u_{11}^{(1)} + u_{21}^{(1)} &&= -u_{10} - u_{12}^{(0)} \\
&(i = 2) \quad u_{11}^{(1)} - 4u_{21}^{(1)} + u_{31} &&= -u_{20} - u_{22}^{(0)}.
\end{aligned}$$

The solution is $u_{11}^{(1)} = u_{21}^{(1)} = 100$. For $j = 2$ (second row) we obtain from (15a) the system

$$\begin{aligned}
&(i = 1) \quad u_{02} - 4u_{12}^{(1)} + u_{22}^{(1)} &&= -u_{11}^{(0)} - u_{13} \\
&(i = 2) \quad u_{12}^{(1)} - 4u_{22}^{(1)} + u_{32} &&= -u_{21}^{(0)} - u_{23}.
\end{aligned}$$

The solution is $u_{12}^{(1)} = u_{22}^{(1)} = 66.667$.

Second approximations $u_{11}^{(2)}, u_{21}^{(2)}, u_{12}^{(2)}, u_{22}^{(2)}$ are now obtained from (15b) with $m = 1$ by using the first approximations just computed and the boundary values. For $i = 1$ (first column) we obtain from (15b) the system

$$\begin{aligned}
&(j = 1) \quad u_{10} - 4u_{11}^{(2)} + u_{12}^{(2)} &&= -u_{01} - u_{21}^{(1)} \\
&(j = 2) \quad u_{11}^{(2)} - 4u_{12}^{(2)} + u_{13} &&= -u_{02} - u_{22}^{(1)}.
\end{aligned}$$

The solution is $u_{11}^{(2)} = 91.11$, $u_{12}^{(2)} = 64.44$, For $i = 2$ (second column) we obtain from (15b) the system

$$\begin{aligned}
&(j = 1) \quad u_{20} - 4u_{21}^{(2)} + u_{22}^{(2)} &&= -u_{11}^{(1)} - u_{31} \\
&(j = 2) \quad u_{21}^{(2)} - 4u_{22}^{(2)} + u_{23} &&= -u_{12}^{(1)} - u_{32}.
\end{aligned}$$

The solution is $u_{21}^{(2)} = 91.11$, $u_{22}^{(2)} = 64.44$.

In this example, which merely serves to explain the practical procedure in the ADI method, the accuracy of the second approximations is about the same as that of two Gauss–Seidel steps in Sec. 20.3 (where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$), as the following table shows.

Method	u_{11}	u_{21}	u_{12}	u_{22}
ADI, 2nd approximations	91.11	91.11	64.44	64.44
Gauss–Seidel, 2nd approximations	93.75	90.62	65.62	64.06
Exact solution of (12)	87.50	87.50	62.50	62.50

■

Improving Convergence. Additional improvement of the convergence of the ADI method results from the following interesting idea. Introducing a parameter p, we can also write (11) in the form

$$
(16) \quad \begin{aligned}
&\text{(a)} \quad u_{i-1,j} - (2+p)u_{ij} + u_{i+1,j} = -u_{i,j-1} + (2-p)u_{ij} - u_{i,j+1} \\
&\text{(b)} \quad u_{i,j-1} - (2+p)u_{ij} + u_{i,j+1} = -u_{i-1,j} + (2-p)u_{ij} - u_{i+1,j}.
\end{aligned}
$$

This gives the more general ADI iteration formulas

$$
(17) \quad \begin{aligned}
&\text{(a)} \quad u_{i-1,j}^{(m+1)} - (2+p)u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} + (2-p)u_{ij}^{(m)} - u_{i,j+1}^{(m)} \\
&\text{(b)} \quad u_{i,j-1}^{(m+2)} - (2+p)u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} + (2-p)u_{ij}^{(m+1)} - u_{i+1,j}^{(m+1)}.
\end{aligned}
$$

For $p = 2$, this is (15). The parameter p may be used for improving convergence. Indeed, one can show that the ADI method converges for positive p, and that the optimum value for maximum rate of convergence is

$$
(18) \qquad p_0 = 2 \sin \frac{\pi}{K}
$$

where K is the larger of $M + 1$ and $N + 1$ (see above). Even better results can be achieved by letting p vary from step to step. More details of the ADI method and variants are discussed in Ref. [E25] listed in App. 1.

PROBLEM SET 21.4

1. Derive (5b), (6b), and (6c).
2. Verify the calculations in Example 1 of the text. Find out experimentally how many steps you need to obtain the solution of the linear system with an accuracy of 3S.
3. **Use of symmetry.** Conclude from the boundary values in Example 1 that $u_{21} = u_{11}$ and $u_{22} = u_{12}$. Show that this leads to a system of two equations and solve it.
4. **Finer grid** of 3×3 inner points. Solve Example 1, choosing $h = \frac{12}{4} = 3$ (instead of $h = \frac{12}{3} = 4$) and the same starting values.

5–10 GAUSS ELIMINATION, GAUSS–SEIDEL ITERATION

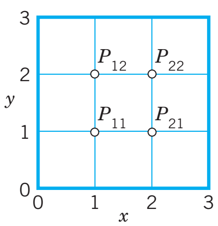


Fig. 456. Problems 5–10

For the grid in Fig. 456 compute the potential at the four internal points by Gauss and by 5 Gauss–Seidel steps with starting values 100, 100, 100, 100 (showing the details of your work) if the boundary values on the edges are:

5. $u(1, 0) = 60$, $u(2, 0) = 300$, $u = 100$ on the other three edges.
6. $u = 0$ on the left, x^3 on the lower edge, $27 - 9y^2$ on the right, $x^3 - 27x$ on the upper edge.
7. U_0 on the upper and lower edges, $-U_0$ on the left and right. Sketch the equipotential lines.
8. $u = 220$ on the upper and lower edges, 110 on the left and right.
9. $u = \sin \frac{1}{3}\pi x$ on the upper edge, 0 on the other edges, 10 steps.
10. $u = x^4$ on the lower edge, $81 - 54y^2 + y^4$ on the right, $x^4 - 54x^2 + 81$ on the upper edge, y^4 on the left. Verify the exact solution $x^4 - 6x^2y^2 + y^4$ and determine the error.

11. Find the potential in Fig. 457 using **(a)** the coarse grid, **(b)** the fine grid 5×3, and Gauss elimination. *Hint.* In (b), use symmetry; take $u = 0$ as boundary value at the two points at which the potential has a jump.

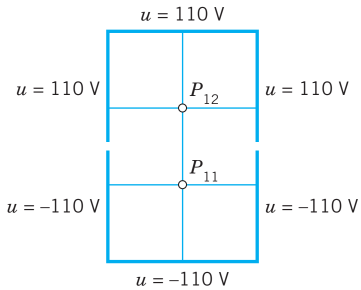


Fig. 457. Region and grids in Problem 11

12. Influence of starting values. Do Prob. 9 by Gauss–Seidel, starting from **0**. Compare and comment.

13. For the square $0 \leqq x \leqq 4$, $0 \leqq y \leqq 4$ let the boundary temperatures be 0°C on the horizontal and 50°C on the vertical edges. Find the temperatures at the interior points of a square grid with $h = 1$.

14. Using the answer to Prob. 13, try to sketch some isotherms.

15. Find the isotherms for the square and grid in Prob. 13 if $u = \sin \frac{1}{4}\pi x$ on the horizontal and $-\sin \frac{1}{4}\pi y$ on the vertical edges. Try to sketch some isotherms.

16. ADI. Apply the ADI method to the Dirichlet problem in Prob. 9, using the grid in Fig. 456, as before and starting values zero.

17. What p_0 in (18) should we choose for Prob. 16? Apply the ADI formulas (17) with that value of p_0 to Prob. 16, performing 1 step. Illustrate the improved convergence by comparing with the corresponding values 0.077, 0.308 after the first step in Prob. 16. (Use the starting values zero.)

18. CAS PROJECT. Laplace Equation. (a) Write a program for Gauss–Seidel with 16 equations in 16 unknowns, composing the matrix (13) from the indicated 4×4 submatrices and including a transformation of the vector of the boundary values into the vector $\mathbf{b}$ of $\mathbf{Ax} = \mathbf{b}$.

(b) Apply the program to the square grid in $0 \leqq x \leqq 5$, $0 \leqq y \leqq 5$ with $h = 1$ and $u = 220$ on the upper and lower edges, $u = 110$ on the left edge and $u = -10$ on the right edge. Solve the linear system also by Gauss elimination. What accuracy is reached in the 20th Gauss–Seidel step?

21.5 Neumann and Mixed Problems. Irregular Boundary

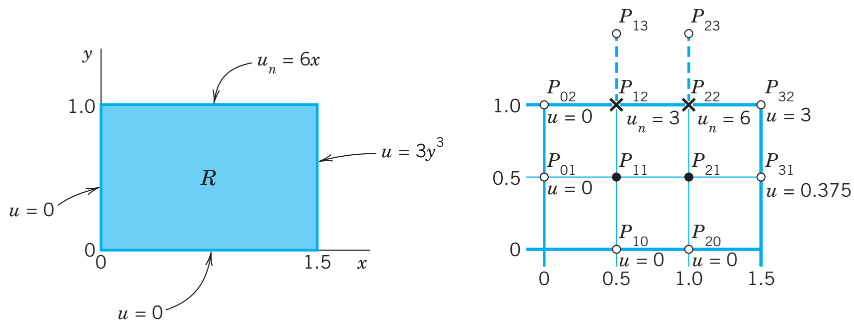
We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. The Dirichlet problem was studied in the last section. In solving **Neumann** and **mixed problems** (defined in the last section) we are confronted with a new situation, because there are boundary points at which the (outer) **normal derivative** $u_n = \partial u/\partial n$ of the solution is given, but u itself is unknown since it is not given. To handle such points we need a new idea. This idea is the same for Neumann and mixed problems. Hence we may explain it in connection with one of these two types of problems. We shall do so and consider a typical example as follows.

EXAMPLE 1 Mixed Boundary Value Problem for a Poisson Equation

Solve the mixed boundary value problem for the Poisson equation

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y) = 12xy$$

shown in Fig. 458a.



(a) Region R and boundary values (b) Grid ($h = 0.5$)

Fig. 458. Mixed boundary value problem in Example 1

Solution. We use the grid shown in Fig. 458b, where $h = 0.5$. We recall that (7) in Sec. 21.4 has the right side $h^2 f(x, y) = 0.5^2 \cdot 12xy = 3xy$. From the formulas $u = 3y^3$ and $u_n = 6x$ given on the boundary we compute the boundary data

$$(1) \qquad u_{31} = 0.375, \qquad u_{32} = 3, \qquad \frac{\partial u_{12}}{\partial n} = \frac{\partial u_{12}}{\partial y} = 6 \cdot 0.5 = 3. \qquad \frac{\partial u_{22}}{\partial n} = \frac{\partial u_{22}}{\partial y} = 6 \cdot 1 = 6.$$

P_{11} and P_{21} are internal mesh points and can be handled as in the last section. Indeed, from (7), Sec. 21.4, with $h^2 = 0.25$ and $h^2 f(x, y) = 3xy$ and from the given boundary values we obtain two equations corresponding to P_{11} and P_{21}, as follows (with -0 resulting from the left boundary).

$$\text{(2a)} \qquad \begin{aligned} -4u_{11} + u_{21} + u_{12} \qquad\qquad &= 12(0.5 \cdot 0.5) \cdot \tfrac{1}{4} - 0 = 0.75 \\ u_{11} - 4u_{21} \qquad + u_{22} &= 12(1 \cdot 0.5) \cdot \tfrac{1}{4} - 0.375 = 1.125. \end{aligned}$$

The only difficulty with these equations seems to be that they involve the unknown values u_{12} and u_{22} of u at P_{12} and P_{22} on the boundary, where the normal derivative $u_n = \partial u/\partial n = \partial u/\partial y$ is given, instead of u; but we shall overcome this difficulty as follows.

We consider P_{12} and P_{22}. The idea that will help us here is this. We imagine the region R to be extended above to the first row of external mesh points (corresponding to $y = 1.5$), and we assume that the Poisson equation also holds in the extended region. Then we can write down two more equations as before (Fig. 458b)

$$\text{(2b)} \qquad \begin{aligned} u_{11} \quad - 4u_{12} + u_{22} + u_{13} \qquad &= 1.5 - 0 = 1.5 \\ u_{21} + u_{12} - 4u_{22} \qquad + u_{23} &= 3 - 3 = 0. \end{aligned}$$

On the right, 1.5 is $12xyh^2$ at (0.5, 1) and 3 is $12xyh^2$ at (1, 1) and 0 (at P_{02}) and 3 (at P_{32}) are given boundary values. We remember that we have not yet used the boundary condition on the upper part of the boundary of R, and we also notice that in (2b) we have introduced two more unknowns u_{13}, u_{23}. But we can now use that condition and get rid of u_{13}, u_{23} by applying the central difference formula for du/dy. From (1) we then obtain (see Fig. 458b)

$$3 = \frac{\partial u_{12}}{\partial y} \approx \frac{u_{13} - u_{11}}{2h} = u_{13} - u_{11}, \qquad \text{hence} \qquad u_{13} = u_{11} + 3$$

$$6 = \frac{\partial u_{22}}{\partial y} \approx \frac{u_{23} - u_{21}}{2h} = u_{23} - u_{21}, \qquad \text{hence} \qquad u_{23} = u_{21} + 6.$$

Substituting these results into (2b) and simplifying, we have

$$\begin{aligned} 2u_{11} \quad - 4u_{12} + u_{22} &= 1.5 - 3 = -1.5 \\ 2u_{21} + u_{12} - 4u_{22} &= 3 - 3 - 6 = -6. \end{aligned}$$

Together with (2a) this yields, written in matrix form,

$$(3)\qquad \begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 2 & 0 & -4 & 1 \\ 0 & 2 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ 1.5 - 3 \\ 0 - 6 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ -1.5 \\ -6 \end{bmatrix}.$$

(The entries 2 come from u_{13} and u_{23}, and so do -3 and -6 on the right). The solution of (3) (obtained by Gauss elimination) is as follows; the exact values of the problem are given in parentheses.

$$u_{12} = 0.866 \quad \text{(exact 1)} \qquad u_{22} = 1.812 \quad \text{(exact 2)}$$
$$u_{11} = 0.077 \quad \text{(exact 0.125)} \qquad u_{21} = 0.191 \quad \text{(exact 0.25)}.$$

Irregular Boundary

We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. If R has a simple geometric shape, we can usually arrange for certain mesh points to lie on the boundary C of R, and then we can approximate partial derivatives as explained in the last section. However, if C intersects the grid at points that are not mesh points, then at points close to the boundary we must proceed differently, as follows.

The mesh point O in Fig. 459 is of that kind. For O and its neighbors A and P we obtain from Taylor's theorem

$$(4)\qquad \begin{aligned} &\text{(a)} \quad u_A = u_O + ah\frac{\partial u_O}{\partial x} + \frac{1}{2}(ah)^2\frac{\partial^2 u_O}{\partial x^2} + \cdots \\ &\text{(b)} \quad u_P = u_O - h\frac{\partial u_O}{\partial x} + \frac{1}{2}h^2\frac{\partial^2 u_O}{\partial x^2} + \cdots . \end{aligned}$$

We disregard the terms marked by dots and eliminate $\partial u_O/\partial x$. Equation (4b) times a plus equation (4a) gives

$$u_A + au_P \approx (1 + a)\,u_O + \frac{1}{2}a\,(a + 1)\,h^2\frac{\partial^2 u_O}{\partial x^2}.$$

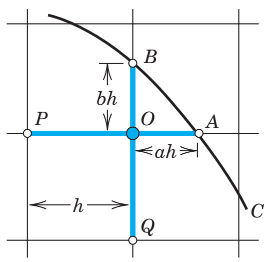


Fig. 459. Curved boundary C of a region R, a mesh point O near C, and neighbors A, B, P, Q

We solve this last equation algebraically for the derivative, obtaining

$$\frac{\partial^2 u_O}{\partial x^2} \approx \frac{2}{h^2}\left[\frac{1}{a(1 + a)}\,u_A + \frac{1}{1 + a}\,u_P - \frac{1}{a}\,u_O\right].$$

Similarly, by considering the points O, B, and Q,

$$\frac{\partial^2 u_O}{\partial y^2} \approx \frac{2}{h^2}\left[\frac{1}{b(1+b)}u_B + \frac{1}{1+b}u_Q - \frac{1}{b}u_O\right].$$

By addition,

$$\text{(5)} \qquad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(1+a)} + \frac{u_B}{b(1+b)} + \frac{u_P}{1+a} + \frac{u_Q}{1+b} - \frac{(a+b)u_O}{ab}\right].$$

For example, if $a = \frac{1}{2}$, $b = \frac{1}{2}$, instead of the stencil (see Sec. 21.4)

$$\left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\} \qquad \text{we now have} \qquad \left\{\begin{matrix} & \frac{4}{3} & \\ \frac{2}{3} & -4 & \frac{4}{3} \\ & \frac{2}{3} & \end{matrix}\right\}.$$

because $1/[a(1+a)] = \frac{4}{3}$, etc. The sum of all five terms still being zero (which is useful for checking).

Using the same ideas, you may show that in the case of Fig. 460.

$$\textbf{(6)} \quad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(a+p)} + \frac{u_B}{b(b+q)} + \frac{u_P}{p(p+a)} + \frac{u_Q}{q(q+b)} - \frac{ap+bq}{abpq}u_O\right],$$

a formula that takes care of all conceivable cases.

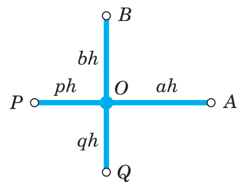


Fig. 460. Neighboring points A, B, P, Q of a mesh point O and notations in formula (6)

EXAMPLE 2 Dirichlet Problem for the Laplace Equation. Curved Boundary

Find the potential u in the region in Fig. 461 that has the boundary values given in that figure; here the curved portion of the boundary is an arc of the circle of radius 10 about (0,0). Use the grid in the figure.

Solution. u is a solution of the Laplace equation. From the given formulas for the boundary values $u = x^3$, $u = 512 - 24y^2, \cdots$ we compute the values at the points where we need them; the result is shown in the figure. For P_{11} and P_{12} we have the usual regular stencil, and for P_{21} and P_{22} we use (6), obtaining

$$\text{(7)} \qquad P_{11}, P_{12}\text{:} \left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\}, \quad P_{21}\text{:} \left\{\begin{matrix} & 0.5 & \\ 0.6 & -2.5 & 0.9 \\ & 0.5 & \end{matrix}\right\}, \quad P_{22}\text{:} \left\{\begin{matrix} & 0.9 & \\ 0.6 & -3 & 0.9 \\ & 0.6 & \end{matrix}\right\}.$$

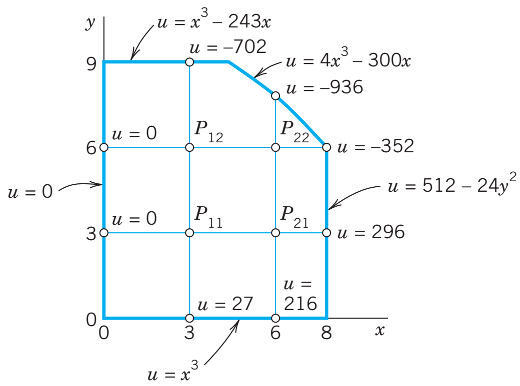


Fig. 461. Region, boundary values of the potential, and grid in Example 2

We use this and the boundary values and take the mesh points in the usual order $P_{11}, P_{21}, P_{12}, P_{22}$. Then we obtain the system

$$\begin{aligned}
-4u_{11} + u_{21} + u_{12} \phantom{+0.5u_{22}} &= 0 - 27 &&= -27\\
0.6u_{11} - 2.5u_{21} \phantom{+u_{12}} + 0.5u_{22} &= -0.9 \cdot 296 - 0.5 \cdot 216 &&= -374.4\\
u_{11} - 4u_{12} + u_{22} &= 702 + 0 &&= 702\\
0.6u_{21} + 0.6u_{12} - 3u_{22} &= 0.9 \cdot 352 + 0.9 \cdot 936 &&= 1159.2
\end{aligned}$$

In matrix form,

$$\begin{bmatrix} -4 & 1 & 1 & 0 \\ 0.6 & -2.5 & 0 & 0.5 \\ 1 & 0 & -4 & 1 \\ 0 & 0.6 & 0.6 & -3 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -27 \\ -374.4 \\ 702 \\ 1159.2 \end{bmatrix}. \tag{8}$$

Gauss elimination yields the (rounded) values

$$u_{11} = -55.6, \qquad u_{21} = 49.2, \qquad u_{12} = -298.5, \qquad u_{22} = -436.3.$$

Clearly, from a grid with so few mesh points we cannot expect great accuracy. The exact solution of the PDE (not of the difference equation) having the given boundary values is $u = x^3 - 3xy^2$ and yields the values

$$u_{11} = -54, \qquad u_{21} = 54, \qquad u_{12} = -297, \qquad u_{22} = -432.$$

In practice one would use a much finer grid and solve the resulting large system by an indirect method. ■

PROBLEM SET 21.5

1–7 MIXED BOUNDARY VALUE PROBLEMS

1. Check the values for the Poisson equation at the end of Example 1 by solving (3) by Gauss elimination.

2. Solve the mixed boundary value problem for the Poisson equation $\nabla^2 u = 2(x^2 + y^2)$ in the region and for the boundary conditions shown in Fig. 462, using the indicated grid.

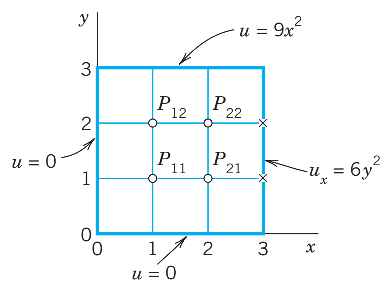


Fig. 462. Problems 2 and 6

3. CAS EXPERIMENT. Mixed Problem. Do Example 1 in the text with finer and finer grids of your choice and study the accuracy of the approximate values by comparing with the exact solution $u = 2xy^3$. Verify the latter.

4. Solve the mixed boundary value problem for the Laplace equation $\nabla^2 u = 0$ in the rectangle in Fig. 458a (using the grid in Fig. 458b) and the boundary conditions $u_x = 0$ on the left edge, $u_x = 3$ on the right edge, $u = x^2$ on the lower edge, and $u = x^2 - 1$ on the upper edge.

5. Do Example 1 in the text for the Laplace equation (instead of the Poisson equation) with grid and boundary data as before.

6. Solve $\nabla^2 u = -\pi^2 y \sin \frac{1}{3}\pi x$ for the grid in Fig. 462 and $u_y(1, 3) = u_y(2, 3) = \frac{1}{2}\sqrt{243}$, $u = 0$ on the other three sides of the square.

7. Solve Prob. 4 when $u_n = 110$ on the upper edge and $u = 110$ on the other edges.

8–16 IRREGULAR BOUNDARY

8. Verify the stencil shown after (5).

9. Derive (5) in the general case.

10. Derive the general formula (6) in detail.

11. Derive the linear system in Example 2 of the text.

12. Verify the solution in Example 2.

13. Solve the Laplace equation in the region and for the boundary values shown in Fig. 463, using the indicated grid. (The sloping portion of the boundary is $y = 4.5 - x$.)

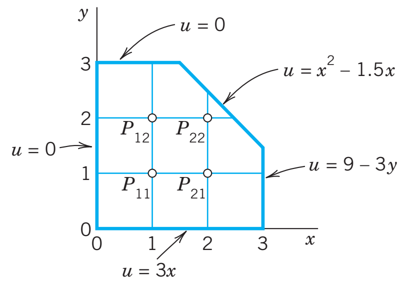


Fig. 463. Problem 13

14. If, in Prob. 13, the axes are grounded ($u = 0$), what constant potential must the other portion of the boundary have in order to produce 220 V at P_{11}?

15. What potential do we have in Prob. 13 if $u = 100$ V on the axes and $u = 0$ on the other portion of the boundary?

16. Solve the Poisson equation $\nabla^2 u = 2$ in the region and for the boundary values shown in Fig. 464, using the grid also shown in the figure.

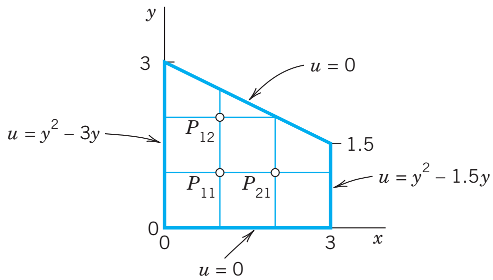


Fig. 464. Problem 16

21.6 Methods for Parabolic PDEs

The last two sections concerned elliptic PDEs, and we now turn to parabolic PDEs. Recall that the definitions of elliptic, parabolic, and hyperbolic PDEs were given in Sec. 21.4. There it was also mentioned that the general behavior of solutions differs from type to type, and so do the problems of practical interest. This reflects on numerics as follows.

For all three types, one replaces the PDE by a corresponding difference equation, but for *parabolic* and *hyperbolic* PDEs this does not automatically guarantee the **convergence** of the approximate solution to the exact solution as the mesh $h \to 0$; in fact, it does not even guarantee convergence at all. For these two types of PDEs one needs additional conditions (inequalities) to assure convergence and **stability**, the latter meaning that small perturbations in the initial data (or small errors at any time) cause only small changes at later times.

In this section we explain the numeric solution of the prototype of parabolic PDEs, the one-dimensional heat equation

$$u_t = c^2 u_{xx} \qquad (c \text{ constant}).$$

This PDE is usually considered for x in some fixed interval, say, $0 \leqq x \leqq L$, and time $t \geqq 0$, and one prescribes the initial temperature $u(x, 0) = f(x)$ (f given) and boundary conditions at $x = 0$ and $x = L$ for all $t \geqq 0$, for instance, $u(0, t) = 0$, $u(L, t) = 0$. We may assume $c = 1$ and $L = 1$; this can always be accomplished by a linear transformation of x and t (Prob. 1). Then the **heat equation** and those conditions are

(1) $$u_t = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

(2) $$u(x, 0) = f(x) \qquad \text{(Initial condition)}$$

(3) $$u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

A simple finite difference approximation of (1) is [see (6a) in Sec. 21.4; j is the number of the ***time step***]

(4) $$\frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}).$$

Figure 465 shows a corresponding grid and mesh points. The mesh size is h in the x-direction and k in the t-direction. Formula (4) involves the four points shown in Fig. 466. On the left in (4) we have used a *forward* difference quotient since we have no information for negative t at the start. From (4) we calculate $u_{i,j+1}$, which corresponds to time row $j + 1$, in terms of the three other u that correspond to time row j. Solving (4) for $u_{i,j+1}$, we have

(5) $$u_{i,j+1} = (1 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}), \qquad r = \frac{k}{h^2}.$$

Computations by this **explicit method** based on (5) are simple. However, it can be shown that crucial to the convergence of this method is the condition

(6) $$r = \frac{k}{h^2} \leqq \frac{1}{2}.$$

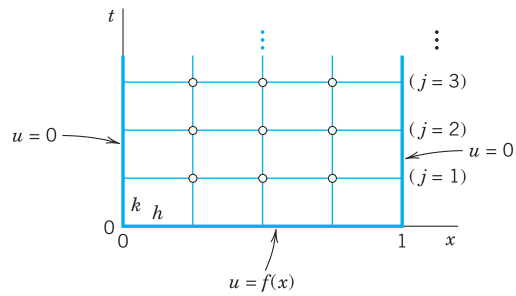


Fig. 465. Grid and mesh points corresponding to (4), (5)

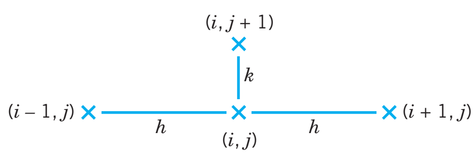


Fig. 466. The four points in (4) and (5)

That is, u_{ij} should have a positive coefficient in (5) or (for $r = \frac{1}{2}$) be absent from (5). Intuitively, (6) means that we should not move too fast in the t-direction. An example is given below.

Crank–Nicolson Method

Condition (6) is a handicap in practice. Indeed, to attain sufficient accuracy, we have to choose h small, which makes k very small by (6). For example, if $h = 0.1$, then $k \leqq 0.005$. Accordingly, we should look for a more satisfactory discretization of the heat equation.

A method that imposes no restriction on $r = k/h^2$ is the **Crank–Nicolson (CN) method**,[5] which uses values of u at the six points in Fig. 467. The idea of the method is the replacement of the difference quotient on the right side of (4) by $\frac{1}{2}$ times the sum of two such difference quotients at two time rows (see Fig. 467). Instead of (4) we then have

$$
\begin{aligned}
\frac{1}{k}(u_{i,j+1} - u_{ij}) &= \frac{1}{2h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \\
&\quad + \frac{1}{2h^2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}).
\end{aligned} \tag{7}
$$

Multiplying by $2k$ and writing $r = k/h^2$ as before, we collect the terms corresponding to time row $j + 1$ on the left and the terms corresponding to time row j on the right:

$$
(2 + 2r)u_{i,j+1} - r(u_{i+1,j+1} + u_{i-1,j+1}) = (2 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}). \tag{8}
$$

How do we use (8)? In general, the three values on the left are unknown, whereas the three values on the right are known. If we divide the x-interval $0 \leqq x \leqq 1$ in (1) into n equal intervals, we have $n - 1$ internal mesh points per time row (see Fig. 465, where $n = 4$). Then for $j = 0$ and $i = 1, \cdots, n - 1$, formula (8) gives a linear system of $n - 1$ equations for the $n - 1$ unknown values $u_{11}, u_{21}, \cdots, u_{n-1,1}$ in the first time row in terms of the initial values $u_{00}, u_{10}, \cdots, u_{n0}$ and the boundary values $u_{01} (= 0), u_{n1} (= 0)$. Similarly for $j = 1, j = 2$, and so on; that is, for each time row we have to solve such a linear system of $n - 1$ equations resulting from (8).

Although $r = k/h^2$ is no longer restricted, smaller r will still give better results. In practice, one chooses a k by which one can save a considerable amount of work, without

[5] JOHN CRANK (1916–2006), English mathematician and physicist at Courtaulds Fundamental Research Laboratory, professor at Brunel University, England. Student of Sir WILLIAM LAWRENCE BRAGG (1890–1971), Australian British physicist, who with his father, Sir WILLIAM HENRY BRAGG (1862–1942) won the Nobel Prize in physics in 1915 for their fundamental work in X-ray crystallography. (This is the only case where a father and a son shared the Nobel Prize for the same research. Furthermore, W. L. Bragg is the youngest Nobel laureate ever.) PHYLLIS NICOLSON (1917–1968), English mathematician, professor at the University of Leeds, England.

making r too large. For instance, often a good choice is $r = 1$ (which would be impossible in the previous method). Then (8) becomes simply

$$4u_{i,j+1} - u_{i+1,j+1} - u_{i-1,j+1} = u_{i+1,j} + u_{i-1,j}. \tag{9}$$

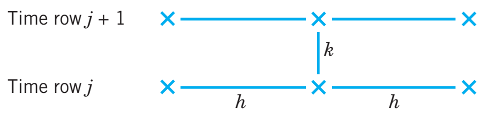


Fig. 467. The six points in the Crank–Nicolson formulas (7) and (8)

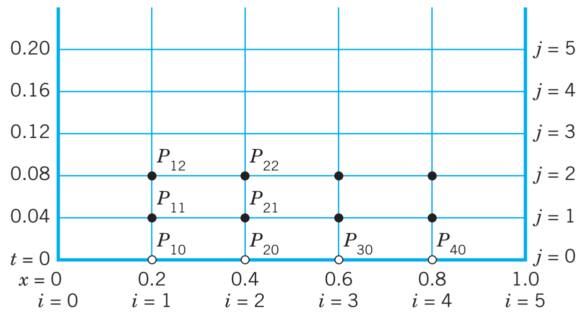


Fig. 468. Grid in Example 1

EXAMPLE 1 Temperature in a Metal Bar. Crank–Nicolson Method, Explicit Method

Consider a laterally insulated metal bar of length 1 and such that $c^2 = 1$ in the heat equation. Suppose that the ends of the bar are kept at temperature $u = 0°C$ and the temperature in the bar at some instant—call it $t = 0$—is $f(x) = \sin \pi x$. Applying the Crank–Nicolson method with $h = 0.2$ and $r = 1$, find the temperature $u(x, t)$ in the bar for $0 \leqq t \leqq 0.2$. Compare the results with the exact solution. Also apply (5) with an r satisfying (6), say, $r = 0.25$, and with values not satisfying (6), say, $r = 1$ and $r = 2.5$.

Solution by Crank–Nicolson. Since $r = 1$, formula (8) takes the form (9). Since $h = 0.2$ and $r = k/h^2 = 1$, we have $k = h^2 = 0.04$. Hence we have to do 5 steps. Figure 468 shows the grid. We shall need the initial values

$$u_{10} = \sin 0.2\pi = 0.587785, \qquad u_{20} = \sin 0.4\pi = 0.951057.$$

Also, $u_{30} = u_{20}$ and $u_{40} = u_{10}$. (Recall that u_{10} means u at P_{10} in Fig. 468, etc.) In each time row in Fig. 468 there are 4 internal mesh points. Hence in each time step we would have to solve 4 equations in 4 unknowns. But since the initial temperature distribution is symmetric with respect to $x = 0.5$, and $u = 0$ at both ends for all t, we have $u_{31} = u_{21}$, $u_{41} = u_{11}$ in the first time row and similarly for the other rows. This reduces each system to 2 equations in 2 unknowns. By (9), since $u_{31} = u_{21}$ and $u_{01} = 0$, for $j = 0$ these equations are

$$\begin{aligned} &(i = 1) \qquad 4u_{11} - u_{21} \phantom{{}- u_{21}} = u_{00} + u_{20} = 0.951057 \\ &(i = 2) \qquad -u_{11} + 4u_{21} - u_{21} = u_{10} + u_{20} = 1.538842. \end{aligned}$$

The solution is $u_{11} = 0.399274$, $u_{21} = 0.646039$. Similarly, for time row $j = 1$ we have the system

$$\begin{aligned} &(i = 1) \qquad 4u_{12} - u_{22} = u_{01} + u_{21} = 0.646039 \\ &(i = 2) \qquad -u_{12} + 3u_{22} = u_{11} + u_{21} = 1.045313. \end{aligned}$$

The solution is $u_{12} = 0.271221$, $u_{22} = 0.438844$, and so on. This gives the temperature distribution (Fig. 469):

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.00	0	0.588	0.951	0.951	0.588	0
0.04	0	0.399	0.646	0.646	0.399	0
0.08	0	0.271	0.439	0.439	0.271	0
0.12	0	0.184	0.298	0.298	0.184	0
0.16	0	0.125	0.202	0.202	0.125	0
0.20	0	0.085	0.138	0.138	0.085	0

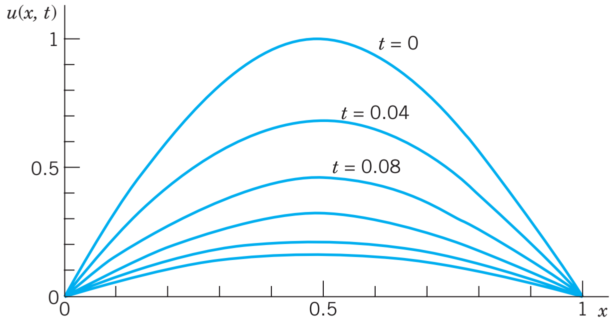


Fig. 469. Temperature distribution in the bar in Example 1

Comparison with the exact solution. The present problem can be solved exactly by separating variables (Sec. 12.5); the result is

$$u(x, t) = \sin \pi x \, e^{-\pi^2 t}. \tag{10}$$

Solution by the explicit method (5) with $r = 0.25$. For $h = 0.2$ and $r = k/h^2 = 0.25$ we have $k = rh^2 = 0.25 \cdot 0.04 = 0.01$. Hence we have to perform 4 times as many steps as with the Crank–Nicolson method! Formula (5) with $r = 0.25$ is

$$u_{i,j+1} = 0.25(u_{i-1,j} + 2u_{ij} + u_{i+1,j}). \tag{11}$$

We can again make use of the symmetry. For $j = 0$ we need $u_{00} = 0$, $u_{10} = 0.587785$ (see p. 939), $u_{20} = u_{30} = 0.951057$ and compute

$$u_{11} = 0.25(u_{00} + 2u_{10} + u_{20}) = 0.531657$$

$$u_{21} = 0.25(u_{10} + 2u_{20} + u_{30}) = 0.25(u_{10} + 3u_{20}) = 0.860239.$$

Of course we can omit the boundary terms $u_{01} = 0$, $u_{02} = 0, \cdots$ from the formulas. For $j = 1$ we compute

$$u_{12} = 0.25(2u_{11} + u_{21}) = 0.480888$$

$$u_{22} = 0.25(u_{11} + 3u_{21}) = 0.778094$$

and so on. We have to perform 20 steps instead of the 5 CN steps, but the numeric values show that the accuracy is only about the same as that of the Crank–Nicolson values CN. The exact 3D-values follow from (10).

t	$x = 0.2$			$x = 0.4$		
	CN	By (11)	Exact	CN	By (11)	Exact
0.04	0.399	0.393	0.396	0.646	0.637	0.641
0.08	0.271	0.263	0.267	0.439	0.426	0.432
0.12	0.184	0.176	0.180	0.298	0.285	0.291
0.16	0.125	0.118	0.121	0.202	0.191	0.196
0.20	0.085	0.079	0.082	0.138	0.128	0.132

Failure of (5) with r violating (6). Formula (5) with $h = 0.2$ and $r = 1$—which violates (6)—is

$$u_{i,j+1} = u_{i-1,j} - u_{ij} + u_{i+1,j}$$

and gives very poor values; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.04	0.363	0.396	0.588	0.641
0.12	0.139	0.180	0.225	0.291
0.20	0.053	0.082	0.086	0.132

Formula (5) with an even larger $r = 2.5$ (and $h = 0.2$ as before) gives completely nonsensical results; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.1	0.0265	0.2191	0.0429	0.3545
0.3	0.0001	0.0304	0.0001	0.0492.

PROBLEM SET 21.6

1. **Nondimensional form.** Show that the heat equation $\tilde{u}_{\tilde{t}} = c^2\tilde{u}_{\tilde{x}\tilde{x}}$, $0 \leqq \tilde{x} \leqq L$, can be transformed to the "nondimensional" standard form $u_t = u_{xx}$, $0 \leqq x \leqq 1$, by setting $x = \tilde{x}/L$, $t = c^2\tilde{t}/L^2$, $u = \tilde{u}/u_0$, where u_0 is any constant temperature.
2. **Difference equation.** Derive the difference approximation (4) of the heat equation.
3. **Explicit method.** Derive (5) by solving (4) for $u_{i,j+1}$.
4. **CAS EXPERIMENT. Comparison of Methods.**

   (a) Write programs for the explicit and the Crank—Nicolson methods.

   (b) Apply the programs to the heat problem of a laterally insulated bar of length 1 with $u(x, 0) = \sin \pi x$ and $u(0, t) = u(1, t) = 0$ for all t, using $h = 0.2$, $k = 0.01$ for the explicit method (20 steps), $h = 0.2$ and (9) for the Crank–Nicolson method (5 steps). Obtain exact 6D-values from a suitable series and compare.

   (c) Graph temperature curves in (b) in two figures similar to Fig. 299 in Sec. 12.7.

   (d) Experiment with smaller h (0.1, 0.05, etc.) for both methods to find out to what extent accuracy increases under systematic changes of h and k.

EXPLICIT METHOD

5. Using (5) with $h = 1$ and $k = 0.5$, solve the heat problem (1)–(3) to find the temperature at $t = 2$ in a laterally insulated bar of length 10 ft and initial temperature $f(x) = x(1 - 0.1x)$.
6. Solve the heat problem (1)–(3) by the explicit method with $h = 0.2$ and $k = 0.01$, 8 time steps, when $f(x) = x$ if $0 \leqq x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} \leqq x \leqq 1$. Compare with the 3S-values 0.108, 0.175 for $t = 0.08$, $x = 0.2, 0.4$ obtained from the series (2 terms) in Sec. 12.5.
7. The accuracy of the explicit method depends on $r\ (\leqq \frac{1}{2})$. Illustrate this for Prob. 6, choosing $r = \frac{1}{2}$ (and $h = 0.2$ as before). Do 4 steps. Compare the values for $t = 0.04$ and 0.08 with the 3S-values in Prob. 6, which are 0.156, 0.254 ($t = 0.04$), 0.105, 0.170 ($t = 0.08$).

8. In a laterally insulated bar of length 1 let the initial temperature be $f(x) = x$ if $0 \leqq x < 0.5$, $f(x) = 1 - x$ if $0.5 \leqq x \leqq 1$. Let (1) and (3) hold. Apply the explicit method with $h = 0.2$, $k = 0.01$, 5 steps. Can you expect the solution to satisfy $u(x, t) = u(1 - x, t)$ for all t?

9. Solve Prob. 8 with $f(x) = x$ if $0 \leqq x \leqq 0.2$, $f(x) = 0.25(1 - x)$ if $0.2 < x \leqq 1$, the other data being as before.

10. Insulated end. If the left end of a laterally insulated bar extending from $x = 0$ to $x = 1$ is insulated, the boundary condition at $x = 0$ is $u_n(0, t) = u_x(0, t) = 0$. Show that, in the application of the explicit method given by (5), we can compute u_{0j+1} by the formula

$$u_{0j+1} = (1 - 2r)u_{0j} + 2ru_{1j}.$$

Apply this with $h = 0.2$ and $r = 0.25$ to determine the temperature $u(x, t)$ in a laterally insulated bar extending from $x = 0$ to 1 if $u(x, 0) = 0$, the left end is insulated and the right end is kept at temperature $g(t) = \sin \frac{50}{3}\pi t$. *Hint.* Use $0 = \partial u_{0j}/\partial x = (u_{1j} - u_{-1j})/2h$.

CRANK–NICOLSON METHOD

11. Solve Prob. 9 by (9) with $h = 0.2$, 2 steps. Compare with exact values obtained from the series in Sec. 12.5 (2 terms) with suitable coefficients.

12. Solve the heat problem (1)–(3) by Crank–Nicolson for $0 \leqq t \leqq 0.20$ with $h = 0.2$ and $k = 0.04$ when $f(x) = x$ if $0 \leqq x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} \leqq x \leqq 1$. Compare with the exact values for $t = 0.20$ obtained from the series (2 terms) in Sec. 12.5.

13–15

Solve (1)–(3) by Crank–Nicolson with $r = 1$ (5 steps), where:

13. $f(x) = 5x$ if $0 \leqq x < 0.25$, $f(x) = 1.25(1 - x)$ if $0.25 \leqq x \leqq 1$, $h = 0.2$

14. $f(x) = x(1 - x)$, $h = 0.1$. (Compare with Prob. 15.)

15. $f(x) = x(1 - x)$, $h = 0.2$

21.7 Method for Hyperbolic PDEs

In this section we consider the numeric solution of problems involving hyperbolic PDEs. We explain a standard method in terms of a typical setting for the prototype of a hyperbolic PDE, the **wave equation**:

(1)	$u_{tt} = u_{xx}$	$0 \leqq x \leqq 1, t \geqq 0$
(2)	$u(x, 0) = f(x)$	(Given initial displacement)
(3)	$u_t(x, 0) = g(x)$	(Given initial velocity)
(4)	$u(0, t) = u(1, t) = 0$	(Boundary conditions).

Note that an equation $u_{tt} = c^2u_{xx}$ and another x-interval can be reduced to the form (1) by a linear transformation of x and t. This is similar to Sec. 21.6, Prob. 1.

For instance, (1)–(4) is the model of a vibrating elastic string with fixed ends at $x = 0$ and $x = 1$ (see Sec. 12.2). Although an analytic solution of the problem is given in (13), Sec. 12.4, we use the problem for explaining basic ideas of the numeric approach that are also relevant for more complicated hyperbolic PDEs.

Replacing the derivatives by difference quotients as before, we obtain from (1) [see (6) in Sec. 21.4 with $y = t$]

$$\frac{1}{k^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \tag{5}$$

where h is the mesh size in x, and k is the mesh size in t. This difference equation relates 5 points as shown in Fig. 470a. It suggests a rectangular grid similar to the grids for parabolic equations in the preceding section. We choose $r^* = k^2/h^2 = 1$. Then u_{ij} drops out and we have

$$u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{1,j-1} \qquad \text{(Fig. 470b)}. \tag{6}$$

It can be shown that for $0 < r^* \leqq 1$ the present **explicit method** is stable, so that from (6) we may expect reasonable results for initial data that have no discontinuities. (For a hyperbolic PDE the latter would propagate into the solution domain—a phenomenon that would be difficult to deal with on our present grid. For unconditionally stable **implicit methods** see [E1] in App. 1.)

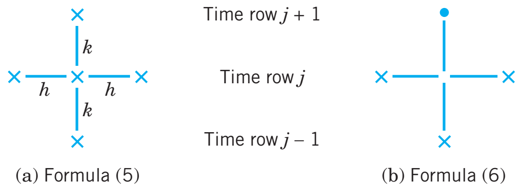


Fig. 470. Mesh points used in (5) and (6)

Equation (6) still involves 3 time steps $j - 1, j, j + 1$, whereas the formulas in the parabolic case involved only 2 time steps. Furthermore, we now have 2 initial conditions. So we ask how we get started and how we can use the initial condition (3). This can be done as follows.

From $u_t(x, 0) = g(x)$ we derive the difference formula

$$\frac{1}{2k}(u_{i1} - u_{i,-1}) = g_i, \qquad \text{hence} \qquad u_{i,-1} = u_{i1} - 2kg_i \tag{7}$$

where $g_i = g(ih)$. For $t = 0$, that is, $j = 0$, equation (6) is

$$u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i,-1}.$$

Into this we substitute $u_{i,-1}$ as given in (7). We obtain $u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i1} + 2kg_i$ and by simplification

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}) + kg_i, \tag{8}$$

This expresses u_{i1} in terms of the initial data. It is for the beginning only. Then use (6).

EXAMPLE 1 Vibrating String, Wave Equation

Apply the present method with $h = k = 0.2$ to the problem (1)–(4), where

$$f(x) = \sin \pi x, \qquad g(x) = 0.$$

Solution. The grid is the same as in Fig. 468, Sec. 21.6, except for the values of t, which now are $0.2, 0.4, \cdots$ (instead of $0.04, 0.08, \cdots$). The initial values $u_{00}, u_{10}, \cdots$ are the same as in Example 1, Sec. 21.6. From (8) and $g(x) = 0$ we have

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}).$$

From this we compute, using $u_{10} = u_{40} = \sin 0.2\pi = 0.587785$, $u_{20} = u_{30} = 0.951057$,

$$(i = 1) \quad u_{11} = \tfrac{1}{2}(u_{00} + u_{20}) = \tfrac{1}{2} \cdot 0.951057 = 0.475528$$

$$(i = 2) \quad u_{21} = \tfrac{1}{2}(u_{10} + u_{30}) = \tfrac{1}{2} \cdot 1.538842 = 0.769421$$

and $u_{31} = u_{21}$, $u_{41} = u_{11}$ by symmetry as in Sec. 21.6, Example 1. From (6) with $j = 1$ we now compute, using $u_{01} = u_{02} = \cdots = 0$,

$$(i = 1) \qquad u_{12} = u_{01} + u_{21} - u_{10} = 0.769421 - 0.587785 \qquad = 0.181636$$

$$(i = 2) \qquad u_{22} = u_{11} + u_{31} - u_{20} = 0.475528 + 0.769421 - 0.951057 = 0.293892,$$

and $u_{32} = u_{22}$, $u_{42} = u_{12}$ by symmetry; and so on. We thus obtain the following values of the displacement $u(x, t)$ of the string over the first half-cycle:

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.0	0	0.588	0.951	0.951	0.588	0
0.2	0	0.476	0.769	0.769	0.476	0
0.4	0	0.182	0.294	0.294	0.182	0
0.6	0	−0.182	−0.294	−0.294	−0.182	0
0.8	0	−0.476	−0.769	−0.769	−0.476	0
1.0	0	−0.588	−0.951	−0.951	−0.588	0

These values are exact to 3D (3 decimals), the exact solution of the problem being (see Sec. 12.3)

$$u(x, t) = \sin \pi x \cos \pi t.$$

The reason for the exactness follows from d'Alembert's solution (4), Sec. 12.4. (See Prob. 4, below.) ■

This is the end of Chap. 21 on numerics for ODEs and PDEs, a field that continues to develop rapidly in both applications and theoretical research. Much of the activity in the field is due to the computer serving as an invaluable tool for solving large-scale and complicated practical problems as well as for testing and experimenting with innovative ideas. These ideas could be small or major improvements on existing numeric algorithms or testing new algorithms as well as other ideas.

PROBLEM SET 21.7

VIBRATING STRING

1–3 Using the present method, solve (1)–(4) with $h = k = 0.2$ for the given initial deflection $f(x)$ and initial velocity 0 on the given t-interval.

1. $f(x) = x$ if $0 = x < \frac{1}{5}$, $\quad f(x) = \frac{1}{4}(1 - x)$ if $\frac{1}{5} \leqq x \leqq 1$, $0 \leqq t \leqq 1$

2. $f(x) = x^2 - x^3$, $\quad 0 \leqq t \leqq 2$

3. $f(x) = 0.2(x - x^2)$, $\quad 0 \leqq t \leqq 2$

4. **Another starting formula.** Show that (12) in Sec. 12.4 gives the starting formula

$$u_{i,1} = \frac{1}{2}(u_{i+1,0} + u_{i-1,0}) + \frac{1}{2}\int_{x_i-k}^{x_i+k} g(s)\, ds$$

(where one can evaluate the integral numerically if necessary). In what case is this identical with (8)?

5. **Nonzero initial displacement and speed.** Illustrate the starting procedure when both f and g are not identically

zero, say, $f(x) = 1 - \cos 2\pi x$, $g(x) = x(1 - x)$, $h = k = 0.1$, 2 time steps.

6. Solve (1)–(3) ($h = k = 0.2$, 5 time steps) subject to $f(x) = x^2$, $g(x) = 2x$, $u_x(0, t) = 2t$, $u(1, t) = (1 + t)^2$.

7. Zero initial displacement. If the string governed by the wave equation (1) starts from its equilibrium position with initial velocity $g(x) = \sin \pi x$, what is its displacement at time $t = 0.4$ and $x = 0.2, 0.4, 0.6, 0.8$? (Use the present method with $h = 0.2$, $k = 0.2$. Use (8). Compare with the exact values obtained from (12) in Sec. 12.4.)

8. Compute approximate values in Prob. 7, using a finer grid ($h = 0.1$, $k = 0.1$), and notice the increase in accuracy.

9. Compute u in Prob. 5 for $t = 0.1$ and $x = 0.1, 0.2, \cdots, 0.9$, using the formula in Prob. 8, and compare the values.

10. Show that from d'Alembert's solution (13) in Sec. 12.4 with $c = 1$ it follows that (6) in the present section gives the exact value $u_{i,j+1} = u(ih, (j + 1)h)$.

CHAPTER 21 REVIEW QUESTIONS AND PROBLEMS

1. Explain the Euler and improved Euler methods in geometrical terms. Why did we consider these methods?

2. How did we obtain numeric methods from the Taylor series?

3. What are the local and the global orders of a method? Give examples.

4. Why did we compute auxiliary values in each Runge–Kutta step? How many?

5. What is adaptive integration? How does its idea extend to Runge–Kutta?

6. What are one-step methods? Multistep methods? The underlying ideas? Give examples.

7. What does it mean that a method is not self-starting? How do we overcome this problem?

8. What is a predictor–corrector method? Give an important example.

9. What is automatic step size control? When is it needed? How is it done in practice?

10. How do we extend Runge–Kutta to systems of ODEs?

11. Why did we have to treat the main types of PDEs in separate sections? Make a list of types of problems and numeric methods.

12. When and how did we use finite differences? Give as many details as you can remember without looking into the text.

13. How did we approximate the Laplace and Poisson equations?

14. How many initial conditions did we prescribe for the wave equation? For the heat equation?

15. Can we expect a difference equation to give the exact solution of the corresponding PDE?

16. In what method for PDEs did we have convergence problems?

17. Solve $y' = y$, $y(0) = 1$ by Euler's method, 10 steps, $h = 0.1$.

18. Do Prob. 17 with $h = 0.01$, 10 steps. Compute the errors. Compare the error for $x = 0.1$ with that in Prob. 17.

19. Solve $y' = 1 + y^2$, $y(0) = 0$ by the improved Euler method, $h = 0.1$, 10 steps.

20. Solve $y' + y = (x + 1)^2$, $y(0) = 3$ by the improved Euler method, 10 steps with $h = 0.1$. Determine the errors.

21. Solve Prob. 19 by RK with $h = 0.1$, 5 steps. Compute the error. Compare with Prob. 19.

22. Fair comparison. Solve $y' = 2x^{-1}\sqrt{y - \ln x} + x^{-1}$, $y(1) = 0$ for $1 \leqq x \leqq 1.8$ **(a)** by the Euler method with $h = 0.1$, **(b)** by the improved Euler method with $h = 0.2$, and **(c)** by RK with $h = 0.4$. Verify that the exact solution is $y = (\ln x)^2 + \ln x$. Compute and compare the errors. Why is the comparison fair?

23. Apply the Adams–Moulton method to $y' = \sqrt{1 - y^2}$, $y(0) = 0$, $h = 0.2$, $x = 0, \cdots, 1$, starting with 0.198668, 0.389416, 0.564637.

24. Apply the A–M method to $y' = (x + y - 4)^2$, $y(0) = 4$, $h = 0.2$, $x = 0, \cdots, 1$, starting with 4.00271, 4.02279, 4.08413.

25. Apply Euler's method for systems to $y'' = x^2 y$, $y(0) = 1$, $y'(0) = 0$, $h = 0.1$, 5 steps.

26. Apply Euler's method for systems to $y_1' = y_2$, $y_2' = -4y_1$, $y_1(0) = 2$, $y_2(0) = 0$, $h = 0.2$, 10 steps. Sketch the solution.

27. Apply Runge–Kutta for systems to $y'' + y = 2e^x$, $y(0) = 0$, $y'(0) = 1$, $h = 0.2$, 5 steps. Determine the errors.

28. Apply Runge–Kutta for systems to $y_1' = 6y_1 + 9y_2$, $y_2' = y_1 + 6y_2$, $y_1(0) = -3$, $y_2(0) = -3$, $h = 0.05$, 3 steps.

29. Find rough approximate values of the electrostatic potential at P_{11}, P_{12}, P_{13} in Fig. 471 that lie in a field between conducting plates (in Fig. 471 appearing as sides of a rectangle) kept at potentials 0 and 220 V as shown. (Use the indicated grid.)

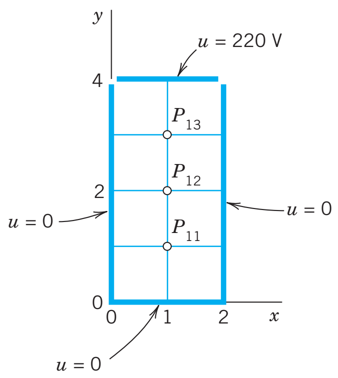


Fig. 471. Problem 29

30. A laterally insulated homogeneous bar with ends at $x = 0$ and $x = 1$ has initial temperature 0. Its left end is kept at 0, whereas the temperature at the right end varies sinusoidally according to

$$u(t, 1) = g(t) = \sin \tfrac{25}{3} \pi t.$$

Find the temperature $u(x, t)$ in the bar [solution of (1) in Sec. 21.6] by the explicit method with $h = 0.2$ and $r = 0.5$ (one period, that is, $0 \leqq t \leqq 0.24$).

31. Find the solution of the vibrating string problem $u_{tt} = u_{xx}$, $u(x, 0) = x(1 - x)$, $u_t = 0$, $u(0, t) = u(1, t) = 0$ by the method in Sec. 21.7 with $h = 0.1$ and $k = 0.1$ for $t = 0.3$.

32–34 POTENTIAL

Find the potential in Fig. 472, using the given grid and the boundary values:

32. $u(P_{01}) = u(P_{03}) = u(P_{41}) = u(P_{43}) = 200$, $u(P_{10}) = u(P_{30}) = -400$, $u(P_{20}) = 1600$, $u(P_{02}) = u(P_{42}) = u(P_{14}) = u(P_{24}) = u(P_{34}) = 0$

33. $u(P_{10}) = u(P_{30}) = 960$, $u(P_{20}) = -480$, $u = 0$ elsewhere on the boundary

34. $u = 70$ on the upper and left sides, $u = 0$ on the lower and right sides

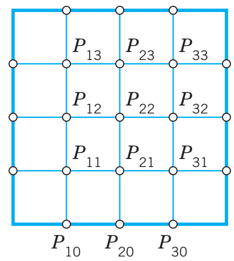


Fig. 472. Problems 32–34

35. Solve $u_t = u_{xx}$ $(0 \leqq x \leqq 1, t \geqq 0)$, $u(x, 0) = x^2(1 - x)$, $u(0, t) = u(1, t) = 0$ by Crank–Nicolson with $h = 0.2$, $k = 0.04$, 5 time steps.

SUMMARY OF CHAPTER 21
Numerics for ODEs and PDEs

In this chapter we discussed numerics for ODEs (Secs. 21.1–21.3) and PDEs (Secs. 21.4–21.7). Methods for initial value problems

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

involving a first-order ODE are obtained by truncating the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots$$

where, by (1), $y' = f, y'' = f' = \partial f/\partial x + (\partial f/\partial y)y'$, etc. Truncating after the term hy', we get the *Euler method,* in which we compute step by step

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots). \tag{2}$$

Taking one more term into account, we obtain the *improved Euler method.* Both methods show the basic idea but are too inaccurate in most cases.

Truncating after the term in h^4, we get the important classical **Runge–Kutta (RK) method** of fourth order. The crucial idea in this method is the replacement of the cumbersome evaluation of derivatives by the evaluation of $f(x, y)$ at suitable points (x, y); thus in each step we first compute four auxiliary quantities (Sec. 21.1)

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1) \\ k_3 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_2) \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned} \tag{3a}$$

and then the new value

$$y_{n+1} = y_n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \tag{3b}$$

Error and step size control are possible by step halving or by **RKF** (Runge–Kutta–Fehlberg).

The methods in Sec. 21.1 are **one-step methods** since they get y_{n+1} from the result y_n of a single step. A **multistep method** (Sec. 21.2) uses the values of $y_n, y_{n-1}, \cdots$ of several steps for computing y_{n+1}. Integrating cubic interpolation polynomials gives the **Adams–Bashforth predictor** (Sec. 21.2)

$$y^*_{n+1} = y_n + \tfrac{1}{24}h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \tag{4a}$$

where $f_j = f(x_j, y_j)$, and an **Adams–Moulton corrector** (the actual new value)

$$y_{n+1} = y_n + \tfrac{1}{24}h(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}), \tag{4b}$$

where $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$. Here, to get started, y_1, y_2, y_3 must be computed by the Runge–Kutta method or by some other accurate method.

Section 19.3 concerned the extension of Euler and RK methods to systems

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \text{thus} \qquad y_j' = f_j(x, y_1, \cdots, y_m), \qquad j = 1, \cdots, m.$$

This includes single mth-order ODEs, which are reduced to systems. Second-order equations can also be solved by **RKN** (Runge–Kutta–Nyström) **methods**. These are particularly advantageous for $y'' = f(x, y)$ with f not containing y'.

Numeric methods for PDEs are obtained by replacing partial derivatives by difference quotients. This leads to approximating difference equations, for the **Laplace equation** to

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0 \qquad (5) \qquad \text{(Sec. 21.4)}$$

for the **heat equation** to

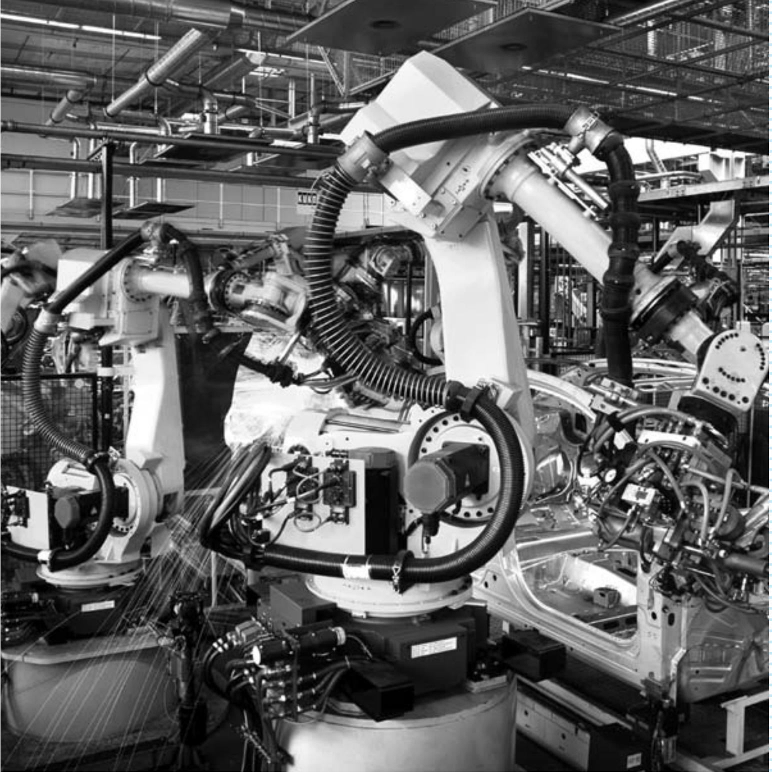
$$\frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad (6) \qquad \text{(Sec. 21.6)}$$

and for the **wave equation** to

$$\frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad (7) \qquad \text{(Sec. 21.7);}$$

here h and k are the mesh sizes of a grid in the x- and y-directions, respectively, where in (6) and (7) the variable y is time t.

These PDEs are *elliptic, parabolic,* and *hyperbolic,* respectively. Corresponding numeric methods differ, for the following reason. For elliptic PDEs we have boundary value problems, and we discussed for them the ***Gauss–Seidel method*** (also known as ***Liebmann's method***) and the ***ADI method*** (Secs. 21.4, 21.5). For parabolic PDEs we are given one initial condition and boundary conditions, and we discussed an *explicit method* and the ***Crank–Nicolson method*** (Sec. 21.6). For hyperbolic PDEs, the problems are similar but we are given a second initial condition (Sec. 21.7).



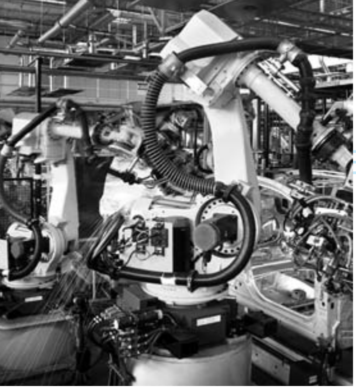
PART F

Optimization, Graphs

The material of Part F is particularly useful in modeling large-scale real-world problems. Just as it is in numerics in Part E, where the greater availability of quality software and computing power is a deciding factor in the continued growth of the field, so it is also in the fields of optimization and combinatorial optimization. Problems, such as optimizing production plans for different industries (microchips, pharmaceuticals, cars, aluminum, steel, chemicals), optimizing usage of transportation systems (usage of runways in airports, tracks of subways), efficiency in running of power plants, optimal shipping (delivery services, shipping of containers, shipping goods from factories to warehouses and from warehouses to stores), designing optimal financial portfolios, and others are all examples where the size of the problem usually requires the use of optimization software. More recently, environmental concerns have put new aspects into the picture, where an important concern, added to these problems, is the minimization of environmental impact. The main task becomes to model these problems correctly. The purpose of Part F is to introduce the main ideas and methods of unconstrained and constrained optimization (Chap. 22), and graphs and combinatorial optimization (Chap. 23).

Chapter 22 introduces unconstrained optimization by the method of ***steepest descent*** and constrained optimization by the versatile ***simplex method***. The simplex method (Secs. 22.3, 22.4) is very useful for solving many linear optimization problems (also called linear programming problems).

Graphs let us model problems in transportation logistics, efficient use of communication networks, best assignment of workers to jobs, and others. We consider shortest path problems (Secs. 22.2, 22.3), shortest spanning trees (Secs. 23.4, 23.5), flow problems in networks (Secs. 23.6, 23.7), and assignment problems (Sec. 23.8). We discuss algorithms of Moore, Dijkstra (both for shortest path), Kruskal, Prim (shortest spanning trees), and Ford–Fulkerson (for flow).



CHAPTER 22

Unconstrained Optimization. Linear Programming

Optimization is a general term used to describe types of problems and solution techniques that are concerned with the best ("optimal") allocation of limited resources in projects. The problems are called optimization problems and the methods optimization methods. Typical problems are concerned with planning and making decisions, such as selecting an optimal production plan. A company has to decide how many units of each product from a choice of (distinct) products it should make. The objective of the company may be to maximize overall profit when the different products have different individual profits. In addition, the company faces certain limitations (constraints). It may have a certain number of machines, it takes a certain amount of time and usage of these machines to make a product, it requires a certain number of workers to handle the machines, and other possible criteria. To solve such a problem, you assign the first variable to number of units to be produced of the first product, the second variable to the second product, up to the number of different (distinct) products the company makes. When you multiply these, for example, by the price, you obtain a linear function called the objective function. You also express the constraints in terms of these variables, thereby obtaining several inequalities, called the constraints. Because the variables in the objective function also occur in the constraints, the objective function and the constraints are tied mathematically to each other and you have set up a linear optimization problem, also called a ***linear programming problem***.

The main focus of this chapter is to set up (Sec. 22.2) and solve (Secs. 22.3, 22.4) such linear programming problems. A famous and versatile method for doing so is the simplex method. In the ***simplex method***, the objective function and the constraints are set up in the form of an augmented matrix as in Sec. 7.3, however, the method of solving such linear constrained optimization problems is a new approach.

The beauty of the simplex method is that it allows us to scale problems up to thousands or more constraints, thereby modeling real-world situations. We can start with a small model and gradually add more and more constraints. The most difficult part is modeling the problem correctly. The actual task of solving large optimization problems is done by software implementations for the simplex method or perhaps by other optimization methods.

Besides optimal production plans, problems in optimal shipping, optimal location of warehouses and stores, easing traffic congestion, efficiency in running power plants are all examples of applications of optimization. More recent applications are in minimizing environmental damages due to pollutants, carbon dioxide emissions, and other factors. Indeed, new fields of green logistics and green manufacturing are evolving and naturally make use of optimization methods.

Prerequisite: a modest working knowledge of linear systems of equations.
References and Answers to Problems: App. 1 Part F, App. 2.

22.1 Basic Concepts. Unconstrained Optimization: Method of Steepest Descent

In an **optimization problem** the objective is to *optimize* (*maximize* or *minimize*) some function f. This function f is called the **objective function**. It is the focal point or goal of our optimization problem.

For example, an objective function f to be *maximized* may be the revenue in a production of TV sets, the rate of return of a financial portfolio, the yield per minute in a chemical process, the mileage per gallon of a certain type of car, the hourly number of customers served in a bank, the hardness of steel, or the tensile strength of a rope.

Similarly, we may want to *minimize* f if f is the cost per unit of producing certain cameras, the operating cost of some power plant, the daily loss of heat in a heating system, CO_2 emissions from a fleet of trucks for freight transport, the idling time of some lathe, or the time needed to produce a fender.

In most optimization problems the objective function f depends on several variables

$$x_1, \cdots, x_n.$$

These are called **control variables** because we can "control" them, that is, choose their values.

For example, the yield of a chemical process may depend on pressure x_1 and temperature x_2. The efficiency of a certain air-conditioning system may depend on temperature x_1, air pressure x_2, moisture content x_3, cross-sectional area of outlet x_4, and so on.

Optimization theory develops methods for optimal choices of $x_1, \cdots, x_n$, which maximize (or minimize) the objective function f, that is, methods for finding optimal values of $x_1, \cdots, x_n$.

In many problems the choice of values of $x_1, \cdots, x_n$ is not entirely free but is subject to some **constraints**, that is, additional restrictions arising from the nature of the problem and the variables.

For example, if x_1 is production cost, then $x_1 \geqq 0$, and there are many other variables (time, weight, distance traveled by a salesman, etc.) that can take nonnegative values only. Constraints can also have the form of equations (instead of inequalities).

We first consider **unconstrained optimization** in the case of a function $f(x_1, \cdots, x_n)$. We also write $\mathbf{x} = (x_1, \cdots, x_n)$ and $f(\mathbf{x})$, for convenience.

By definition, f has a **minimum** at a point $\mathbf{x} = \mathbf{X}_0$ in a region R (where f is defined) if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Similarly, f has a **maximum** at $\mathbf{X}_0$ in R if

$$f(\mathbf{x}) \leqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Minima and maxima together are called **extrema**.

Furthermore, f is said to have a **local minimum** at $\mathbf{X}_0$ if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in a neighborhood of $\mathbf{X}_0$, say, for all $\mathbf{x}$ satisfying

$$|\mathbf{x} - \mathbf{X}_0| = [(x_1 - X_1)^2 + \cdots + (x_n - X_n)^2]^{1/2} < r,$$

where $\mathbf{X}_0 = (X_1, \cdots, X_n)$ and $r > 0$ is sufficiently small.

Similarly, f has a **local maximum** at $\mathbf{X}_0$ if $f(\mathbf{x}) \leqq f(\mathbf{X}_0)$ for all $\mathbf{x}$ satisfying $|\mathbf{x} - \mathbf{X}_0| < r$.

If f is differentiable and has an extremum at a point $\mathbf{X}_0$ in the *interior of a region R* (that is, not on the boundary), then the partial derivatives $\partial f/\partial x_1, \cdots, \partial f/\partial x_n$ must be zero at $\mathbf{X}_0$. These are the components of a vector that is called the **gradient** of f and denoted by grad f or ∇f. (For $n = 3$ this agrees with Sec. 9.7.) Thus

$$\nabla f(\mathbf{X}_0) = \mathbf{0}. \tag{1}$$

A point $\mathbf{X}_0$ at which (1) holds is called a **stationary point** of f.

Condition (1) is necessary for an extremum of f at $\mathbf{X}_0$ in the interior of R, but is not sufficient. Indeed, if $n = 1$, then for $y = f(x)$, condition (1) is $y' = f'(X_0) = 0$; and, for instance, $y = x^3$ satisfies $y' = 3x^2 = 0$ at $x = X_0 = 0$ where f has no extremum but a point of inflection. Similarly, for $f(\mathbf{x}) = x_1x_2$ we have $\nabla f(\mathbf{0}) = \mathbf{0}$, and f does not have an extremum but has a saddle point at $\mathbf{0}$. Hence, after solving (1), one must still find out whether one has obtained an extremum. In the case $n = 1$ the conditions $y'(X_0) = 0$, $y''(X_0) > 0$ guarantee a local minimum at X_0 and the conditions $y'(X_0) = 0$, $y''(X_0) < 0$ a local maximum, as is known from calculus. For $n > 1$ there exist similar criteria. However, in practice, even solving (1) will often be difficult. For this reason, one generally prefers solution by iteration, that is, by a search process that starts at some point and moves stepwise to points at which f is smaller (if a minimum of f is wanted) or larger (in the case of a maximum).

The **method of steepest descent** or **gradient method** is of this type. We present it here in its standard form. (For refinements see Ref. [E25] listed in App. 1.)

The idea of this method is to find a minimum of $f(\mathbf{x})$ by repeatedly computing minima of a function $g(t)$ of a single variable t, as follows. Suppose that f has a minimum at $\mathbf{X}_0$ and we start at a point $\mathbf{x}$. Then we look for a minimum of f closest to $\mathbf{x}$ along the straight line in the direction of $-\nabla f(\mathbf{x})$, which is the direction of steepest descent (= direction of maximum decrease) of f at $\mathbf{x}$. That is, we determine the value of t and the corresponding point

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) \tag{2}$$

at which the function

$$g(t) = f(\mathbf{z}(t)) \tag{3}$$

has a minimum. We take this $\mathbf{z}(t)$ as our next approximation to $\mathbf{X}_0$.

EXAMPLE 1 Method of Steepest Descent

Determine a minimum of

$$f(\mathbf{x}) = x_1^2 + 3x_2^2, \tag{4}$$

starting from $\mathbf{x}_0 = (6, 3) = 6\mathbf{i} + 3\mathbf{j}$ and applying the method of steepest descent.

Solution. Clearly, inspection shows that $f(\mathbf{x})$ has a minimum at $\mathbf{0}$. Knowing the solution gives us a better feel of how the method works. We obtain $\nabla f(\mathbf{x}) = 2x_1\mathbf{i} + 6x_2\mathbf{j}$ and from this

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) = (1 - 2t)x_1\mathbf{i} + (1 - 6t)x_2\mathbf{j}$$

$$g(t) = f(\mathbf{z}(t)) = (1 - 2t)^2x_1^2 + 3(1 - 6t)^2x_2^2.$$

We now calculate the derivative

$$g'(t) = 2(1 - 2t)x_1^2(-2) + 6(1 - 6t)x_2^2(-6),$$

set $g'(t) = 0$, and solve for t, finding

$$t = \frac{x_1^2 + 9x_2^2}{2x_1^2 + 54x_2^2}.$$

Starting from $x_0 = 6\mathbf{i} + 3\mathbf{j}$, we compute the values in Table 22.1, which are shown in Fig. 473.

Figure 473 suggests that in the case of slimmer ellipses ("a long narrow valley"), convergence would be poor. You may confirm this by replacing the coefficient 3 in (4) with a large coefficient. For more sophisticated descent and other methods, some of them also applicable to vector functions of vector variables, we refer to the references listed in Part F of App. 1; see also [E25]. ■

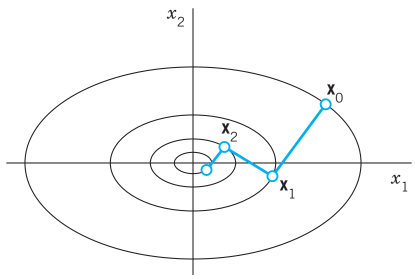


Fig. 473. Method of steepest descent in Example 1

Table 22.1 Method of Steepest Descent, Computations in Example 1

n	$\mathbf{x}$		t	$1 - 2t$	$1 - 6t$
0	6.000	3.000	0.210	0.581	−0.258
1	3.484	−0.774	0.310	0.381	−0.857
2	1.327	0.664	0.210	0.581	−0.258
3	0.771	−0.171	0.310	0.381	−0.857
4	0.294	0.147	0.210	0.581	−0.258
5	0.170	−0.038	0.310	0.381	−0.857
6	0.065	0.032			

PROBLEM SET 22.1

1. **Orthogonality.** Show that in Example 1, successive gradients are orthogonal (perpendicular). Why?
2. What happens if you apply the method of steepest descent to $f(\mathbf{x}) = x_1^2 + x_2^2$? First guess, then calculate.

3–9 STEEPEST DESCENT

Do steepest descent steps when:

3. $f(\mathbf{x}) = 2x_1^2 + x_2^2 - 4x_1 + 4x_2$, $\mathbf{x}_0 = \mathbf{0}$, 3 steps
4. $f(\mathbf{x}) = x_1^2 + 0.5x_2^2 - 5.0x_1 - 3.0x_2 + 24.95$, $x_0 = (3, 4)$, 5 steps
5. $f(\mathbf{x}) = ax_1 + bx_2$, $a \neq 0, b \neq 0$. First guess, then compute.
6. $f(\mathbf{x}) = x_1^2 - x_2^2$, $\mathbf{x}_0 = (1, 2)$, 5 steps. First guess, then compute. Sketch the path. What if $\mathbf{x}_0 = (2, 1)$?
7. $f(\mathbf{x}) = x_1^2 + cx_2^2$, $\mathbf{x}_0 = (c, 1)$. Show that 2 steps give $(c, 1)$ times a factor, $-4c^2/(c^2 - 1)^2$. What can you conclude from this about the speed of convergence?
8. $f(\mathbf{x}) = x_1^2 - x_2$, $\mathbf{x}_0 = (1, 1)$; 3 steps. Sketch your path. Predict the outcome of further steps.
9. $f(\mathbf{x}) = 0.1x_1^2 + x_2^2 - 0.02x_1$, $\mathbf{x}_0 = (3, 3)$, 5 steps

10. CAS EXPERIMENT. Steepest Descent. (a) Write a program for the method.

(b) Apply your program to $f(\mathbf{x}) = x_1^2 + 4x_2^2$, experimenting with respect to speed of convergence depending on the choice of $\mathbf{x}_0$.

(c) Apply your program to $f(\mathbf{x}) = x_1^2 + x_2^4$ and to $f(\mathbf{x}) = x_1^4 + x_2^4$, $\mathbf{x}_0 = (2, 1)$. Graph level curves and your path of descent. (Try to include graphing directly in your program.)

22.2 Linear Programming

Linear programming or **linear optimization** consists of methods for solving optimization problems ***with constraints***, that is, methods for finding a maximum (or a minimum) $\mathbf{x} = (x_1, \cdots, x_n)$ of a ***linear*** objective function

$$z = f(\mathbf{x}) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

satisfying the constraints. The latter are **linear inequalities**, such as $3x_1 + 4x_2 \leqq 36$, or $x_1 \geqq 0$, etc. (examples below). Problems of this kind arise frequently, almost daily, for instance, in production, inventory management, bond trading, operation of power plants, routing delivery vehicles, airplane scheduling, and so on. Progress in computer technology has made it possible to solve programming problems involving hundreds or thousands or more variables. Let us explain the setting of a linear programming problem and the idea of a "geometric" solution, so that we shall see what is going on.

EXAMPLE 1 Production Plan

Energy Savers, Inc., produces heaters of types S and L. The wholesale price is \$40 per heater for S and \$88 for L. Two time constraints result from the use of two machines M_1 and M_2. On M_1 one needs 2 min for an S heater and 8 min for an L heater. On M_2 one needs 5 min for an S heater and 2 min for an L heater. Determine production figures x_1 and x_2 for S and L, respectively (number of heaters produced per hour), so that the hourly revenue

$$z = f(\mathbf{x}) = 40x_1 + 88x_2$$

is maximum.

Solution. Production figures x_1 and x_2 must be nonnegative. Hence the objective function (to be maximized) and the four constraints are

$$
\begin{array}{llrl}
(0) & z = 40x_1 + 88x_2 & & \\
(1) & 2x_1 + 8x_2 & \leqq 60 & \text{min time on machine } M_1 \\
(2) & 5x_1 + 2x_2 & \leqq 60 & \text{min time on machine } M_2 \\
(3) & x_1 & \geqq 0 & \\
(4) & x_2 & \geqq 0. &
\end{array}
$$

Figure 474 shows (0)–(4) as follows. Constancy lines

$$z = \text{const}$$

are marked (0). These are **lines of constant revenue**. Their slope is $-40/88 = -5/11$. To increase z we must move the line upward (parallel to itself), as the arrow shows. Equation (1) with the equality sign is marked (1). It intersects the coordinate axes at $x_1 = 60/2 = 30$ (set $x_2 = 0$) and $x_2 = 60/8 = 7.5$ (set $x_1 = 0$). The arrow marks the side on which the points (x_1, x_2) lie that satisfy the inequality in (1). Similarly for Eqs. (2)–(4). The blue quadrangle thus obtained is called the **feasibility region**. It is the set of all **feasible solutions**, meaning solutions that satisfy all four constraints. The figure also lists the revenue at O, A, B, C. The optimal solution is obtained by moving the line of constant revenue up as much as possible without leaving the feasibility region completely. Obviously, this optimum is reached when that line passes through B, the intersection (10, 5) of (1) and (2). We see that the optimal revenue

$$z_{\max} = 40 \cdot 10 + 88 \cdot 5 = \$840$$

is obtained by producing twice as many S heaters as L heaters. ■

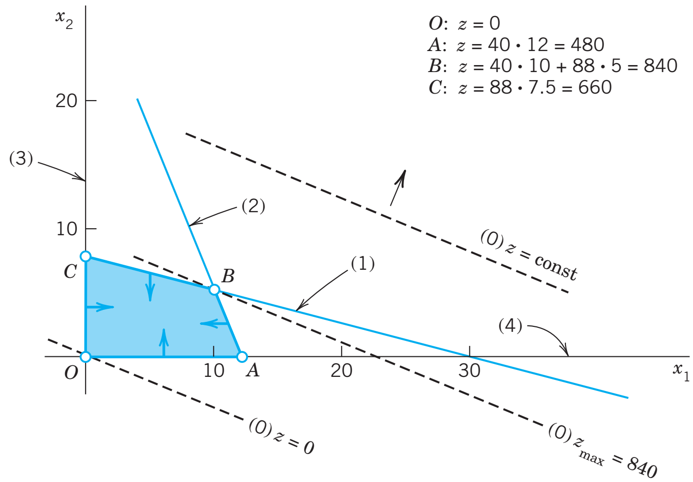


Fig. 474. Linear programming in Example 1

Note well that the problem in Example 1 or similar optimization problems *cannot* be solved by setting certain partial derivatives equal to zero, because crucial to such problems is the region in which the control variables are allowed to vary.

Furthermore, our "geometric" or graphic method illustrated in Example 1 is confined to two variables x_1, x_2. However, most practical problems involve much more than two variables, so that we need other methods of solution.

Normal Form of a Linear Programming Problem

To prepare for general solution methods, we show that constraints can be written more uniformly. Let us explain the idea in terms of (1),

$$2x_1 + 8x_2 \leqq 60.$$

This inequality implies $60 - 2x_1 - 8x_2 \geqq 0$ (and conversely), that is, the quantity

$$x_3 = 60 - 2x_1 - 8x_2$$

is nonnegative. Hence, our original inequality can now be written as an equation

$$2x_1 + 8x_2 + x_3 = 60,$$

where

$$x_3 \geqq 0.$$

x_3 is a nonnegative auxiliary variable introduced for converting inequalities to equations. Such a variable is called a **slack variable**, because it "takes up the slack" or difference between the two sides of the inequality.

EXAMPLE 2 Conversion of Inequalities by the Use of Slack Variables

With the help of two slack variables x_3, x_4 we can write the linear programming problem in Example 1 in the following form. ***Maximize***

$$f = 40x_1 + 88x_2$$

subject to the constraints

$$\begin{aligned} 2x_1 + 8x_2 + x_3 \qquad &= 60 \\ 5x_1 + 2x_2 \qquad + x_4 &= 60 \\ x_i \geqq 0 \qquad (i = 1, \cdots, 4)&. \end{aligned}$$

We now have $n = 4$ variables and $m = 2$ (linearly independent) equations, so that two of the four variables, for example, x_1, x_2, determine the others. Also note that each of the four sides of the quadrangle in Fig. 474 now has an equation of the form $x_i = 0$:

$$\begin{aligned} OA&: x_2 = 0, \\ AB&: x_4 = 0, \\ BC&: x_3 = 0, \\ CO&: x_1 = 0, \end{aligned}$$

A vertex of the quadrangle is the intersection of two sides. Hence at a vertex, $n - m = 4 - 2 = 2$ of the variables are zero and the others are nonnegative. Thus at A we have $x_2 = 0, x_4 = 0$, and so on. ■

Our example suggests that a general linear optimization problem can be brought to the following **normal form**. *Maximize*

$$f = c_1x_1 + c_2x_2 + \cdots + c_nx_n \tag{5}$$

subject to the constraints

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots&\cdots\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\ x_i \geqq 0 \qquad (i = 1, \cdots&, n) \end{aligned} \tag{6}$$

with all b_j nonnegative. (If a $b_j < 0$, multiply the equation by -1.) Here $x_1, \cdots, x_n$ include the slack variables (for which the c_j's in f are zero). We assume that the equations in (6) are linearly independent. Then, if we choose values for $n - m$ of the variables, the system uniquely determines the others. Of course, since we must have

$$x_1 \geqq 0, \cdots, x_n \geqq 0,$$

this choice is not entirely free.

Our problem also includes the **minimization** of an objective function f since this corresponds to maximizing $-f$ and thus needs no separate consideration.

An n-tuple $(x_1, \cdots, x_n)$ that satisfies all the constraints in (6) is called a *feasible point* or **feasible solution**. A feasible solution is called an **optimal solution** if, for it, the objective function f becomes maximum, compared with the values of f at all feasible solutions.

Finally, by a **basic feasible solution** we mean a feasible solution for which at least $n - m$ of the variables $x_1, \cdots, x_n$ are zero. For instance, in Example 2 we have $n = 4$, $m = 2$, and the basic feasible solutions are the four vertices O, A, B, C in Fig. 474. Here B is an optimal solution (the only one in this example).

The following theorem is fundamental.

THEOREM 1

Optimal Solution

Some optimal solution of a linear programming problem (5), (6) *is also a basic feasible solution of* (5), (6).

For a proof, see Ref. [F5], Chap. 3 (listed in App. 1). A problem can have many optimal solutions and not all of them may be *basic* feasible solutions; but the theorem guarantees that we can find an optimal solution by searching through the basic feasible solutions only. This is a great simplification; but since there are $\binom{n}{n-m} = \binom{n}{m}$ different ways of equating $n - m$ of the n variables to zero, considering all these possibilities, dropping those which are not feasible and then searching through the rest would still involve very much work, even when n and m are relatively small. Hence a systematic search is needed. We shall explain an important method of this type in the next section.

PROBLEM SET 22.2

1–6 REGIONS, CONSTRAINTS

Describe and graph the regions in the first quadrant of the x_1x_2-plane determined by the given inequalities.

1. $$\begin{aligned} x_1 - 3x_2 &\geqq -6 \\ x_1 + x_2 &\leqq 6 \end{aligned}$$

2. $$\begin{aligned} 2x_1 - x_2 &\geqq 6 \\ 8x_1 + 10x_2 &\leqq 80 \\ x_1 - 2x_2 &\geqq -3 \end{aligned}$$

3. $$\begin{aligned} -0.5x_1 + x_2 &\leqq 2 \\ x_1 + x_2 &\geqq 2 \\ -x_1 + 5x_2 &\geqq 5 \end{aligned}$$

4. $$\begin{aligned} -x_1 + x_2 &\leqq 5 \\ 2x_1 + x_2 &\geqq 10 \\ x_2 &\geqq 4 \\ 10x_1 + 15x_2 &\leqq 150 \end{aligned}$$

5. $$\begin{aligned} -x_1 + x_2 &\geqq 0 \\ x_1 + x_2 &\leqq 5 \\ -2x_1 + x_2 &\leqq 16 \end{aligned}$$

6. $$\begin{aligned} x_1 + x_2 &\geqq 2 \\ 3x_1 + 5x_2 &\geqq 15 \\ 2x_1 - x_2 &\geqq -2 \\ -x_1 + 2x_2 &\leqq 10 \end{aligned}$$

7. Location of maximum. Could we find a profit $f(x_1, x_2) = a_1x_1 + a_2x_2$ whose maximum is at an interior point of the quadrangle in Fig. 474? Give reason for your answer.

8. Slack variables. Why are slack variables always nonnegative? How many of them do we need?

9. What is the meaning of the slack variables x_3, x_4 in Example 2 in terms of the problem in Example 1?

10. Uniqueness. Can we always expect a unique solution (as in Example 1)?

11–16 **MAXIMIZATION, MINIMIZATION**

Maximize or minimize the given objective function f subject to the given constraints.

11. Maximize $f = 30x_1 + 10x_2$ in the region in Prob. 5.

12. Minimize $f = 45.0x_1 + 22.5x_2$ in the region in Prob. 4.

13. Maximize $f = 5x_1 + 25x_2$ in the region in Prob. 5.

14. Minimize $f = 5x_1 + 25x_2$ in the region in Prob. 3.

15. Maximize $f = 20x_1 + 30x_2$ subject to $4x_1 + 3x_2 \geqq 12$, $x_1 - x_2 \geqq -3$, $x_2 \leqq 6$, $2x_1 - 3x_2 \leqq 0$.

16. Maximize $f = -10x_1 + 2x_2$ subject to $x_1 \geqq 0$, $x_2 \geqq 0$, $-x_1 + x_2 \geqq -1$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 5$.

17. Maximum profit. United Metal, Inc., produces alloys B_1 (special brass) and B_2 (yellow tombac). B_1 contains 50% copper and 50% zinc. (Ordinary brass contains about 65% copper and 35% zinc.) B_2 contains 75% copper and 25% zinc. Net profits are \$120 per ton of B_1 and \$100 per ton of B_2. The daily copper supply is 45 tons. The daily zinc supply is 30 tons. Maximize the net profit of the daily production.

18. Maximum profit. The DC Drug Company produces two types of liquid pain killer, N (normal) and S (Super). Each bottle of N requires 2 units of drug A, 1 unit of drug B, and 1 unit of drug C. Each bottle of S requires 1 unit of A, 1 unit of B, and 3 units of C. The company is able to produce, each week, only 1400 units of A, 800 units of B, and 1800 units of C. The profit per bottle of N and S is \$11 and \$15, respectively. Maximize the total profit.

19. Maximum output. Giant Ladders, Inc., wants to maximize its daily total output of large step ladders by producing x_1 of them by a process P_1 and x_2 by a process P_2, where P_1 requires 2 hours of labor and 4 machine hours per ladder, and P_2 requires 3 hours of labor and 2 machine hours. For this kind of work, 1200 hours of labor and 1600 hours on the machines are, at most, available per day. Find the optimal x_1 and x_2.

20. Minimum cost. Hardbrick, Inc., has two kilns. Kiln I can produce 3000 gray bricks, 2000 red bricks, and 300 glazed bricks daily. For Kiln II the corresponding figures are 2000, 5000, and 1500. Daily operating costs of Kilns I and II are \$400 and \$600, respectively. Find the number of days of operation of each kiln so that the operation cost in filling an order of 18,000 gray, 34,000 red, and 9000 glazed bricks is minimized.

21. Maximum profit. Universal Electric, Inc., manufactures and sells two models of lamps, L_1 and L_2, the profit being \$150 and \$100, respectively. The process involves two workers W_1 and W_2 who are available for this kind of work 100 and 80 hours per month, respectively. W_1 assembles L_1 in 20 min and L_2 in 30 min. W_2 paints L_1 in 20 min and L_2 in 10 min. Assuming that all lamps made can be sold without difficulty, determine production figures that maximize the profit.

22. Nutrition. Foods A and B have 600 and 500 calories, contain 15 g and 30 g of protein, and cost \$1.80 and \$2.10 per unit, respectively. Find the minimum cost diet of at least 3900 calories containing at least 150 g of protein.

22.3 Simplex Method

From the last section we recall the following. A linear optimization problem (linear programming problem) can be written in normal form; that is:

Maximize

$$z = f(x) = c_1x_1 + \cdots + c_nx_n \tag{1}$$

subject to the constraints

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\ x_i \geqq 0 \qquad (i &= 1, \cdots, n). \end{aligned} \tag{2}$$

For finding an optimal solution of this problem, we need to consider only the **basic feasible solutions** (defined in Sec. 22.2), but there are still so many that we have to follow a systematic search procedure. In 1948 G. B. Dantzig[1] published an iterative method, called the **simplex method**, for that purpose. In this method, one proceeds stepwise from one basic feasible solution to another in such a way that the objective function f always increases its value. Let us explain this method in terms of the example in the last section.

In its original form the problem concerned the maximization of the objective function

$$z = 40x_1 + 88x_2$$

subject to

$$\begin{aligned} 2x_1 + 8x_2 &\leqq 60 \\ 5x_1 + 2x_2 &\leqq 60 \\ x_1 \qquad &\geqq 0 \\ x_2 &\geqq 0. \end{aligned}$$

Converting the first two inequalities to equations by introducing two slack variables x_3, x_4, we obtained the **normal form** of the problem in Example 2. Together with the objective function (written as an equation $z - 40x_1 - 88x_2 = 0$) this normal form is

$$\begin{aligned} z - 40x_1 - 88x_2 \qquad\qquad &= 0 \\ 2x_1 + 8x_2 + x_3 \qquad &= 60 \\ 5x_1 + 2x_2 \qquad + x_4 &= 60 \end{aligned} \tag{3}$$

where $x_1 \geqq 0, \cdots, x_4 \geqq 0$. This is a linear system of equations. To find an optimal solution of it, we may consider its **augmented matrix** (see Sec. 7.3)

$$\mathbf{T}_0 = \begin{array}{c} \begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\ \left[\begin{array}{c|cc|cc|c} 1 & -40 & -88 & 0 & 0 & 0 \\ \hline 0 & 2 & 8 & 1 & 0 & 60 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] \end{array} \tag{4}$$

[1]GEORGE BERNARD DANTZIG (1914–2005), American mathematician, who is one of the pioneers of linear programming and inventor of the simplex method. According to Dantzig himself (see G. B. Dantzig, Linear programming: The story of how it began, in J. K. Lenestra et al., *History of Mathematical Programming: A Collection of Personal Reminiscences*. Amsterdam: Elsevier, 1991, pp. 19–31), he was particularly fascinated by Wassilly Leontief's input–output model (Sec. 8.2) and invented his famous method to solve large-scale planning (logistics) problems. Besides Leontief, Dantzig credits others for their pioneering work in linear programming, that is, JOHN VON NEUMANN (1903–1957), Hungarian American mathematician, Institute for Advanced Studies, Princeton University, who made major contributions to game theory, computer science, functional analysis, set theory, quantum mechanics, ergodic theory, and other areas, the Nobel laureates LEONID VITALIYEVICH KANTOROVICH (1912–1986), Russian economist, and TJALLING CHARLES KOOPMANS (1910–1985), Dutch–American economist, who shared the 1975 Nobel Prize in Economics for their contributions to the theory of optimal allocation of resources. Dantzig was a driving force in establishing the field of linear programming and became professor of transportation sciences, operations research, and computer science at Stanford University. For his work see R. W. Cottle (ed.), *The Basic George B. Dantzig.* Palo Alto, CA: Stanford University Press, 2003.

This matrix is called a **simplex tableau** or **simplex table** (the ***initial*** *simplex table*). These are standard names. The dashed lines and the letters

$$z, \quad x_1, \quad \cdots, \quad b$$

are for ease in further manipulation.

Every simplex table contains two kinds of variables x_j. By **basic variables** we mean those whose columns have only one nonzero entry. Thus x_3, x_4 in (4) are basic variables and x_1, x_2 are **nonbasic variables**.

Every simplex table gives a basic feasible solution. It is obtained by setting the nonbasic variables to zero. Thus (4) gives the basic feasible solution

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 60/1 = 60, \qquad x_4 = 60/1 = 60, \qquad z = 0$$

with x_3 obtained from the second row and x_4 from the third.

The optimal solution (its location and value) is now obtained stepwise by pivoting, designed to take us to basic feasible solutions with higher and higher values of z until the maximum of z is reached. Here, the choice of the **pivot equation** and **pivot** are quite different from that in the Gauss elimination. The reason is that x_1, x_2, x_3, x_4 are restricted to nonnegative values.

***Step 1. Operation O_1*: Selection of the *Column* of the Pivot**

Select as the column of the pivot the first column with a negative entry in Row 1. In (4) this is Column 2 (because of the -40).

Operation O_2*: Selection of the *Row* of the Pivot.** Divide the right sides [60 and 60 in (4)] by the corresponding entries of the column just selected ($60/2 = 30$, $60/5 = 12$). Take as the pivot equation the equation that gives the ***smallest quotient. Thus the pivot is 5 because $60/5$ is smallest.

***Operation O_3*: Elimination by Row Operations.** This gives zeros above and below the pivot (as in Gauss–Jordan, Sec. 7.8).

With the notation for row operations as introduced in Sec. 7.3, the calculations in Step 1 give from the simplex table $\mathbf{T}_0$ in (4) the following simplex table (augmented matrix), with the blue letters referring to the ***previous table***.

$$\begin{array}{ccc}
 & \begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} & \\
(5) \qquad \mathbf{T}_1 = & \left[\begin{array}{c|cc|cc|c} 1 & 0 & -72 & 0 & 8 & 480 \\ \hline 0 & 0 & 7.2 & 1 & -0.4 & 36 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] & \begin{array}{l} \text{Row 1} + 8 \text{ Row 3} \\ \text{Row 2} - 0.4 \text{ Row 3} \\ \end{array}
\end{array}$$

We see that basic variables are now x_1, x_3 and nonbasic variables are x_2, x_4. Setting the latter to zero, we obtain the basic feasible solution given by $\mathbf{T}_1$,

$$x_1 = 60/5 = 12, \qquad x_2 = 0, \qquad x_3 = 36/1 = 36, \qquad x_4 = 0, \qquad z = 480.$$

This is A in Fig. 474 (Sec. 22.2). We thus have moved from O: $(0, 0)$ with $z = 0$ to A: $(12, 0)$ with the greater $z = 480$. The reason for this increase is our elimination of a term $(-40x_1)$ with a negative coefficient. Hence ***elimination is applied only to negative entries*** in Row 1 but to no others. This motivates the selection of the *column* of the pivot.

We now motivate the selection of the *row* of the pivot. Had we taken the second row of $\mathbf{T}_0$ instead (thus 2 as the pivot), we would have obtained $z = 1200$ (verify!), but this line of constant revenue $z = 1200$ lies entirely outside the feasibility region in Fig. 474. This motivates our cautious choice of the entry 5 as our pivot because it gave the smallest quotient $(60/5 = 12)$.

Step 2. The basic feasible solution given by (5) is not yet optimal because of the negative entry -72 in Row 1. Accordingly, we perform the operations O_1 to O_3 again, choosing a pivot in the column of -72.

Operation O_1. Select Column 3 of $\mathbf{T}_1$ in (5) as the column of the pivot (because $-72 < 0$).

Operation O_2. We have $36/7.2 = 5$ and $60/2 = 30$. Select 7.2 as the pivot (because $5 < 30$).

Operation O_3. Elimination by row operations gives

$$
(6) \qquad \mathbf{T}_2 = \begin{array}{c} \begin{matrix} z & x_1 & x_2 & x_3 & x_4 & b \end{matrix} \\ \left[\begin{array}{c|cc|cc|c} 1 & 0 & 0 & 10 & 4 & 840 \\ \hline 0 & 0 & 7.2 & 1 & -0.4 & 36 \\ 0 & 5 & 0 & -\dfrac{1}{3.6} & \dfrac{1}{0.9} & 50 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1} + 10 \text{ Row 2} \\ \\ \text{Row 3} - \dfrac{2}{7.2} \text{ Row 2} \end{array}
$$

We see that now x_1, x_2 are basic and x_3, x_4 nonbasic. Setting the latter to zero, we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 50/5 = 10, \qquad x_2 = 36/7.2 = 5, \qquad x_3 = 0, \qquad x_4 = 0, \qquad z = 840.$$

This is B in Fig. 474 (Sec. 22.2). In this step, z has increased from 480 to 840, due to the elimination of -72 in $\mathbf{T}_1$. Since $\mathbf{T}_2$ contains no more negative entries in Row 1, we conclude that $z = f(10, 5) = 40 \cdot 10 + 88 \cdot 5 = 840$ is the maximum possible revenue. It is obtained if we produce twice as many S heaters as L heaters. This is the solution of our problem by the simplex method of linear programming. ■

Minimization. If we want to ***minimize*** $z = f(\mathbf{x})$ (instead of maximize), we take as the columns of the pivots those whose entry in Row 1 is ***positive*** (instead of negative). In such a Column k we consider only positive entries t_{jk} and take as pivot a t_{jk} for which b_j/t_{jk} is smallest (as before). For examples, see the problem set.

PROBLEM SET 22.3

1. Verify the calculations in Example 1 of the text.

2–14 SIMPLEX METHOD

Write in normal form and solve by the simplex method, assuming all x_j to be nonnegative.

2. The problem in the example in the text with the constraints interchanged.

3. Maximize $f = 3x_1 + 2x_2$ subject to $3x_1 + 4x_2 \leqq 60$, $4x_1 + 3x_2 \leqq 60$, $10x_1 + 2x_2 \leqq 120$.

4. Maximize the daily output in producing x_1 chairs by Process P_1 and x_2 chairs by Process P_2 subject to $3x_1 + 4x_2 \leqq 550$ (machine hours), $5x_1 + 4x_2 \leqq 650$ (labor).

5. Minimize $f = 5x_1 - 20x_2$ subject to $-2x_1 + 10x_2 \leqq 5$, $2x_1 + 5x_2 \leqq 10$.

6. Prob. 19 in Sec. 22.2.

7. Suppose we produce x_1 AA batteries by Process P_1 and x_2 by Process P_2, furthermore x_3 A batteries by Process P_3 and x_4 by Process P_4. Let the profit for 100 batteries be \$10 for AA and \$20 for A. Maximize the total profit subject to the constraints

$$12x_1 + 8x_2 + 6x_3 + 4x_4 \leqq 120 \quad \text{(Material)}$$
$$3x_1 + 6x_2 + 12x_3 + 24x_4 \leqq 180 \quad \text{(Labor)}.$$

8. Maximize the daily profit in producing x_1 metal frames F_1 (profit \$90 per frame) and x_2 frames F_2 (profit \$50 per frame) subject to $x_1 + 3x_2 \leqq 18$ (material), $x_1 + x_2 \leqq 10$ (machine hours), $3x_1 + x_2 \leqq 24$ (labor).

9. Maximize $f = 2x_1 + x_2 + 3x_3$ subject to $4x_1 + 3x_2 + 6x_3 = 12$.

10. Minimize $f = 4x_1 - 10x_2 - 20x_3$ subject to $3x_1 + 4x_2 + 5x_3 \leqq 60$, $2x_1 + x_2 \leqq 20$, $2x_1 + 3x_3 \leqq 30$.

11. Prob. 22 in Problem Set 22.2.

12. Maximize $f = 2x_1 + 3x_2 + x_3$ subject to $x_1 + x_2 + x_3 \leqq 4.8$, $10x_1 + x_3 \leqq 9.9$, $x_2 - x_3 \leqq 0.2$.

13. Maximize $f = 34x_1 + 29x_2 + 32x_3$ subject to $8x_1 + 2x_2 + x_3 \leqq 54$, $3x_1 + 8x_2 + 2x_3 \leqq 59$, $x_1 + x_2 + 5x_3 \leqq 39$.

14. Maximize $f = 2x_1 + 3x_2$ subject to $5x_1 + 3x_2 \leqq 105$, $3x_1 + 6x_2 \leqq 126$.

15. CAS PROJECT. Simple Method. (a) Write a program for graphing a region R in the first quadrant of the x_1x_2-plane determined by linear constraints.

(b) Write a program for maximizing $z = a_1x_1 + a_2x_2$ in R.

(c) Write a program for maximizing $z = a_1x_1 + \cdots + a_nx_n$ subject to linear constraints.

(d) Apply your programs to problems in this problem set and the previous one.

22.4 Simplex Method: Difficulties

In solving a linear optimization problem by the simplex method, we proceed stepwise from one basic feasible solution to another. By so doing, we increase the value of the objective function f. We continue this stepwise procedure, until we reach an optimal solution. This was all explained in Sec. 22.3. However, the method does not always proceed so smoothly. Occasionally, but rather infrequently in practice, we encounter two kinds of difficulties. The first one is the degeneracy and the second one concerns difficulties in starting.

Degeneracy

A **degenerate feasible solution** is a feasible solution at which more than the usual number $n - m$ of variables are zero. Here n is the number of variables (slack and others) and m the number of constraints (not counting the $x_j \geqq 0$ conditions). In the last section, $n = 4$ and $m = 2$, and the occurring basic feasible solutions were nondegenerate; $n - m = 2$ variables were zero in each such solution.

In the case of a degenerate feasible solution we do an extra elimination step in which a basic variable that is zero for that solution becomes nonbasic (and a nonbasic variable becomes basic instead). We explain this in a typical case. For more complicated cases and techniques (rarely needed in practice) see Ref. [F5] in App. 1.

EXAMPLE 1 Simplex Method, Degenerate Feasible Solution

AB Steel, Inc., produces two kinds of iron I_1, I_2 by using three kinds of raw material R_1, R_2, R_3 (scrap iron and two kinds of ore) as shown. Maximize the daily profit.

Raw Material	Raw Material Needed per Ton		Raw Material Available per Day (tons)
	Iron I_1	Iron I_2	
R_1	2	1	16
R_2	1	1	8
R_3	0	1	3.5
Net profit per ton	\$150	\$300	

Solution. Let x_1 and x_2 denote the amount (in tons) of iron I_1 and I_2, respectively, produced per day. Then our problem is as follows. Maximize

$$z = f(x) = 150x_1 + 300x_2 \tag{1}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and

$$\begin{aligned} 2x_1 + x_2 &\leqq 16 \qquad \text{(raw material } R_1\text{)} \\ x_1 + x_2 &\leqq \ \ 8 \qquad \text{(raw material } R_2\text{)} \\ x_2 &\leqq \ \ 3.5 \quad \text{(raw material } R_3\text{)}. \end{aligned}$$

By introducing slack variables x_3, x_4, x_5 we obtain the normal form of the constraints

$$\begin{aligned} 2x_1 + x_2 + x_3 \qquad\qquad &= 16 \\ x_1 + x_2 \qquad + x_4 \qquad &= \ \ 8 \\ x_2 \qquad\qquad + x_5 &= \ \ 3.5 \\ x_i \geqq 0 \qquad (i = 1, \cdots, 5). \end{aligned} \tag{2}$$

As in the last section we obtain from (1) and (2) the initial simplex table

$$\mathbf{T}_0 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|ccc|c} 1 & -150 & -300 & 0 & 0 & 0 & 0 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 1 & 1 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array}\right] \end{array}. \tag{3}$$

We see that x_1, x_2 are nonbasic variables and x_3, x_4, x_5 are basic. With $x_1 = x_2 = 0$ we have from (3) the basic feasible solution

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 16/1 = 16, \quad x_4 = 8/1 = 8, \quad x_5 = 3.5/1 = 3.5, \quad z = 0.$$

This is O: (0, 0) in Fig. 475. We have $n = 5$ variables x_j, $m = 3$ constraints, and $n - m = 2$ variables equal to zero in our solution, which thus is nondegenerate.

Step 1 of Pivoting

***Operation O_1*:** Column Selection of Pivot. Column 2 (since $-150 < 0$).

***Operation O_2*:** Row Selection of Pivot. $16/2 = 8$, $8/1 = 8$; $3.5/0$ is not possible. Hence we could choose Row 2 or Row 3. We choose Row 2. The pivot is 2.

***Operation O_3*:** Elimination by Row Operations. This gives the simplex table

$$\text{(4)}\qquad \mathbf{T}_1 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \hline \left[\begin{array}{c|cc|ccc|c} 1 & 0 & -225 & 75 & 0 & 0 & 1200 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1} + 75 \text{ Row 2} \\ \\ \text{Row 3} - \frac{1}{2}\text{ Row 2} \\ \text{Row 4} \end{array}$$

We see that the basic variables are x_1, x_4, x_5 and the nonbasic are x_2, x_3. Setting the nonbasic variables to zero, we obtain from $\mathbf{T}_1$ the basic feasible solution

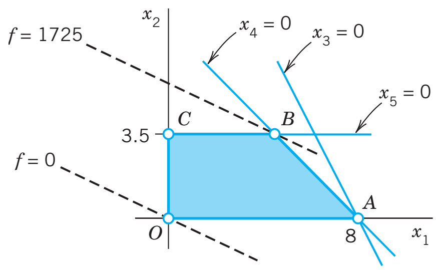


Fig. 475. Example 1, where A is degenerate

$$x_1 = 16/2 = 8, \qquad x_2 = 0, \qquad x_3 = 0, \qquad x_4 = 0/1 = 0, \qquad x_5 = 3.5/1 = 3.5, \qquad z = 1200.$$

This is A: $(8, 0)$ in Fig. 475. This solution in degenerate because $x_4 = 0$ (in addition to $x_2 = 0, x_3 = 0$); geometrically: the straight line $x_4 = 0$ also passes through A. This requires the next step, in which x_4 will become nonbasic.

Step 2 of Pivoting

***Operation O_1*:** Column Selection of Pivot. Column 3 (since $-225 < 0$).

***Operation O_2*:** Row Selection of Pivot. $16/1 = 16, 0/\frac{1}{2} = 0$. Hence $\frac{1}{2}$ must serve as the pivot.

***Operation O_3*:** Elimination by Row Operations. This gives the following simplex table.

$$\text{(5)}\qquad \mathbf{T}_2 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \hline \left[\begin{array}{c|cc|ccc|c} 1 & 0 & 0 & -150 & 450 & 0 & 1200 \\ \hline 0 & 2 & 0 & 2 & -2 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1} + 450 \text{ Row 3} \\ \text{Row 2} - 2 \text{ Row 3} \\ \\ \text{Row 4} - 2\text{ Row 3} \end{array}$$

We see that the basic variables are x_1, x_2, x_5 and the nonbasic are x_3, x_4. Hence x_4 has become nonbasic, as intended. By equating the nonbasic variables to zero we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 16/2 = 8, \qquad x_2 = 0/\tfrac{1}{2} = 0, \qquad x_3 = 0, \qquad x_4 = 0, \qquad x_5 = 3.5/1 = 3.5, \qquad z = 1200.$$

This is still A: $(8, 0)$ in Fig. 475 and z has not increased. But this opens the way to the maximum, which we reach in the next step.

Step 3 of Pivoting

Operation O_1: Column Selection of Pivot. Column 4 (since $-150 < 0$).

Operation O_2: Row Selection of Pivot. $16/2 = 8$, $0/(-\frac{1}{2}) = 0$, $3.5/1 = 3.5$. We can take 1 as the pivot. (With $-\frac{1}{2}$ as the pivot we would not leave A. Try it.)

Operation O_3: Elimination by Row Operations. This gives the simplex table

$$
\begin{array}{c} \\ \\ \end{array}
(6) \qquad \mathbf{T}_3 = \begin{array}{c} \begin{array}{cccccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|ccc|c} 1 & 0 & 0 & 0 & 150 & 150 & 1725 \\ \hline 0 & 2 & 0 & 0 & 2 & -2 & 9 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 1.75 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \\ \text{Row 1 + 150 Row 4} \\ \text{Row 2 − 2 Row 4} \\ \text{Row 3 + } \tfrac{1}{2} \text{ Row 4} \\ \\ \end{array}
$$

We see that basic variables are x_1, x_2, x_3 and nonbasic x_4, x_5. Equating the latter to zero we obtain from $\mathbf{T}_3$ the basic feasible solution

$$x_1 = 9/2 = 4.5, \qquad x_2 = 1.75/\tfrac{1}{2} = 3.5, \qquad x_3 = 3.5/1 = 3.5, \qquad x_4 = 0, \qquad x_5 = 0, \qquad z = 1725.$$

This is B: (4.5, 3.5) in Fig. 475. Since Row 1 of $\mathbf{T}_3$ has no negative entries, we have reached the maximum daily profit $z_{\max} = f(4.5, 3.5) = 150 \cdot 4.5 + 300 \cdot 3.5 = \$1725$. This is obtained by using 4.5 tons of iron I_1 and 3.5 tons of iron I_2. ■

Difficulties in Starting

As a second kind of difficulty, it may sometimes be hard to find a basic feasible solution to start from. In such a case the idea of an **artificial variable** (or several such variables) is helpful. We explain this method in terms of a typical example.

EXAMPLE 2 Simplex Method: Difficult Start, Artificial Variable

Maximize

$$z = f(\mathbf{x}) = 2x_1 + x_2 \tag{7}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and (Fig. 476)

$$
\begin{aligned}
x_1 - \tfrac{1}{2}x_2 &\geqq 1 \\
x_1 - \phantom{\tfrac{1}{2}}x_2 &\leqq 2 \\
x_1 + \phantom{\tfrac{1}{2}}x_2 &\leqq 4.
\end{aligned}
$$

Solution. By means of slack variables we achieve the normal form of the constraints

$$
\begin{aligned}
z - 2x_1 - x_2 \qquad\qquad\qquad &= 0 \\
x_1 - \tfrac{1}{2}x_2 - x_3 \qquad\qquad &= 1 \\
x_1 - x_2 \qquad + x_4 \qquad &= 2 \\
x_1 + x_2 \qquad\qquad + x_5 &= 4 \\
x_i \geqq 0 \quad (i = 1, \cdots, 5).&
\end{aligned} \tag{8}
$$

Note that the first slack variable is negative (or zero), which makes x_3 nonnegative within the feasibility region (and negative outside). From (7) and (8) we obtain the simplex table

$$\begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array}$$
$$\left[\begin{array}{c|cc|ccc|c} 1 & -2 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 0 & 0 & 1 & 4 \end{array}\right].$$

x_1, x_2 are nonbasic, and we would like to take x_3, x_4, x_5 as basic variables. By our usual process of equating the nonbasic variables to zero we obtain from this table

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 1/(-1) = -1, \qquad x_4 = \tfrac{2}{1} = 2, \qquad x_5 = \tfrac{4}{1} = 4, \qquad z = 0.$$

$x_3 < 0$ indicates that $(0, 0)$ lies outside the feasibility region. Since $x_3 < 0$, we cannot proceed immediately. Now, instead of searching for other basic variables, we use the following idea. Solving the second equation in (8) for x_3, we have

$$x_3 = -1 + x_1 - \tfrac{1}{2}x_2.$$

To this we now add a variable x_6 on the right,

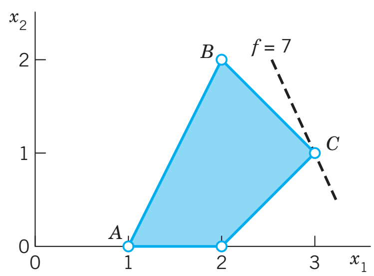


Fig. 476. Feasibility region in Example 2

$$x_3 = -1 + x_1 - \tfrac{1}{2}x_2 + x_6. \tag{9}$$

x_6 is called an **artificial variable** and is subject to the constraint $x_6 \geqq 0$.

We must take care that x_6 (which is not part of the given problem!) will disappear eventually. We shall see that we can accomplish this by adding a term $-Mx_6$ with very large M to the objective function. Because of (7) and (9) (solved for x_6) this gives the modified objective function for this "**extended problem**"

$$\hat{z} = z - Mx_6 = 2x_1 + x_2 - Mx_6 = (2 + M)x_1 + (1 - \tfrac{1}{2}M)x_2 - Mx_3 - M. \tag{10}$$

We see that the simplex table corresponding to (10) and (8) is

$$\begin{array}{cccccccc} \hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \end{array}$$
$$\mathbf{T}_0 = \left[\begin{array}{c|cc|cccc|c} 1 & -2 - M & -1 + \frac{1}{2}M & M & 0 & 0 & 0 & -M \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 1 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 & 1 \end{array}\right].$$

The last row of this table results from (9) written as $x_1 - \frac{1}{2}x_2 - x_3 + x_6 = 1$. We see that we can now start, taking x_4, x_5, x_6 as the basic variables and x_1, x_2, x_3 as the nonbasic variables. Column 2 has a negative first entry. We can take the second entry (1 in Row 2) as the pivot. This gives

$$\mathbf{T}_1 = \begin{array}{c} \begin{array}{cccccccc} \hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \end{array} \\ \left[\begin{array}{c|cc|cccc|c} 1 & 0 & -2 & -2 & 0 & 0 & 0 & 2 \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}.$$

This corresponds to $x_1 = 1, x_2 = 0$ (point A in Fig. 476), $x_3 = 0, x_4 = 1, x_5 = 3, x_6 = 0$. We can now drop Row 5 and Column 7. In this way we get rid of x_6, as wanted, and obtain

$$\mathbf{T}_2 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|ccc|c} 1 & 0 & -2 & -2 & 0 & 0 & 2 \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 1 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3 \end{array}\right] \end{array}.$$

In Column 3 we choose $\frac{3}{2}$ as the next pivot. We obtain

$$\mathbf{T}_3 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|ccc|c} 1 & 0 & 0 & -\frac{2}{3} & 0 & \frac{4}{3} & 6 \\ \hline 0 & 1 & 0 & -\frac{2}{3} & 0 & \frac{1}{3} & 2 \\ 0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3 \end{array}\right] \end{array}.$$

This corresponds to $x_1 = 2, x_2 = 2$ (this is B in Fig. 476), $x_3 = 0, x_4 = 2, x_5 = 0$. In Column 4 we choose $\frac{4}{3}$ as the pivot, by the usual principle. This gives

$$\mathbf{T}_4 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|ccc|c} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{3}{2} & 7 \\ \hline 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 3 \\ 0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\ 0 & 0 & \frac{3}{2} & 0 & -\frac{3}{4} & \frac{3}{4} & \frac{3}{2} \end{array}\right] \end{array}.$$

This corresponds to $x_1 = 3, x_2 = 1$ (point C in Fig. 476), $x_3 = \frac{3}{2}, x_4 = 0, x_5 = 0$. This is the maximum $f_{\max} = f(3, 1) = 7$. ■

We have reached the end of our discussion on linear programming. We have presented the simplex method in great detail as this method has many beautiful applications and works well on most practical problems. Indeed, problems of optimization appear in civil engineering, chemical engineering, environmental engineering, management science, logistics, strategic planning, operations management, industrial engineering, finance, and other areas. Furthermore, the simplex method allows your problem to be *scaled up* from a small modeling attempt to a larger modeling attempt, by adding more constraints and variables, thereby making your model more realistic. The area of optimization is an active field of development and research and optimization methods, besides the simplex method, are being explored and experimented with.

PROBLEM SET 22.4

1. Maximize $z = f_1(\mathbf{x}) = 7x_1 + 14x_2$ subject to $0 \leqq x_1 \leqq 6$, $0 \leqq x_2 \leqq 3$, $7x_1 + 14x_2 \leqq 84$.

2. Do Prob. 1 with the last two constraints interchanged.

3. Maximize the daily output in producing x_1 steel sheets by process P_A and x_2 steel sheets by process P_B subject to the constraints of labor hours, machine hours, and raw material supply:

$$3x_1 + 2x_2 \leqq 180, \qquad 4x_1 + 6x_2 \leqq 200,$$
$$5x_1 + 3x_2 \leqq 160.$$

4. Maximize $z = 300x_1 + 500x_2$ subject to $2x_1 + 8x_2 \leqq 60$, $2x_1 + x_2 \leqq 30$, $4x_1 + 4x_2 \leqq 60$.

5. Do Prob. 4 with the last two constraints interchanged. Comment on the resulting simplification.

6. Maximize the total output $f = x_1 + x_2 + x_3$ (production from three distinct processes) subject to input constraints (limitation of time available for production)

$$5x_1 + 6x_2 + 7x_3 \leqq 12,$$
$$7x_1 + 4x_2 + x_3 \leqq 12.$$

7. Maximize $f = 5x_1 + 8x_2 + 4x_3$ subject to $x_j \geqq 0$ $(j = 1, \cdots, 5)$ and $x_1 + x_3 + x_5 = 1$, $x_2 + x_3 + x_4 = 1$.

8. Using an artificial variable, minimize $f = 4x_1 - x_2$ subject to $x_1 + x_2 \geqq 2$, $-2x_1 + 3x_2 \leqq 1$, $5x_1 + 4x_2 \leqq 50$.

9. Maximize $f = 2x_1 + 3x_2 + 2x_3$, $x_1 \geqq 0$, $x_2 \geqq 0$, $x_3 \geqq 0$, $x_1 + 2x_2 - 4x_3 \leqq 2$, $x_1 + 2x_2 + 2x_3 \leqq 5$.

CHAPTER 22 REVIEW QUESTIONS AND PROBLEMS

1. What is unconstrained optimization? Constraint optimization? To which one do methods of calculus apply?

2. State the idea and the formulas of the method of steepest descent.

3. Write down an algorithm for the method of steepest descent.

4. Design a "method of steepest ascent" for determining maxima.

5. What is the method of steepest descent for a function of a single variable?

6. What is the basic idea of linear programming?

7. What is an objective function? A feasible solution?

8. What are slack variables? Why did we introduce them?

9. What happens in Example 1 of Sec. 22.1 if you replace $f(\mathbf{x}) = x_1^2 + 3x_2^2$ with $f(\mathbf{x}) = x_1^2 + 5x_2^2$? Start from $\mathbf{x}_0 = [6 \quad 3]^\mathsf{T}$. Do 5 steps. Is the convergence faster or slower?

10. Apply the method of steepest descent to $f(\mathbf{x}) = 9x_1^2 + x_2^2 + 18x_1 - 4x_2$, 5 steps. Start from $\mathbf{x}_0 = [2 \quad 4]^\mathsf{T}$.

11. In Prob. 10, could you start from $[0 \quad 0]^\mathsf{T}$ and do 5 steps?

12. Show that the gradients in Prob. 11 are orthogonal. Give a reason.

13–16 Graph or sketch the region in the first quadrant of the x_1x_2-plane determined by the following inequalities.

13.
$$\begin{aligned} x_1 - 2x_2 &\leqq -2 \\ 0.8x_1 + x_2 &\leqq 6 \end{aligned}$$

14.
$$\begin{aligned} x_1 - 2x_2 &\geqq -4 \\ 2x_1 + x_2 &\leqq 12 \\ x_1 + x_2 &\leqq 8 \end{aligned}$$

15.
$$\begin{aligned} x_1 + x_2 &\leqq 5 \\ x_2 &\leqq 3 \\ -x_1 + x_2 &\leqq 2 \end{aligned}$$

16.
$$\begin{aligned} x_1 + x_2 &\geqq 2 \\ 2x_1 - 3x_2 &\geqq -12 \\ x_1 &\leqq 15 \end{aligned}$$

17–20 Maximize or minimize as indicated.

17. Maximize $f = 10x_1 + 20x_2$ subject to $x_1 \leqq 5$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 4$.

18. Maximize $f = x_1 + x_2$ subject to $x_1 + 2x_2 \leqq 10$, $2x_2 + x_2 \leqq 10$, $x_2 \leqq 4$.

19. Minimize $f = 2x_1 - 10x_2$ subject to $x_1 - x_2 \leqq 4$, $2x_1 + x_2 \leqq 14$, $x_1 + x_2 \leqq 9$, $-x_1 + 3x_2 \leqq 15$.

20. A factory produces two kinds of gaskets, G_1, G_2, with net profit of \$60 and \$30, respectively, Maximize the total daily profit subject to the constraints (x_j = number of gaskets G_j produced per day):

$$\begin{aligned} 40x_1 + 40x_2 &\leqq 1800 \quad \text{(Machine hours)}, \\ 200x_1 + 20x_2 &\leqq 6300 \quad \text{(Labor)}. \end{aligned}$$

SUMMARY OF CHAPTER 22
Unconstrained Optimization. Linear Programming

In optimization problems we maximize or minimize an ***objective function*** $z = f(\mathbf{x})$ depending on control variables $x_1, \cdots, x_m$ whose domain is either unrestricted ("***unconstrained optimization***," Sec. 22.1) or restricted by constraints in the form of inequalities or equations or both ("***constrained optimization***," Sec. 22.2).

If the objective function is *linear* and the constraints are *linear inequalities* in $x_1, \cdots, x_m$, then by introducing **slack variables** $x_{m+1}, \cdots, x_n$ we can write the optimization problem in **normal form** with the objective function given by

$$f_1 = c_1 x_1 + \cdots + c_n x_n \tag{1}$$

(where $c_{m+1} = \cdots = c_n = 0$) and the constraints given by

$$\begin{gathered}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\
x_1 \geqq 0, \cdots, x_n \geqq 0.
\end{gathered} \tag{2}$$

In this case we can then apply the widely used ***simplex method*** (Sec. 22.3), a systematic stepwise search through a very much reduced subset of all feasible solutions. Section 22.4 shows how to overcome difficulties with this method.

CHAPTER 23

Graphs. Combinatorial Optimization

Many problems in electrical engineering, civil engineering, operations research, industrial engineering, management, logistics, marketing, and economics can be modeled by ***graphs*** and directed ***graphs***, called digraphs. This is not surprising as they allow us to model networks, such as roads and cables, where the nodes may be cities or computers. The task then is to find the shortest path through the network or the best way to connect computers. Indeed, many researchers who made contributions to combinatorial optimization and graphs, and whose names lend themselves to fundamental algorithms in this chapter, such as Fulkerson, Kruskal, Moore, and Prim, all worked at Bell Laboratories in New Jersey, the major R&D facilities of the huge telephone and telecommunication company AT&T. As such, they were interested in methods of optimally building computer networks and telephone networks. The field has progressed into looking for more and more efficient algorithms for very large problems.

Combinatorial optimization deals with optimization problems that are of a pronounced discrete or combinatorial nature. Often the problems are very large and so a direct search may not be possible. Just like in linear programming (Chap. 22), the computer is an indispensible tool and makes solving large-scale modeling problems possible. Because the area has a distinct flavor, different from ODEs, linear algebra, and other areas, we start with the basics and gradually introduce algorithms for shortest path problems (Secs. 22.2, 22.3), shortest spanning trees (Secs. 23.4, 23.5), flow problems in networks (Secs. 23.6, 23.7), and assignment problems (Sec. 23.8).

Prerequisite: none.
References and Answers to Problems: App. 1 Part F, App. 2.

23.1 Graphs and Digraphs

Roughly, a *graph* consists of points, called *vertices*, and lines connecting them, called *edges*. For example, these may be four cities and five highways connecting them, as in Fig. 477. Or the points may represent some people, and we connect by an edge those who do business with each other. Or the vertices may represent computers in a network and the edge connections between them. Let us now give a formal definition.

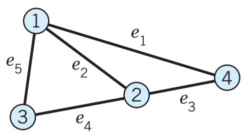


Fig. 477. Graph consisting of 4 vertices and 5 edges

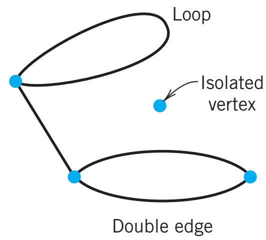


Fig. 478. Isolated vertex, loop, double edge. (Excluded by definition.)

DEFINITION

Graph

A **graph** G consists of two finite sets (sets having finitely many elements), a set V of points, called **vertices**, and a set E of connecting lines, called **edges**, such that each edge connects two vertices, called the *endpoints* of the edge. We write

$$G = (V, E).$$

Excluded are *isolated vertices* (vertices that are not endpoints of any edge), *loops* (edges whose endpoints coincide), and *multiple edges* (edges that have both endpoints in common). See Fig. 478.

CAUTION! Our three exclusions are practical and widely accepted, but not uniformly. For instance, some authors permit multiple edges and call graphs without them *simple graphs.* ■

We denote vertices by letters, $u, v, \cdots$ or $v_1, v_2, \cdots$ or simply by numbers $1, 2, \cdots$ (as in Fig. 477). We denote edges by $e_1, e_2, \cdots$ or by their two endpoints; for instance, $e_1 = (1, 4)$, $e_2 = (1, 2)$ in Fig. 477.

An edge (v_i, v_j) is called **incident** with the vertex v_i (and conversely); similarly, (v_i, v_j) is *incident* with v_j. The number of edges incident with a vertex v is called the **degree** of v. Two vertices are called **adjacent** in G if they are connected by an edge in G (that is, if they are the two endpoints of some edge in G).

We meet graphs in different fields under different names: as "networks" in electrical engineering, "structures" in civil engineering, "molecular structures" in chemistry, "organizational structures" in economics, "sociograms," "road maps," "telecommunication networks," and so on.

Digraphs (Directed Graphs)

Nets of one-way streets, pipeline networks, sequences of jobs in construction work, flows of computation in a computer, producer–consumer relations, and many other applications suggest the idea of a "digraph" (= directed graph), in which each edge has a direction (indicated by an arrow, as in Fig. 479).

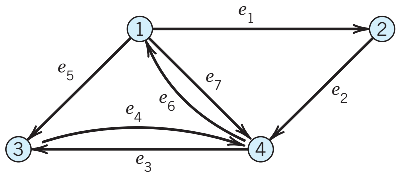


Fig. 479. Digraph

DEFINITION

Digraph (Directed Graph)

A **digraph** $G = (V, E)$ is a graph in which each edge $e = (i, j)$ has a direction from its "*initial point*" i to its "*terminal point*" j.

Two edges connecting the same two points i, j are now permitted, provided they have opposite directions, that is, they are (i, j) and (j, i). *Example.* (1, 4) and (4, 1) in Fig. 479.

A **subgraph** or subdigraph of a given graph or digraph $G = (V, E)$, respectively, is a graph or digraph obtained by deleting some of the edges and vertices of G, retaining the other edges of G (together with their pairs of endpoints). For instance, e_1, e_3 (together with the vertices 1, 2, 4) form a subgraph in Fig. 477, and e_3, e_4, e_5 (together with the vertices 1, 3, 4) form a subdigraph in Fig. 479.

Computer Representation of Graphs and Digraphs

Drawings of graphs are useful to people in explaining or illustrating specific situations. Here one should be aware that a graph may be sketched in various ways; see Fig. 480. For handling graphs and digraphs in computers, one uses matrices or lists as appropriate data structures, as follows.

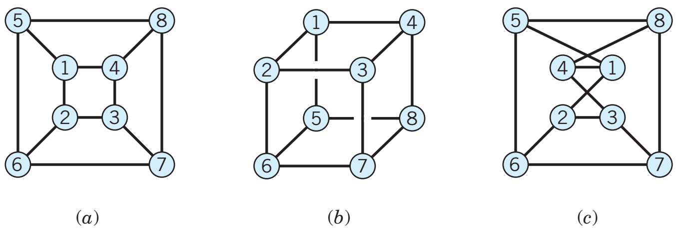


Fig. 480. Different sketches of the same graph

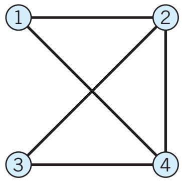
Adjacency Matrix of a Graph G: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has an edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

Thus $a_{ij} = 1$ if and only if two vertices i and j are adjacent in G. Here, by definition, no vertex is considered to be adjacent to itself; thus, $a_{ii} = 0$. $\mathbf{A}$ is symmetric, $a_{ij} = a_{ji}$. (Why?)

The adjacency matrix of a graph is generally much smaller than the so-called *incidence matrix* (see Prob. 18) and is preferred over the latter if one decides to store a graph in a computer in matrix form.

EXAMPLE 1 Adjacency Matrix of a Graph



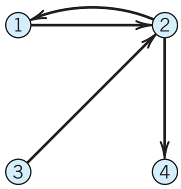
$$\begin{array}{r} \text{Vertex} \\ \text{Vertex } 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Adjacency Matrix of a Digraph G: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has a directed edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

This matrix $\mathbf{A}$ need not be symmetric. (Why?)

EXAMPLE 2 Adjacency Matrix of a Digraph

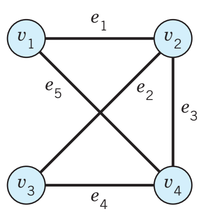


$$\begin{array}{r} \text{To vertex} \\ \text{From vertex } 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Lists. The **vertex incidence list** of a graph shows, for each vertex, the incident edges. The **edge incidence list** shows for each edge its two endpoints. Similarly for a *digraph*; in the vertex list, outgoing edges then get a minus sign, and in the edge list we now have *ordered* pairs of vertices.

EXAMPLE 3 Vertex Incidence List and Edge Incidence List of a Graph

This graph is the same as in Example 1, except for notation.



Vertex	Incident Edges
v_1	e_1, e_5
v_2	e_1, e_2, e_3
v_3	e_2, e_4
v_4	e_3, e_4, e_5

Edge	Endpoints
e_1	v_1, v_2
e_2	v_2, v_3
e_3	v_2, v_4
e_4	v_3, v_4
e_5	v_1, v_4

Sparse graphs are graphs with few edges (far fewer than the maximum possible number $n(n-1)/2$, where n is the number of vertices). For these graphs, matrices are not efficient. *Lists* then have the advantage of requiring much less storage and being easier to handle; they can be ordered, sorted, or manipulated in various other ways directly within the computer. For instance, in tracing a "walk" (a connected sequence of edges with pairwise common endpoints), one can easily go back and forth between the two lists just discussed, instead of scanning a large column of a matrix for a single 1.

Computer science has developed more refined lists, which, in addition to the actual content, contain "pointers" indicating the preceding item or the next item to be scanned or both items (in the case of a "walk": the preceding edge or the subsequent one). For details, see Refs. [E16] and [F7].

This section was devoted to basic concepts and notations needed throughout this chapter, in which we shall discuss some of the most important classes of combinatorial optimization problems. This will at the same time help us to become more and more familiar with graphs and digraphs.

PROBLEM SET 23.1

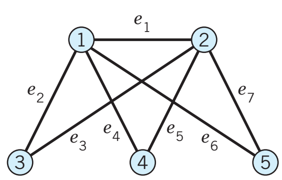
1. Explain how the following can be regarded as a graph or a digraph: a family tree, air connections between given cities, trade relations between countries, a tennis tournament, and memberships of some persons in some committees.
2. Sketch the graph consisting of the vertices and edges of a triangle. Of a pentagon. Of a tetrahedron.
3. How would you represent a net of two-way and one-way streets by a digraph?
4. Worker W_1 can do jobs J_1, J_3, J_4, worker W_2 job J_3, and worker W_3 jobs J_2, J_3, J_4. Represent this by a graph.
5. Find further situations that can be modeled by a graph or diagraph.

ADJACENCY MATRIX

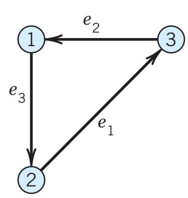
6. Show that the adjacency matrix of a graph is symmetric.
7. When will the adjacency matrix of a digraph be symmetric?

8–13 Find the adjacency matrix of the given graph or digraph.

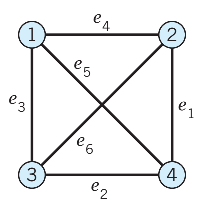
8.



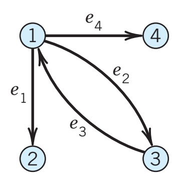
9.



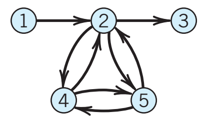
10.



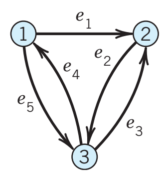
11.



12.



13.



14–15 Sketch the graph for the given adjacency matrix.

14. $$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

15. $$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

16. **Complete graph.** Show that a graph G with n vertices can have at most $n(n-1)/2$ edges, and G has exactly $n(n-1)/2$ edges if G is *complete*, that is, if every pair of vertices of G is joined by an edge. (Recall that loops and multiple edges are excluded.)

17. In what case are all the off-diagonal entries of the adjacency matrix of a graph G equal to one?

18. Incidence matrix B of a graph. The definition is $\mathbf{B} = [b_{jk}]$, where

$$b_{jk} = \begin{cases} 1 & \text{if vertex } j \text{ is an endpoint of edge } e_k, \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of the graph in Prob. **8**.

19. Incidence matrix $\widetilde{\mathbf{B}}$ of a digraph. The definition is $\widetilde{\mathbf{B}} = [\tilde{b}_{jk}]$, where

$$\tilde{b}_{jk} = \begin{cases} -1 & \text{if edge } e_k \text{ leaves vertex } j, \\ 1 & \text{if edge } e_k \text{ enters vertex } j, \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of the digraph in Prob. 11.

20. Make the vertex incidence list of the digraph in Prob. 11.

23.2 Shortest Path Problems. Complexity

The rest of this chapter is devoted to the most important classes of problems of combinatorial optimization that can be represented by graphs and digraphs. We selected these problems because of their importance in applications, and present their solutions in algorithmic form. Although basic ideas and algorithms will be explained and illustrated by small graphs, you should keep in mind that real-life problems may often involve many thousands or even millions of vertices and edges. Think of computer networks, telephone networks, electric power grids, worldwide air travel, and companies that have offices and stores in all larger cities. You can also think of other ideas for networks related to the Internet, such as electronic commerce (networks of buyers and sellers of goods over the Internet) and social networks and related websites, such as Facebook. Hence reliable and efficient systematic methods are an absolute necessity—solutions by trial and error would no longer work, even if "nearly optimal" solutions were acceptable.

We begin with **shortest path problems**, as they arise, for instance, in designing shortest (or least expensive, or fastest) routes for a traveling salesman, for a cargo ship, etc. Let us first explain what we mean by a path.

In a graph $G = (V, E)$ we can walk from a vertex v_1 along some edges to some other vertex v_k. Here we can

(A) make no restrictions, or

(B) require that each *edge* of G be traversed at most once, or

(C) require that each *vertex* be visited at most once.

In case (A) we call this a **walk**. Thus a walk from v_1 to v_k is of the form

$$(v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k), \tag{1}$$

where some of these edges or vertices may be the same. In case (B), where each *edge* may occur at most once, we call the walk a **trail**. Finally, in case (C), where each *vertex* may occur at most once (and thus each edge automatically occurs at most once), we call the trail a **path**.

We admit that a walk, trail, or path may end at the vertex it started from, in which case we call it **closed**; then $v_k = v_1$ in (1).

A closed path is called a **cycle**. *A cycle has at least three edges* (because we do not have double edges; see Sec. 23.1). Figure 481 illustrates all these concepts.

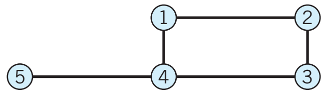


Fig. 481. Walk, trail, path, cycle

$1 - 2 - 3 - 2$ is a walk (not a trail).
$4 - 1 - 2 - 3 - 4 - 5$ is a trail (not a path).
$1 - 2 - 3 - 4 - 5$ is a path (not a cycle).
$1 - 2 - 3 - 4 - 1$ is a cycle.

Shortest Path

To define the concept of a shortest path, we assume that $G = (V, E)$ is a **weighted graph**, that is, each edge (v_i, v_j) in G has a given *weight* or *length* $l_{ij} > 0$. Then a **shortest path** $v_1 \to v_k$ (with fixed v_1 and v_k) is a path (1) such that the sum of the lengths of its edges

$$l_{12} + l_{23} + l_{34} + \cdots + l_{k-1,k}$$

(l_{12} = length of (v_1, v_2), etc.) is minimum (as small as possible among all paths from v_1 to v_k). Similarly, a **longest path** $v_1 \to v_k$ is one for which that sum is maximum.

Shortest (and longest) path problems are among the most important optimization problems. Here, "length" l_{ij} (often also called "cost" or "weight") can be an actual length measured in miles or travel time or fuel expenses, but it may also be something entirely different.

For instance, the *traveling salesman problem* requires the determination of a shortest **Hamiltonian**[1] **cycle** in a graph, that is, a cycle that contains all the vertices of the graph.

In more detail, the traveling salesman problem in its most basic and intuitive form can be stated as follows. You have a salesman who has to drive by car to his customers. He has to drive to n cities. He can start at any city and after completion of the trip he has to return to that city. Furthermore, he can only visit each city once. All the cities are linked by roads to each other, so any city can be visited from any other city directly, that is, if he wants to go from one city to another city, there is only one direct road connecting those two cities. He has to find the optimal route, that is, the route with the shortest total mileage for the overall trip. This is a classic problem in combinatorial optimization and comes up in many different versions and applications. The maximum number of possible paths to be examined in the process of selecting the optimal path for n cities is $(n - 1)!/2$, because, after you pick the first city, you have $n - 1$ choices for the second city, $n - 2$ choices for the third city, etc. You get a total of $(n - 1)!$ (see Sec. 24.4). However, since the mileage does not depend on the direction of the tour (e.g., for $n = 4$ (four cities 1, 2, 3, 4), the tour 1–2–3–4–1 has the same mileage as 1–4–3–2–1, etc., so that we counted all the tours twice!), the final answer is $(n - 1)!/2$. Even for a small number of cities, say $n = 15$, the maximum number of possible paths is very large. Use your calculator or CAS to see for yourself! This means that this is a very difficult problem for larger n and typical of problems in combinatorial optimization, in that you want a discrete solution but where it might become nearly impossible to explicitly search through all the possibilities and therefore some heuristics (rules of thumbs, shortcuts) might be used, and a less than optimal answer suffices.

[1] WILLIAM ROWAN HAMILTON (1805–1865), Irish mathematician, known for his work in dynamics.

A variation of the traveling salesman problem is the following. By choosing the "most profitable" route $v_1 \rightarrow v_k$, a salesman may want to maximize Σl_{ij}, where l_{ij} is his expected commission minus his travel expenses for going from town i to town j.

In an investment problem, i may be the day an investment is made, j the day it matures, and l_{ij} the resulting profit, and one gets a graph by considering the various possibilities of investing and reinvesting over a given period of time.

Shortest Path If All Edges Have Length $l = 1$

Obviously, if all edges have length l, then a shortest path $v_1 \rightarrow v_k$ is one that has the smallest number of edges among all paths $v_1 \rightarrow v_k$ in a given graph G. For this problem we discuss a BFS algorithm. BFS stands for **Breadth First Search**. This means that in each step the algorithm visits *all neighboring* (all adjacent) vertices of a vertex reached, as opposed to a DFS algorithm (**Depth First Search** algorithm), which makes a long trail (as in a maze). This widely used BFS algorithm is shown in Table 23.1.

We want to find a shortest path in G from a vertex s (***start***) to a vertex t (***terminal***). To guarantee that there is a path from s to t, we make sure that G does not consist of separate portions. Thus we assume that G is **connected**, that is, for any two vertices v and w there is a path $v \rightarrow w$ in G. (Recall that a vertex v is called **adjacent** to a vertex u if there is an edge (u, v) in G.)

Table 23.1 Moore's[2] BFS for Shortest Path (All Lengths One)

Proceedings of the International Symposium for Switching Theory, Part II. pp. 285–292. Cambridge: Harvard University Press, 1959.

ALGORITHM MOORE [$G = (V, E)$, s, t]

This algorithm determines a shortest path in a connected graph $G = (V, E)$ from a vertex s to a vertex t.

INPUT: Connected graph $G = (V, E)$, in which one vertex is denoted by s and one by t, and each edge (i, j) has length $l_{ij} = 1$. Initially all vertices are unlabeled.

OUTPUT: A shortest path $s \rightarrow t$ in $G = (V, E)$

1. Label s with 0.
2. Set $i = 0$.
3. Find all *unlabeled* vertices adjacent to a vertex labeled i.
4. Label the vertices just found with $i + 1$.
5. If vertex t is labeled, then "backtracking" gives the shortest path

$$k\ (= \text{label of } t),\ k - 1,\ k - 2,\ \cdots,\ 0$$

OUTPUT $k, k - 1, k - 2, \cdots, 0$. Stop

Else increase i by 1. Go to Step 3.

End MOORE

[2]EDWARD FORREST MOORE (1925–2003), American mathematician and computer scientist, who did pioneering work in theoretical computer science (automata theory, Turing machines).

EXAMPLE 1 Application of Moore's BFS Algorithm

Find a shortest path $s \to t$ in the graph G shown in Fig. 482.

Solution. Figure 482 shows the labels. The blue edges form a shortest path (length 4). There is another shortest path $s \to t$. (Can you find it?) Hence in the program we must introduce a rule that makes backtracking unique because otherwise the computer would not know what to do next if at some step there is a choice (for instance, in Fig. 482 when it got back to the vertex labeled 2). The following rule seems to be natural.

Backtracking rule. Using the numbering of the vertices from 1 to n (not the labeling!), at each step, if a vertex labeled i is reached, take as the next vertex that with the smallest number (not label!) among all the vertices labeled $i - 1$. ■

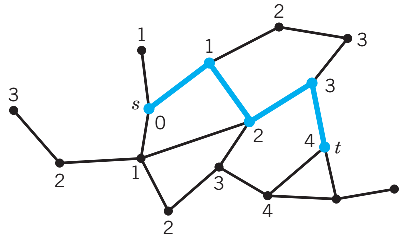


Fig. 482. Example 1, given graph and result of labeling

Complexity of an Algorithm

Complexity *of Moore's algorithm.* To find the vertices to be labeled 1, we have to scan all edges incident with s. Next, when $i = 1$, we have to scan all edges incident with vertices labeled 1, etc. Hence each edge is scanned twice. These are $2m$ operations (m = number of edges of G). This is a function $c(m)$. Whether it is $2m$ or $5m + 3$ or $12m$ is not so essential; it *is* essential that $c(m)$ is proportional to m (not m^2, for example); it is of the "order" m. We write for any function $am + b$ simply $O(m)$, for any function $am^2 + bm + d$ simply $O(m^2)$, and so on; here, O suggests **order**. The underlying idea and practical aspect are as follows.

In judging an algorithm, we are mostly interested in its behavior for very large problems (large m in the present case), since these are going to determine the limits of the applicability of the algorithm. Thus, the essential item is the fastest growing term (am^2 in $am^2 + bm + d$, etc.) since it will overwhelm the others when m is large enough. Also, a constant factor in this term is not very essential; for instance, the difference between two algorithms of orders, say, $5m^2$ and $8m^2$ is generally not very essential and can be made irrelevant by a modest increase in the speed of computers. However, it does make a great practical difference whether an algorithm is of order m or m^2 or of a still higher power m^p. And the biggest difference occurs between these "polynomial orders" and "exponential orders," such as 2^m.

For instance, on a computer that does 10^9 operations per second, a problem of size $m = 50$ will take 0.3 sec with an algorithm that requires m^5 operations, but 13 days with an algorithm that requires 2^m operations. But this is not our only reason for regarding polynomial orders as good and exponential orders as bad. Another reason is the ***gain in using a faster computer***. For example, let two algorithms be $O(m)$ and $O(m^2)$. Then, since $1000 = 31.6^2$, an increase in speed by a factor 1000 has the effect that per hour we can do problems 1000 and 31.6 times as big, respectively. But since $1000 = 2^{9.97}$, with an algorithm that is $O(2^m)$, all we gain is a relatively modest increase of 10 in problem size because $2^{9.97} \cdot 2^m = 2^{m+9.97}$.

The **symbol** $\boldsymbol{O}$ is quite practical and commonly used whenever the order of growth is essential, but not the specific form of a function. Thus if a function $g(m)$ is of the form

$$g(m) = kh(m) + \text{more slowly growing terms} \qquad (k \neq 0, \text{ constant}),$$

we say that $g(m)$ is of the *order* $h(m)$ and write

$$g(m) = O(h(m)).$$

For instance,

$$am + b = O(m), \qquad am^2 + bm + d = O(m^2), \qquad 5 \cdot 2^m + 3m^2 = O(2^m).$$

We want an algorithm $\mathcal{A}$ to be "efficient," that is, "good" with respect to

(i) *Time* (number $c_{\mathcal{A}}(m)$ of computer operations), or

(ii) *Space* (storage needed in the internal memory)

or both. Here $c_{\mathcal{A}}$ suggests "**complexity**" of $\mathcal{A}$. Two popular choices for $c_{\mathcal{A}}$ are

(*Worst case*) $\quad c_{\mathcal{A}}(m) =$ longest time $\mathcal{A}$ takes for a problem of size m,

(*Average case*) $\quad c_{\mathcal{A}}(m) =$ average time $\mathcal{A}$ takes for a problem of size m.

In problems on graphs, the "size" will often be m (number of edges) or n (number of vertices). For Moore's algorithm, $c_{\mathcal{A}}(m) = 2m$ in both cases. Hence the complexity of Moore's algorithm is of order $O(m)$.

For a "good" algorithm $\mathcal{A}$, we want that $c_{\mathcal{A}}(m)$ does not grow too fast. Accordingly, we call $\mathcal{A}$ **efficient** if $c_{\mathcal{A}}(m) = O(m^k)$ for some integer $k \geqq 0$; that is, $c_{\mathcal{A}}$ may contain only powers of m (or functions that grow even more slowly, such as $\ln m$), but no exponential functions. Furthermore, we call $\mathcal{A}$ **polynomially bounded** if $\mathcal{A}$ is efficient when we choose the "worst case" $c_{\mathcal{A}}(m)$. These conventional concepts have intuitive appeal, as our discussion shows.

Complexity should be investigated for every algorithm, so that one can also compare different algorithms for the same task. This may often exceed the level in this chapter; accordingly, we shall confine ourselves to a few occasional comments in this direction.

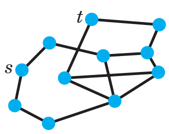
PROBLEM SET 23.2

SHORTEST PATHS, MOORE'S BFS

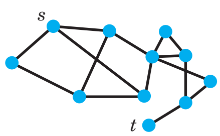
(All edges length one)

1–4 Find a shortest path $P: s \to t$ and its length by Moore's algorithm. Sketch the graph with the labels and indicate P by heavier lines as in Fig. 482.

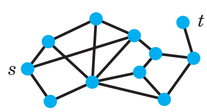
1.



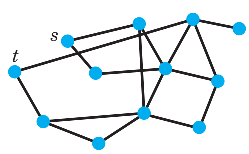
2.



3.



4.



5. Moore's algorithm. Show that if vertex v has label $\lambda(v) = k$, then there is a path $s \to v$ of length k.

6. Maximum length. What is the maximum number of edges that a shortest path between any two vertices in a graph with n vertices can have? Give a reason. In a complete graph with all edges of length 1?

7. **Nonuniqueness.** Find another shortest path from s to t in Example 1 of the text.
8. **Moore's algorithm.** Call the length of a shortest path $s \to v$ the *distance* of v from s. Show that if v has distance l, it has label $\lambda(v) = l$.
9. **CAS PROBLEM. Moore's Algorithm.** Write a computer program for the algorithm in Table 23.1. Test the program with the graph in Example 1. Apply it to Probs. 1–3 and to some graphs of your own choice.

10–12 HAMILTONIAN CYCLE

10. Find and sketch a Hamiltonian cycle in the graph of a dodecahedron, which has 12 pentagonal faces and 20 vertices (Fig. 483). This is a problem Hamilton himself considered.

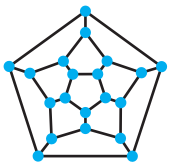


Fig. 483. Problem 10

11. Find and sketch a Hamiltonian cycle in Prob. 1.
12. Does the graph in Prob. 4 have a Hamiltonian cycle?

13–14 POSTMAN PROBLEM

13. The **postman problem** is the problem of finding a closed walk $W: s \to s$ (s the post office) in a graph G with edges (i, j) of length $l_{ij} > 0$ such that every edge of G is traversed at least once and the length of W is minimum. Find a solution for the graph in Fig. 484 by inspection. (The problem is also called the *Chinese postman problem* since it was published in the journal *Chinese Mathematics* 1 (1962), 273–277.)

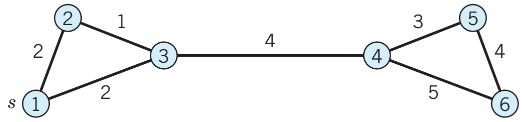


Fig. 484. Problem 13

14. Show that the length of a shortest postman trail is the same for every starting vertex.

15–17 EULER GRAPHS

15. An **Euler graph** G is a graph that has a closed Euler trail. An **Euler trail** is a trail that contains every edge of G exactly once. Which subgraph with four edges of the graph in Example 1, Sec. 23.1, is an Euler graph?
16. Find four different closed Euler trails in Fig. 485.

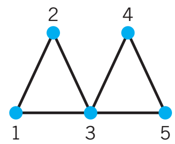


Fig. 485. Problem 16

17. Is the graph in Fig. 484 an Euler graph. Give reason.

18–20 ORDER

18. Show that $O(m^3) + O(m^3) = O(m^3)$ and $kO(m^p) = O(m^p)$.
19. Show that $\sqrt{1 + m^2} = O(m)$, $0.02e^m + 100m^2 = O(e^m)$.
20. If we switch from one computer to another that is 100 times as fast, what is our gain in problem size per hour in the use of an algorithm that is $O(m)$, $O(m^2)$, $O(m^5)$, $O(e^m)$?

23.3 Bellman's Principle. Dijkstra's Algorithm

We continue our discussion of the shortest path problem in a graph G. The last section concerned the special case that all edges had length 1. But in most applications the edges (i, j) will have any lengths $l_{ij} > 0$, and we now turn to this general case, which is of greater practical importance. We write $l_{ij} = \infty$ for any edge (i, j) that does not exist in G (setting $\infty + a = \infty$ for any number a, as usual).

We consider the problem of finding shortest paths from a given vertex, denoted by 1 and called the **origin**, to ***all*** other vertices $2, 3, \cdots, n$ of G. We let L_j denote the length of a shortest path P_j: $1 \to j$ in G.

THEOREM 1

Bellman's Minimality Principle or Optimality Principle[3]

If P_j: $1 \rightarrow j$ is a shortest path from 1 to j in G and (i, j) is the last edge of P_j (Fig. 486), *then P_i: $1 \rightarrow i$ [obtained by dropping (i, j) from P_j] is a shortest path $1 \rightarrow i$.*

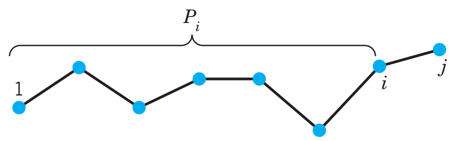


Fig. 486. Paths P and P_i in Bellman's minimality principle

PROOF Suppose that the conclusion is false. Then there is a path P_i^*: $1 \rightarrow i$ that is shorter than P_i. Hence, if we now add (i, j) to P_i^*, we get a path $1 \rightarrow j$ that is shorter than P_j. This contradicts our assumption that P_j is shortest. ■

From Bellman's principle we can derive basic equations as follows. For fixed j we may obtain various paths $1 \rightarrow j$ by taking shortest paths P_i for various i for which there is in G an edge (i, j), and add (i, j) to the corresponding P_i. These paths obviously have lengths $L_i + l_{ij}$ (L_i = length of P_i). We can now take the minimum over i, that is, pick an i for which $L_i + l_{ij}$ is smallest. By the Bellman principle, this gives a shortest path $1 \rightarrow j$. It has the length

$$\text{(1)} \qquad \begin{aligned} L_1 &= 0 \\ L_j &= \min_{i \neq j} (L_i + l_{ij}), \end{aligned} \qquad j = 2, \cdots, n.$$

These are the **Bellman equations**. Since $l_{ii} = 0$ by definition, instead of $\min_{i \neq j}$ we can simply write $\min_i$. These equations suggest the idea of one of the best-known algorithms for the shortest path problem, as follows.

Dijkstra's Algorithm for Shortest Paths

Dijkstra's[4] **algorithm** is shown in Table 23.2, where a **connected graph** G is a graph in which, for any two vertices v and w in G, there is a path $v \rightarrow w$. The algorithm is a labeling procedure. At each stage of the computation, each vertex v gets a label, either

(PL) a *permanent label* = length L_v of a shortest path $1 \rightarrow v$

or

(TL) a *temporary label* = upper bound $\tilde{L}_v$ for the length of a shortest path $1 \rightarrow v$.

[3] RICHARD BELLMAN (1920–1984), American mathematician, known for his work in dynamic programming.
[4] EDSGER WYBE DIJKSTRA (1930–2002), Dutch computer scientist, 1972 recipient of the ACM Turing Award. His algorithm appeared in *Numerische Mathematik* **1** (1959), 269–271.

We denote by $\mathcal{PL}$ and $\mathcal{TL}$ the sets of vertices with a permanent label and with a temporary label, respectively. The algorithm has an initial step in which vertex 1 gets the permanent label $L_1 = 0$ and the other vertices get temporary labels, and then the algorithm alternates between Steps 2 and 3. In Step 2 the idea is to pick k "minimally." In Step 3 the idea is that the upper bounds will in general improve (decrease) and must be updated accordingly. Namely, the new temporary label $\tilde{L}_j$ of vertex j will be the old one if there is no improvement or it will be $L_k + l_{kj}$ if there is.

Table 23.2 Dijkstra's Algorithm for Shortest Paths

ALGORITHM DIJKSTRA [$G = (V, E)$, $V = \{1, \cdots, n\}$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, \cdots, n$ and edges (i, j) having lengths $l_{ij} > 0$, this algorithm determines the lengths of shortest paths from vertex 1 to the vertices $2, \cdots, n$.

INPUT: Number of vertices n, edges (i, j), and lengths l_{ij}

OUTPUT: Lengths L_j of shortest paths $1 \to j$, $j = 2, \cdots, n$

1. *Initial step*

Vertex 1 gets PL: $L_1 = 0$.
Vertex j $(= 2, \cdots, n)$ gets TL: $\tilde{L}_j = l_{1j}$ $(= \infty$ if there is no edge $(1, j)$ in $G)$.
Set $\mathcal{PL} = \{1\}$, $\mathcal{TL} = \{2, 3, \cdots, n\}$.

2. *Fixing a permanent label*

Find a k in $\mathcal{TL}$ for which $\tilde{L}_k$ is miminum, set $L_k = \tilde{L}_k$. Take the smallest k if there are several. Delete k from $\mathcal{TL}$ and include it in $\mathcal{PL}$.
If $\mathcal{TL} = \varnothing$ (that is, $\mathcal{TL}$ is empty) then

OUTPUT $L_2, \cdots, L_n$. Stop

Else continue (that is, go to Step 3).

3. *Updating temporary labels*

For all j in $\mathcal{TL}$, set $\tilde{L}_j = \min_k \{\tilde{L}_j, L_k + l_{kj}\}$ (that is, take the smaller of $\tilde{L}_j$ and $L_k + l_{kj}$ as your new $\tilde{L}_j$).

Go to Step 2.

End DIJKSTRA

EXAMPLE 1 Application of Dijkstra's Algorithm

Applying Dijkstra's algorithm to the graph in Fig. 487a, find shortest paths from vertex 1 to vertices 2, 3, 4.

Solution. We list the steps and computations.

1.	$L_1 = 0, \tilde{L}_2 = 8, \tilde{L}_3 = 5, \tilde{L}_4 = 7,$	$\mathcal{PL} = \{1\},$	$\mathcal{TL} = \{2, 3, 4\}$
2.	$L_3 = \min\{\tilde{L}_2, \tilde{L}_3, \tilde{L}_4\} = 5, k = 3,$	$\mathcal{PL} = \{1, 3\},$	$\mathcal{TL} = \{2, 4\}$
3.	$\tilde{L}_2 = \min\{8, L_3 + l_{32}\} = \min\{8, 5 + 1\} = 6$		
	$\tilde{L}_4 = \min\{7, L_3 + l_{34}\} = \min\{7, \infty\} = 7$		
2.	$L_2 = \min\{\tilde{L}_2, \tilde{L}_4\} = \min\{6, 7\} = 6, k = 2,$	$\mathcal{PL} = \{1, 2, 3\},$	$\mathcal{TL} = \{4\}$
3.	$\tilde{L}_4 = \min\{7, L_2 + l_{24}\} = \min\{7, 6 + 2\} = 7$		
2.	$L_4 = 7, k = 4$	$\mathcal{PL} = \{1, 2, 3, 4\},$	$\mathcal{TL} = \varnothing.$

Figure 487b shows the resulting shortest paths, of lengths $L_2 = 6$, $L_3 = 5$, $L_4 = 7$. ■

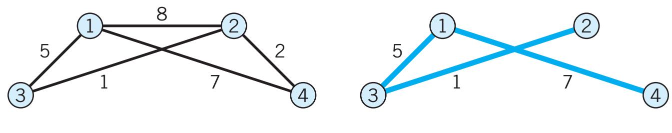


Fig. 487. Example 1

Complexity. *Dijkstra's algorithm is* $O(n^2)$.

PROOF Step 2 requires comparison of elements, first $n - 2$, the next time $n - 3$, etc., a total of $(n - 2)(n - 1)/2$. Step 3 requires the same number of comparisons, a total of $(n - 2)(n - 1)/2$, as well as additions, first $n - 2$, the next time $n - 3$, etc., again a total of $(n - 2)(n -)/2$. Hence the total number of operations is $3(n - 2)(n - 1)/2 = O(n^2)$. ■

PROBLEM SET 23.3

1. The net of roads in Fig. 488 connecting four villages is to be reduced to minimum length, but so that one can still reach every village from every other village. Which of the roads should be retained? Find the solution **(a)** by inspection, **(b)** by Dijkstra's algorithm.

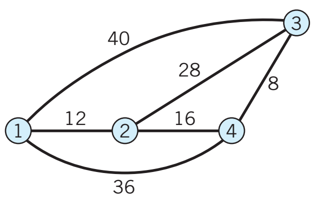


Fig. 488. Problem 1

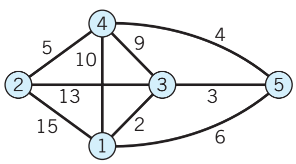
2. Show that in Dijkstra's algorithm, for L_k there is a path P: $1 \rightarrow k$ of length L_k.

3. Show that in Dijkstra's algorithm, at each instant the demand on storage is light (data for fewer than n edges).

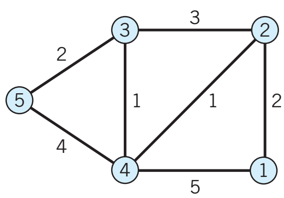
4–9 DIJKSTRA'S ALGORITHM

For each graph find the shortest paths.

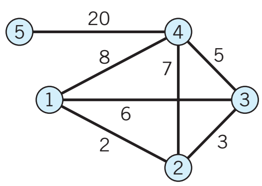
4.



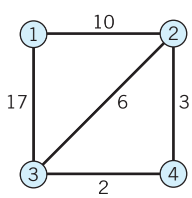
5.

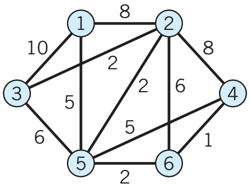


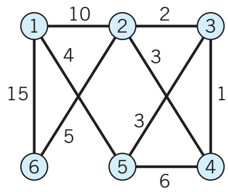
6.



7.



8. 

9. 

23.4 Shortest Spanning Trees: Greedy Algorithm

So far we have discussed shortest path problems. We now turn to a particularly important kind of graph, called a tree, along with related optimization problems that arise quite often in practice.

By definition, a **tree** T is a graph that is connected and has no cycles. "**Connected**" was defined in Sec. 23.3; it means that there is a path from any vertex in T to any other vertex in T. A **cycle** is a path $s \to t$ of at least three edges that is closed ($t = s$); see also Sec. 23.2. Figure 489a shows an example.

CAUTION! The terminology varies; *cycles* are sometimes also called *circuits*.

A **spanning tree** T in a given connected graph $G = (V, E)$ is a tree containing *all* the n vertices of G. See Fig. 489b. Such a tree has $n - 1$ edges. (Proof?)

A **shortest spanning tree** T in a connected graph G (whose edges (i, j) have lengths $l_{ij} > 0$) is a spanning tree for which Σl_{ij} (sum over all edges of T) is minimum compared to Σl_{ij} for any other spanning tree in G.

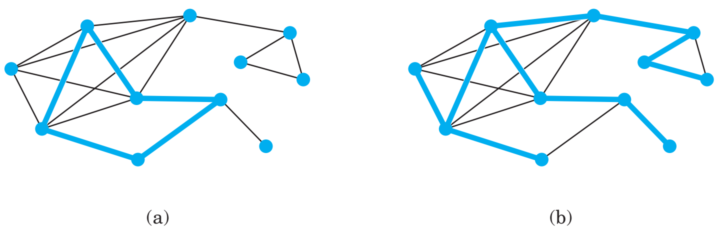


Fig. 489. Example of (a) a cycle, (b) a spanning tree in a graph

Trees are among the most important types of graphs, and they occur in various applications. Familiar examples are family trees and organization charts. Trees can be used to exhibit, organize, or analyze electrical networks, producer–consumer and other business relations, information in database systems, syntactic structure of computer programs, etc. We mention a few specific applications that need no lengthy additional explanations.

The set of shortest paths from vertex 1 to the vertices $2, \cdots, n$ in the last section forms a spanning tree.

Railway lines connecting a number of cities (the vertices) can be set up in the form of a spanning tree, the "length" of a line (edge) being the construction cost, and one wants to minimize the total construction cost. Similarly for bus lines, where "length" may be the average annual operating cost. Or for steamship lines (freight lines), where "length" may be profit and the goal is the maximization of total profit. Or in a network of telephone lines between some cities, a shortest spanning tree may simply represent a selection of lines that connect all the cities at minimal cost. In addition to these examples we could mention others from distribution networks, and so on.

We shall now discuss a simple algorithm for the problem of finding a shortest spanning tree. This algorithm (Table 23.3) is particularly suitable for sparse graphs (graphs with very few edges; see Sec. 23.1).

Table 23.3 Kruskal's[5] Greedy Algorithm for Shortest Spanning Trees

Proceedings of the American Mathematical Society **7** (1956), 48–50.

ALGORITHM KRUSKAL [$G = (V, E)$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, 2, \cdots, n$ and edges (i, j) having length $l_{ij} > 0$, the algorithm determines a shortest spanning tree T in G.

INPUT: Edges (i, j) of G and their lengths l_{ij}

OUTPUT: Shortest spanning tree T in G

1. Order the edges of G in ascending order of length.

2. Choose them in this order as edges of T, rejecting an edge only if it forms a cycle with edges already chosen.

If $n - 1$ edges have been chosen, then

OUTPUT T (= the set of edges chosen). Stop

End KRUSKAL

EXAMPLE 1 Application of Kruskal's Algorithm

Using Kruskal's algorithm, we shall determine a shortest spanning tree in the graph in Fig. 490.

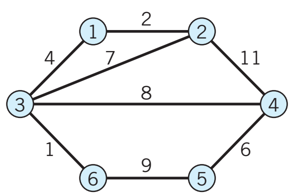


Fig. 490. Graph in Example 1

Table 23.4 Solution in Example 1

Edge	Length	Choice
(3, 6)	1	1st
(1, 2)	2	2nd
(1, 3)	4	3rd
(4, 5)	6	4th
(2, 3)	7	Reject
(3, 4)	8	5th
(5, 6)	9	
(2, 4)	11	

Solution. See Table 23.4. In some of the intermediate stages the edges chosen form a *disconnected* graph (see Fig. 491); this is typical. We stop after $n - 1 = 5$ choices since a spanning tree has $n - 1$ edges. In our problem the edges chosen are in the upper part of the list. This is typical of problems of any size; in general, edges farther down in the list have a smaller chance of being chosen. ■

---

[5]JOSEPH BERNARD KRUSKAL (1928–), American mathematician who worked at Bell Laboratories. He is known for his contributions to graph theory and statistics.

The efficiency of Kruskal's method is greatly increased by double labeling of vertices.

Double Labeling of Vertices. *Each vertex i carries a double label* (r_i, p_i), *where*

$$r_i = \textit{Root of the subtree to which } i \textit{ belongs,}$$

$$p_i = \textit{Predecessor of } i \textit{ in its subtree,}$$

$$p_i = 0 \textit{ for roots.}$$

This simplifies rejecting.

Rejecting. If (i, j) *is next in the list to be considered, reject* (i, j) *if* $r_i = r_j$ (that is, i and j are in the same subtree, so that they are already joined by edges and (i, j) would thus create a cycle). *If* $r_i \neq r_j$, *include* (i, j) *in T.*

If there are several choices for r_i, choose the smallest. If subtrees merge (become a single tree), retain the smallest root as the root of the new subtree.

For Example 1 the double-label list is shown in Table 23.5. In storing it, at each instant one may retain only the latest double label. We show all double labels in order to exhibit the process in all its stages. Labels that remain unchanged are not listed again. Underscored are the two 1's that are the common root of vertices 2 and 3, the reason for rejecting the edge (2, 3). By reading for each vertex the latest label we can read from this list that 1 is the vertex we have chosen as a root and the tree is as shown in the last part of Fig. 491.

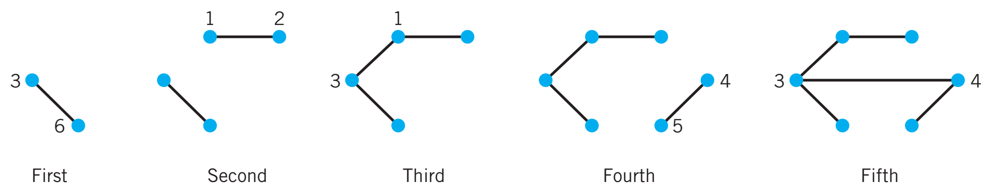


Fig. 491. Choice process in Example 1

Table 23.5 List of Double Labels in Example 1

Vertex	Choice 1 (3, 6)	Choice 2 (1, 2)	Choice 3 (1, 3)	Choice 4 (4, 5)	Choice 5 (3, 4)
1		(1, 0)			
2		$(\underline{1}, 1)$			
3	(3, 0)		$(\underline{1}, 1)$		
4				(4, 0)	(1, 3)
5				(4, 4)	(1, 4)
6	(3, 3)		(1, 3)		

This is made possible by the predecessor label that each vertex carries. Also, for accepting or rejecting an edge we have to make only one comparison (the roots of the two endpoints of the edge).

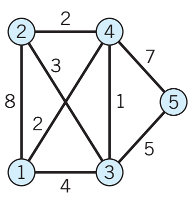
Ordering is the more expensive part of the algorithm. It is a standard process in data processing for which various methods have been suggested (see **Sorting** in Ref. [E25] listed in App. 1). For a complete list of m edges, an algorithm would be $O(m \log_2 m)$, but since the $n - 1$ edges of the tree are most likely to be found earlier, by inspecting the q ($< m$) topmost edges, for such a list of q edges one would have $O(q \log_2 m)$.

PROBLEM SET 23.4

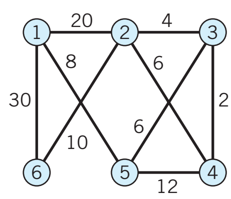
1–6 KRUSKAL'S GREEDY ALGORITHM

Find a shortest spanning tree by Kruskal's algorithm. Sketch it.

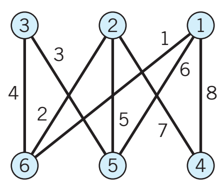
1.



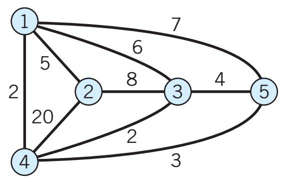
2.



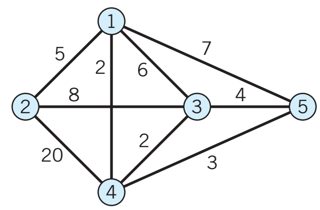
3.



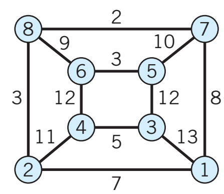
4.



5.



6.



7. CAS PROBLEM. Kruskal's Algorithm. Write a corresponding program. (Sorting is discussed in Ref. [E25] listed in App. 1.)

8. To get a minimum spanning tree, instead of adding shortest edges, one could think of deleting longest edges. For what graphs would this be feasible? Describe an algorithm for this.

9. Apply the method suggested in Prob. 8 to the graph in Example 1. Do you get the same tree?

10. Design an algorithm for obtaining longest spanning trees.

11. Apply the algorithm in Prob. 10 to the graph in Example 1. Compare with the result in Example 1.

12. Forest. A (not necessarily connected) graph without cycles is called a *forest*. Give typical examples of applications in which graphs occur that are forests or trees.

	Dallas	Denver	Los Angeles	New York	Washington, DC
Chicago	800	900	1800	700	650
Dallas		650	1300	1350	1200
Denver			850	1650	1500
Los Angeles				2500	2350
New York					200

13. Air cargo. Find a shortest spanning tree in the complete graph of all possible 15 connections between the six cities given (distances by airplane, in miles, rounded). Can you think of a practical application of the result?

14–20 GENERAL PROPERTIES OF TREES

Prove the following. *Hint.* Use Prob. 14 in proving 15 and 18; use Probs. 16 and 18 in proving 20.

14. Uniqueness. The path connecting any two vertices u and v in a tree is unique.

15. If in a graph any two vertices are connected by a unique path, the graph is a tree.

16. If a graph has no cycles, it must have at least 2 vertices of degree 1 (definition in Sec. 23.1).

17. A tree with exactly two vertices of degree 1 must be a path.

18. A tree with n vertices has $n - 1$ edges. (Proof by induction.)

19. If two vertices in a tree are joined by a new edge, a cycle is formed.

20. A graph with n vertices is a tree if and only if it has $n - 1$ edges and has no cycles.

23.5 Shortest Spanning Trees: Prim's Algorithm

Prim's[6] algorithm, shown in Table 23.6, is another popular algorithm for the shortest spanning tree problem (see Sec. 23.4). This algorithm avoids ordering edges and gives a tree T at each stage, a property that Kruskal's algorithm in the last section did not have (look back at Fig. 491 if you did not notice it).

In Prim's algorithm, starting from any single vertex, which we call 1, we "grow" the tree T by adding edges to it, one at a time, according to some rule (in Table 23.6) until T finally becomes a *spanning* tree, which is shortest.

We denote by U the set of vertices of the growing tree T and by S the set of its edges. Thus, initially $U = \{1\}$ and $S = \varnothing$; at the end, $U = V$, the vertex set of the given graph $G = (V, E)$, whose edges (i, j) have length $l_{ij} > 0$, as before.

[6]ROBERT CLAY PRIM (1921–), American computer scientist at General Electric, Bell Laboratories, and Sandia National Laboratories.

Thus at the beginning (Step 1) the labels

$$\lambda_2, \cdots, \lambda_n \qquad \text{of the vertices} \qquad 2, \cdots, n$$

are the lengths of the edges connecting them to vertex 1 (or ∞ if there is no such edge in G). And we pick (Step 2) the shortest of these as the first edge of the growing tree T and include its other end j in U (choosing the smallest j if there are several, to make the process unique). Updating labels in Step 3 (at this stage and at any later stage) concerns each vertex k not yet in U. Vertex k has label $\lambda_k = l_{i(k),k}$ from before. If $l_{jk} < \lambda_k$, this means that k is closer to the new member j just included in U than k is to its old "closest neighbor" $i(k)$ in U. Then we update the label of k, replacing $\lambda_k = l_{i(k),k}$ by $\lambda_k = l_{jk}$ and setting $i(k) = j$. If, however, $l_{jk} \geqq \lambda_k$ (the *old* label of k), we don't touch the old label. Thus the label λ_k always identifies the closest neighbor of k in U, and this is updated in Step 3 as U and the tree T grow. From the final labels we can backtrack the final tree, and from their numeric values we compute the total length (sum of the lengths of the edges) of this tree.

Prim's algorithm is useful for computer network design, cable, distribution networks, and transportation networks.

Table 23.6 Prim's Algorithm for Shortest Spanning Trees

Bell System Technical Journal **36** (1957), 1389–1401.

For an improved version of the algorithm, see Cheriton and Tarjan, *SIAM Journal on Computation* **5** (1976), 724–742.

ALGORITHM PRIM [$G = (V, E)$, $V = \{1, \cdots, n\}$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, 2, \cdots, n$ and edges (i, j) having length $l_{ij} > 0$, this algorithm determines a shortest spanning tree T in G and its length $L(T)$.

INPUT: n, edges (i, j) of G and their lengths l_{ij}
OUTPUT: Edge set S of a shortest spanning tree T in G; $L(T)$
[*Initially, all vertices are unlabeled.*]

1. *Initial step*
Set $i(k) = 1$, $U = \{1\}$, $S = \varnothing$.
Label vertex k $(= 2, \cdots, n)$ with $\lambda_k = l_{ik}$ [$= \infty$ if G has no edge $(1, k)$].

2. *Addition of an edge to the tree T*
Let λ_j be the smallest λ_k for vertex k not in U. Include vertex j in U and edge $(i(j), j)$ in S.
If $U = V$ then compute
$L(T) = \Sigma l_{ij}$ (sum over all edges in S)
OUTPUT S, $L(T)$. Stop
[*S is the edge set of a shortest spanning tree T in G.*]
Else continue (that is, go to Step 3).

3. *Label updating*
For every k not in U, if $l_{jk} < \lambda_k$, then set $\lambda_k = l_{jk}$ and $i(k) = j$.
Go to Step 2.

End PRIM

EXAMPLE 1 Application of Prim's Algorithm

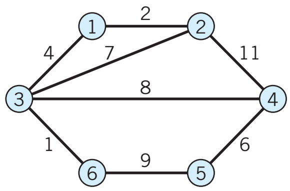


Fig. 492. Graph in Example 1

Find a shortest spanning tree in the graph in Fig. 492 (which is the same as in Example 1, Sec. 23.4, so that we can compare).

Solution. The steps are as follows.

1. $i(k) = 1$, $U = \{1\}$, $S = \varnothing$, initial labels see Table 23.7.

2. $\lambda_2 = l_{12} = 2$ is smallest, $U = \{1, 2\}$, $S = \{(1, 2)\}$.

3. Update labels as shown in Table 23.7, column (I).

2. $\lambda_3 = l_{13} = 4$ is smallest, $U = \{1, 2, 3\}$, $S = \{(1, 2), (1, 3)\}$.

3. Update labels as shown in Table 23.7, column (II).

2. $\lambda_6 = l_{36} = 1$ is smallest, $U = \{1, 2, 3, 6\}$, $S = \{(1, 2), (1, 3), (3, 6)\}$.

3. Update labels as shown in Table 23.7, column (III).

2. $\lambda_4 = l_{34} = 8$ is smallest, $U = \{1, 2, 3, 4, 6\}$, $S = \{(1, 2), (1, 3), (3, 4), (3, 6)\}$.

3. Update labels as shown in Table 23.7, column (IV).

2. $\lambda_5 = l_{45} = 6$ is smallest, $U = V$, $S = (1, 2), (1, 3), (3, 4), (3, 6), (4, 5)$. Stop.

The tree is the same as in Example 1, Sec. 23.4. Its length is 21. You will find it interesting to compare the growth process of the present tree with that in Sec. 23.4. ■

Table 23.7 Labeling of Vertices in Example 1

Vertex	Initial Label	Relabeling (I)	(II)	(III)	(IV)
2	$l_{12} = 2$	—	—	—	—
3	$l_{13} = 4$	$l_{13} = 4$	—	—	—
4	∞	$l_{24} = 11$	$l_{34} = 8$	$l_{34} = 8$	—
5	∞	∞	∞	$l_{65} = 9$	$l_{45} = 6$
6	∞	∞	$l_{36} = 1$	—	—

PROBLEM SET 23.5

SHORTEST SPANNING TREES. PRIM'S ALGORITHM

1. When will $S = E$ at the end in Prim's algorithm?

2. Complexity. Show that Prim's algorithm has complexity $O(n^2)$.

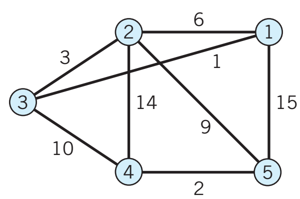
3. What is the result of applying Prim's algorithm to a graph that is not connected?

4. If for a complete graph (or one with very few edges missing), our data is an $n \times n$ distance table (as in Prob. 13, Sec. 23.4), show that the present algorithm [which is $O(n^2)$] cannot easily be replaced by an algorithm of order less than $O(n^2)$.

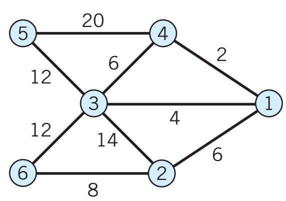
5. How does Prim's algorithm prevent the generation of cycles as you grow T?

6–13 Find a shortest spanning tree by Prim's algorithm.

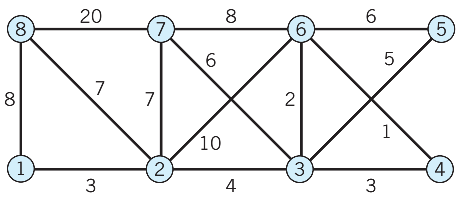
6.



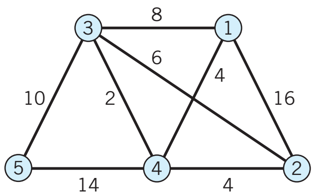
7.



8.



9.



10. For the graph in Prob. 6, Sec. 23.4.

11. For the graph in Prob. 4, Sec. 23.4.

12. For the graph in Prob. 2, Sec. 23.4.

13. CAS PROBLEM. Prim's Algorithm. Write a program and apply it to Probs. 6–9.

14. TEAM PROJECT. Center of a Graph and Related Concepts. (a) Distance, Eccentricity. Call the length of a shortest path $u \rightarrow v$ in a graph $G = (V, E)$ the *distance* $d(u, v)$ from u to v. For fixed u, call the greatest $d(u, v)$ as v ranges over V the *eccentricity* $\epsilon(u)$ of u. Find the eccentricity of vertices 1, 2, 3 in the graph in Prob. 7.

(b) Diameter, Radius, Center. The *diameter* $d(G)$ of a graph $G = (V, E)$ is the maximum of $d(u, v)$ as u and v vary over V, and the *radius* $r(G)$ is the smallest eccentricity $\epsilon(v)$ of the vertices v. A vertex v with $\epsilon(v) = r(G)$ is called a *central vertex*. The set of all central vertices is called the *center* of G. Find $d(G)$, $r(G)$, and the center of the graph in Prob. 7.

(c) What are the diameter, radius, and center of the spanning tree in Example 1 of the text?

(d) Explain how the idea of a center can be used in setting up an emergency service facility on a transportation network. In setting up a fire station, a shopping center. How would you generalize the concepts in the case of two or more such facilities?

(e) Show that a tree T whose edges all have length 1 has center consisting of either one vertex or two adjacent vertices.

(f) Set up an algorithm of complexity $O(n)$ for finding the center of a tree T.

23.6 Flows in Networks

After shortest path problems and problems for trees, as a third large area in combinatorial optimization we discuss **flow problems in networks** (electrical, water, communication, traffic, business connections, etc.), turning from graphs to digraphs (directed graphs; see Sec. 23.1).

By definition, a **network** is a digraph $G = (V, E)$ in which each edge (i, j) has assigned to it a **capacity** $c_{ij} > 0$ [= maximum possible flow along (i, j)], and at one vertex, s, called the **source**, a flow is produced that flows along the edges of the digraph G to another vertex, t, called the **target** or **sink**, where the flow disappears.

In applications, this may be the flow of electricity in wires, of water in pipes, of cars on roads, of people in a public transportation system, of goods from a producer to consumers, of e-mail from senders to recipients over the Internet, and so on.

We denote the flow along a (directed!) edge (i, j) by f_{ij} and impose two conditions:

1. For each edge (i, j) in G the flow does not exceed the capacity c_{ij},

$$0 \leqq f_{ij} \leqq c_{ij} \qquad \text{(“Edge condition”).} \tag{1}$$

2. For each vertex i, not s or t,

$$\text{Inflow} = \text{Outflow} \qquad \text{(“Vertex condition,” “Kirchhoff's law”);}$$

in a formula,

$$\underbrace{\sum_k f_{ki}}_{\text{Inflow}} - \underbrace{\sum_j f_{ij}}_{\text{Outflow}} = \begin{cases} 0 \text{ if vertex } i \neq s, i \neq t, \\ -f \text{ at the source } s, \\ f \text{ at the target (sink) } t, \end{cases} \tag{2}$$

where f is the total flow (and at s the inflow is zero, whereas at t the outflow is zero). Figure 493 illustrates the notation (for some hypothetical figures).

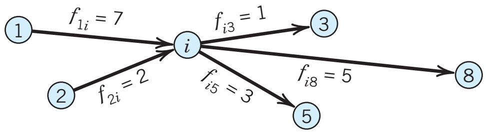


Fig. 493. Notation in (2): inflow and outflow for a vertex i (not s or t)

Paths

By a **path** $v_1 \to v_k$ from a vertex v_1 to a vertex v_k in a digraph G we mean a sequence of edges

$$(v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k),$$

regardless of their directions in G, that forms a path as in a graph (see Sec. 23.2). Hence when we travel along this path from v_1 to v_k we may traverse some edge ***in*** its given direction—then we call it a **forward edge** of our path—or ***opposite to*** its given direction—then we call it a **backward edge** of our path. In other words, our path consists of one-way streets, and forward edges (backward edges) are those that we travel *in the right direction* (*in the wrong direction*). Figure 494 shows a forward edge (u, v) and a backward edge (w, v) of a path $v_1 \to v_k$.

CAUTION! Each edge in a network has a given direction, *which we cannot change*. Accordingly, if (u, v) is a forward edge in a path $v_1 \to v_k$, then (u, v) can become a backward edge only in another path $x_1 \to x_j$ in which it is an edge and is traversed in the opposite direction as one goes from x_1 to x_j; see Fig. 495. Keep this in mind, to avoid misunderstandings.

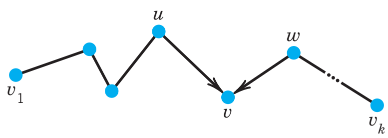


Fig. 494. Forward edge (u, v) and backward edge (w, v) of a path $v_1 \to v_k$

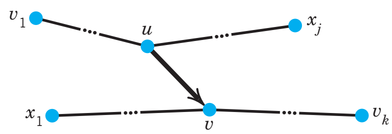


Fig. 495. Edge (u, v) as forward edge in the path $v_1 \to v_k$ and as backward edge in the path $x_1 \to x_j$

Flow Augmenting Paths

Our goal will be to ***maximize the flow*** from the source s to the target t of a given network. We shall do this by developing methods for increasing an existing flow (including the special case in which the latter is zero). The idea then is to find a path P: $s \to t$ all of whose edges are not fully used, so that we can push additional flow through P. This suggests the following concept.

DEFINITION

Flow Augmenting Path

A *flow augmenting path* in a network with a given flow f_{ij} on each edge (i, j) is a path $P: s \to t$ such that

(i) no forward edge is used to capacity; thus $f_{ij} < c_{ij}$ for these;

(ii) no backward edge has flow 0; thus $f_{ij} > 0$ for these.

EXAMPLE 1 **Flow Augmenting Paths**

Find flow augmenting paths in the network in Fig. 496, where the first number is the capacity and the second number a given flow.

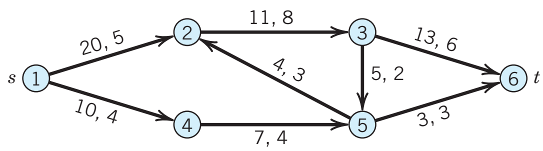


Fig. 496. Network in Example 1
First number = Capacity, Second number = Given flow

Solution. In practical problems, networks are large and one needs a ***systematic method for augmenting flows, which we discuss in the next section.*** In our small network, which should help to illustrate and clarify the concepts and ideas, we can find flow augmenting paths by inspection and augment the existing flow $f = 9$ in Fig. 496. (The outflow from s is $5 + 4 = 9$, which equals the inflow $6 + 3$ into t.)

We use the notation

$$\begin{aligned} \Delta_{ij} &= c_{ij} - f_{ij} && \text{for forward edges} \\ \Delta_{ij} &= f_{ij} && \text{for backward edges} \\ \Delta &= \min \Delta_{ij} && \text{taken over all edges of a path.} \end{aligned}$$

From Fig. 496 we see that a flow augmenting path $P_1: s \to t$ is $P_1: 1 - 2 - 3 - 6$ (Fig. 497), with $\Delta_{12} = 20 - 5 = 15$, etc., and $\Delta = 3$. Hence we can use P_1 to increase the given flow 9 to $f = 9 + 3 = 12$. All three edges of P_1 are forward edges. We augment the flow by 3. Then the flow in each of the edges of P_1 is increased by 3, so that we now have $f_{12} = 8$ (instead of 5), $f_{23} = 11$ (instead of 8), and $f_{36} = 9$ (instead of 6). Edge (2, 3) is now used to capacity. The flow in the other edges remains as before.

We shall now try to increase the flow in this network in Fig. 496 beyond $f = 12$.

There is another flow augmenting path $P_2: s \to t$, namely, $P_2: 1 - 4 - 5 - 3 - 6$ (Fig. 497). It shows how a backward edge comes in and how it is handled. Edge (3, 5) is a backward edge. It has flow 2, so that $\Delta_{36} = 2$. We compute $\Delta_{14} = 10 - 4 = 6$, etc. (Fig. 497) and $\Delta = 2$. Hence we can use P_2 for another augmentation to get $f = 12 + 2 = 14$. The new flow is shown in Fig. 498. No further augmentation is possible. We shall confirm later that $f = 14$ is maximum. ■

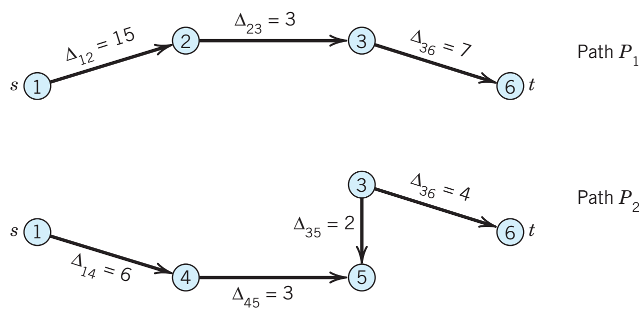


Fig. 497. Flow augmenting paths in Example 1

Cut Sets

A **cut set** is a set of edges in a network. The underlying idea is simple and natural. If we want to find out what is flowing from s to t in a network, we may cut the network somewhere between s and t (Fig. 498 shows an example) and see what is flowing in the edges hit by the cut, because any flow from s to t must sometimes pass through some of these edges. These form what is called a **cut set**. [In Fig. 498, the cut set consists of the edges (2, 3), (5, 2), (4, 5).] We denote this cut set by (S, T). Here S is the set of vertices on that side of the cut on which s lies ($S = \{s, 2, 4\}$ for the cut in Fig. 498) and T is the set of the other vertices ($T = \{3, 5, t\}$ in Fig. 498). We say that a cut ***partitions*** the vertex set V into two parts S and T. Obviously, the corresponding cut set (S, T) consists of all the edges in the network with one end in S and the other end in T.

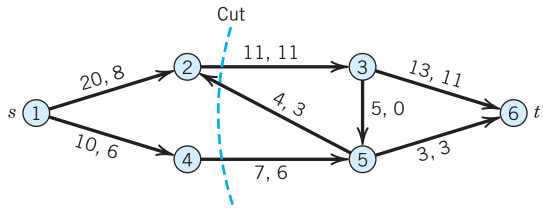


Fig. 498. Maximum flow in Example 1

By definition, the **capacity** cap (S, T) of a cut set (S, T) is the sum of the capacities of all **forward edges** in (S, T) (forward edges only!), that is, the edges that are directed *from S to T*,

$$\text{cap}\,(S, T) = \Sigma c_{ij} \qquad \text{[sum over the forward edges of } (S, T)]. \tag{3}$$

Thus, cap $(S, T) = 11 + 7 = 18$ in Fig. 498.

Explanation. This can be seen as follows. Look at Fig. 498. Recall that for each edge in that figure, the first number denotes capacity and the second number flow. Intuitively, you can think of the edges as roads, where the capacity of the road is how many cars can actually be on the road, and the flow denotes how many cars actually are on the road. To compute capacity cap (S, T) we are only looking at the first number on the edges. Take a look and see that the cut physically cuts three edges, that is, (2, 3), (4, 5), and (5, 2). *The cut concerns only forward edges* that are being cut, so it concerns edges (2, 3) and (4, 5) (and does not include edge (5, 2) which is also being cut, but since it goes backwards, it does not count). Hence (2, 3) contributes 11 and (4, 5) contributes 7 to the capacity cap (S, T), for a total of 18 in Fig. 498. Hence cap $(S, T) = 18$.

The other edges (directed *from T to S*) are called **backward edges** of the cut set (S, T), and by the **net flow** through a cut set we mean the sum of the flows in the forward edges minus the sum of the flows in the backward edges of the cut set.

CAUTION! Distinguish well between forward and backward edges in a cut set and in a path: (5, 2) in Fig. 498 is a backward edge for the cut shown but a forward edge in the path $1 - 4 - 5 - 2 - 3 - 6$.

For the cut in Fig. 498 the net flow is $11 + 6 - 3 = 14$. For the same cut in Fig. 496 (not indicated there), the net flow is $8 + 4 - 3 = 9$. In both cases it equals the flow f.

We claim that this is not just by chance, but cuts do serve the purpose for which we have introduced them:

THEOREM 1

Net Flow in Cut Sets

Any given flow in a network G is the net flow through any cut set (S, T) of G.

PROOF By Kirchhoff's law (2), multiplied by -1, at a vertex i we have

$$(4) \qquad \underbrace{\sum_j f_{ij}}_{\text{Outflow}} - \underbrace{\sum_l f_{li}}_{\text{Inflow}} = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

Here we can sum over j and l from 1 to n (= number of vertices) by putting $f_{ij} = 0$ for $j = i$ and also for edges without flow or nonexisting edges; hence we can write the two sums as one,

$$\sum_j (f_{ij} - f_{ji}) = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

We now sum over all i in S. Since s is in S, this sum equals f:

$$(5) \qquad \sum_{i \in S} \sum_{j \in V} (f_{ij} - f_{ji}) = f.$$

We claim that in this sum, only the edges belonging to the cut set contribute. Indeed, edges with both ends in T cannot contribute, since we sum only over i in S; but edges (i, j) with both ends in S contribute $+f_{ij}$ at one end and $-f_{ij}$ at the other, a total contribution of 0. Hence the left side of (5) equals the net flow through the cut set. By (5), this is equal to the flow f and proves the theorem. ■

This theorem has the following consequence, which we shall also need later in this section.

THEOREM 2

Upper Bound for Flows

A flow f in a network G cannot exceed the capacity of any cut set (S, T) in G.

PROOF By Theorem 1 the flow f equals the net flow through the cut set, $f = f_1 - f_2$, where f_1 is the sum of the flows through the forward edges and f_2 ($\geqq 0$) is the sum of the flows through the backward edges of the cut set. Thus $f \leqq f_1$. Now f_1 cannot exceed the sum of the capacities of the forward edges; but this sum equals the capacity of the cut set, by definition. Together, $f \leqq \text{cap}\,(S, T)$, as asserted. ■

Cut sets will now bring out the full importance of augmenting paths:

THEOREM 3

Main Theorem. Augmenting Path Theorem for Flows

A flow from s to t in a network G is maximum if and only if there does not exist a flow augmenting path $s \to t$ in G.

PROOF **(a)** If there is a flow augmenting path $P: s \to t$, we can use it to push through it an additional flow. Hence the given flow cannot be maximum.

(b) On the other hand, suppose that there is no flow augmenting path $s \to t$ in G. Let S_0 be the set of all vertices i (including s) such that there is a flow augmenting path $s \to i$, and let T_0 be the set of the other vertices in G. Consider any edge (i, j) with i in S_0 and j in T_0. Then we have a flow augmenting path $s \to i$ since i is in S_0, but $s \to i \to j$ is not flow augmenting because j is not in S_0. Hence we must have

$$f_{ij} = \begin{cases} c_{ij} \\ 0 \end{cases} \quad \text{if } (i, j) \text{ is a } \begin{cases} \text{forward} \\ \text{backward} \end{cases} \text{edge of the path } s \to i \to j. \tag{6}$$

Otherwise we could use (i, j) to get a flow augmenting path $s \to i \to j$. Now (S_0, T_0) defines a cut set (since t is in T_0; why?). Since by (6), forward edges are used to capacity and backward edges carry no flow, the net flow through the cut set (S_0, T_0) equals the sum of the capacities of the forward edges, which is cap (S_0, T_0) by definition. This net flow equals the given flow f by Theorem 1. Thus $f = \text{cap}\,(S_0, T_0)$. We also have $f \leqq \text{cap}\,(S_0, T_0)$ by Theorem 2. Hence f must be maximum since we have reached equality. ■

The end of this proof yields another basic result (by Ford and Fulkerson, *Canadian Journal of Mathematics* **8** (1956), 399–404), namely, the so-called

THEOREM 4

Max-Flow Min-Cut Theorem

The maximum flow in any network G equals the capacity of a "**minimum cut set**" (= a cut set of minimum capacity) *in G.*

PROOF We have just seen that $f = \text{cap}\,(S_0, T_0)$ for a maximum flow f and a suitable cut set (S_0, T_0). Now by Theorem 2 we also have $f \leqq \text{cap}\,(S, T)$ for this f and any cut set (S, T) in G. Together, $\text{cap}\,(S_0, T_0) \leqq \text{cap}\,(S, T)$. Hence (S_0, T_0) is a minimum cut set.

The existence of a maximum flow in this theorem follows for rational capacities from the algorithm in the next section and for arbitrary capacities from the Edmonds–Karp BFS also in that section. ■

The two basic tools in connection with networks are flow augmenting paths and cut sets. In the next section we show how flow augmenting paths can be used in an algorithm for maximum flows.

PROBLEM SET 23.6

1–6 CUT SETS, CAPACITY

Find T and cap (S, T) for:

1. Fig. 498, $S = \{1, 2, 4, 5\}$
2. Fig. 499, $S = \{1, 2, 3\}$
3. Fig. 498, $S = \{1, 2, 3\}$
4. Fig. 499, $S = \{1, 2\}$
5. Fig. 499, $S = \{1, 2, 4, 5\}$
6. Fig. 498, $S = \{1, 3, 5\}$

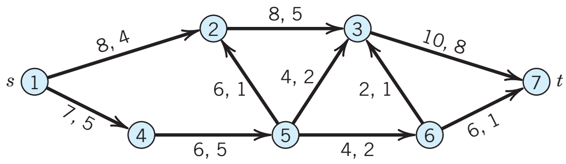


Fig. 499. Problems 2, 4, and 5

7–8 MINIMUM CUT SET

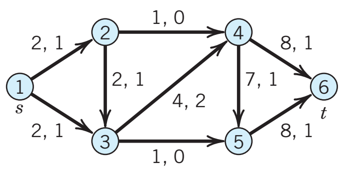
Find a minimum cut set and its capacity for the network:

7. In Fig. 499
8. In Fig. 496. Verify that its capacity equals the maximum flow.
9. Why are backward edges not considered in the definition of the capacity of a cut set?
10. **Incremental network.** Sketch the network in Fig. 499, and on each edge (i, j) write $c_{ij} - f_{ij}$ and f_{ij}. Do you recognize that from this "incremental network" one can more easily see flow augmenting paths?
11. **Omission of edges.** Which edges could be omitted from the network in Fig. 499 without decreasing the maximum flow?

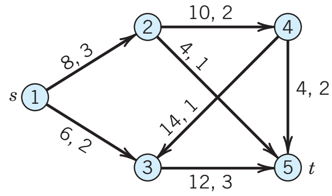
12–15 FLOW AUGMENTING PATHS

Find flow augmenting paths:

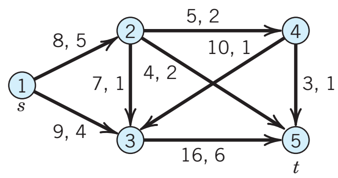
12.



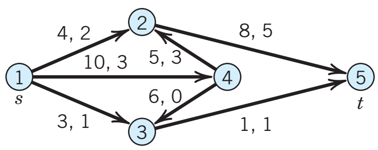
13.



14.



15.

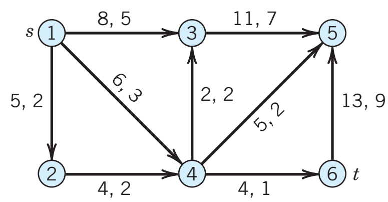


16–19 MAXIMUM FLOW

Find the maximum flow by inspection:

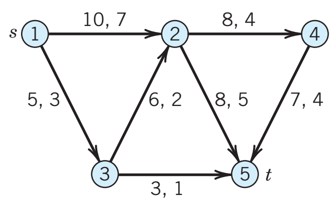
16. In Prob. 13

17.



18. In Prob. 12

19.



20. Find another maximum flow $f = 15$ in Prob. 19.

23.7 Maximum Flow: Ford–Fulkerson Algorithm

Flow augmenting paths, as discussed in the last section, are used as the basic tool in the Ford–Fulkerson[7] algorithm in Table 23.8 in which a given flow (for instance, zero flow in all edges) is increased until it is maximum. The algorithm accomplishes the increase by a stepwise construction of flow augmenting paths, one at a time, until no further such paths can be constructed, which happens precisely when the flow is maximum.

In Step 1, an initial flow may be given. In Step 3, a vertex j can be labeled if there is an edge (i, j) with i labeled and

$$c_{ij} > f_{ij} \qquad \textbf{(“forward edge”)}$$

or if there is an edge (j, i) with i labeled and

$$f_{ji} > 0 \qquad \textbf{(“backward edge”)}.$$

To **scan** a labeled vertex i means to label every unlabeled vertex j adjacent to i that can be labeled. Before scanning a labeled vertex i, scan all the vertices that got labeled before i. This **BFS** (**Breadth First Search**) strategy was suggested by Edmonds and Karp in 1972 (*Journal of the Association for Computing Machinery* **19**, 248–64). It has the effect that one gets shortest possible augmenting paths.

Table 23.8 Ford–Fulkerson Algorithm for Maximum Flow

Canadian Journal of Mathematics **9** (1957), 210–218

ALGORITHM FORD–FULKERSON

$[G = (V, E)$, vertices $1\ (= s), \cdots, n\ (= t)$, edges (i, j), $c_{ij}]$

This algorithm computes the maximum flow in a network G with source s, sink t, and capacities $c_{ij} > 0$ of the edges (i, j).

INPUT: n, $s = 1$, $t = n$, edges (i, j) of G, c_{ij}

OUTPUT: Maximum flow f in G

1. Assign an initial flow f_{ij} (for instance, $f_{ij} = 0$ for all edges), compute f.
2. Label s by $\varnothing$. Mark the other vertices *“unlabeled.”*
3. Find a labeled vertex i that has not yet been scanned. Scan i as follows. For every unlabeled adjacent vertex j, if $c_{ij} > f_{ij}$, compute

$$\Delta_{ij} = c_{ij} - f_{ij} \quad \text{and} \quad \Delta_j = \begin{cases} \Delta_{ij} & \text{if } i = 1 \\ \min(\Delta_i, \Delta_{ij}) & \text{if } i > 1 \end{cases}$$

and label j with a *“forward label”* (i^+, Δ_j); or if $f_{ji} > 0$, compute

$$\Delta_j = \min(\Delta_i, f_{ji})$$

and label j by a “backward label” (i^-, Δ_j).

[7]LESTER RANDOLPH FORD Jr. (1927–) and DELBERT RAY FULKERSON (1924–1976), American mathematicians known for their pioneering work on flow algorithms.

If no such j exists then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 4).

4. Repeat Step 3 until t is reached.

[*This gives a flow augmenting path P: $s \to t$.*]

If it is impossible to reach t then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 5).

5. Backtrack the path P, using the labels.

6. Using P, augment the existing flow by Δ_t. Set $f = f + \Delta_t$.

7. Remove all labels from vertices $2, \cdots, n$. Go to Step 3.

End FORD–FULKERSON

EXAMPLE 1 Ford–Fulkerson Algorithm

Applying the Ford–Fulkerson algorithm, determine the maximum flow for the network in Fig. 500 (which is the same as that in Example 1, Sec. 23.6, so that we can compare).

Solution. The algorithm proceeds as follows.

1. An initial flow $f = 9$ is given.

2. Label s $(= 1)$ by $\varnothing$. Mark 2, 3, 4, 5, 6 "unlabeled."

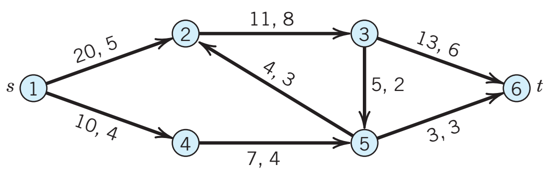


Fig. 500. Network in Example 1 with capacities (first numbers) and given flow

3. Scan 1.

Compute $\Delta_{12} = 20 - 5 = 15 = \Delta_2$. Label 2 by $(1^+, 15)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

Compute $\Delta_{23} = 11 - 8 = 3$, $\Delta_3 = \min(\Delta_2, 3) = 3$. Label 3 by $(2^+, 3)$.

Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

Scan 3.

Compute $\Delta_{36} = 13 - 6 = 7$, $\Delta_6 = \Delta_t = \min(\Delta_3, 7) = 3$. Label 6 by $(3^+, 3)$.

5. P: $1 - 2 - 3 - 6\ (= t)$ is a flow augmenting path.

6. $\Delta_t = 3$. Augmentation gives $f_{12} = 8$, $f_{23} = 11$, $f_{36} = 9$, other f_{ij} unchanged. Augmented flow $f = 9 + 3 = 12$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

3. Scan 1.

Compute $\Delta_{12} = 20 - 8 = 12 = \Delta_2$. Label 2 by $(1^+, 12)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

   Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

   Scan 4. [*No vertex left for labeling.*]

   Scan 5.

   Compute $\Delta_3 = \min(\Delta_5, 2) = 2$. Label 3 by $(5^-, 2)$.

   Scan 3.

   Compute $\Delta_{36} = 13 - 9 = 4$, $\Delta_6 = \min(\Delta_3, 4) = 2$. Label 6 by $(3^+, 2)$.

5. P: $1 - 2 - 5 - 3 - 6 \,(= t)$ is a flow augmenting path.

6. $\Delta_t = 2$. Augmentation gives $f_{12} = 10, f_{32} = 1, f_{35} = 0, f_{36} = 11$, other f_{ij} unchanged. Augmented flow $f = 12 + 2 = 14$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

One can now scan 1 and then scan 2, as before, but in scanning 4 and then 5 one finds that no vertex is left for labeling. Thus one can no longer reach t. Hence the flow obtained (Fig. 501) is maximum, in agreement with our result in the last section.

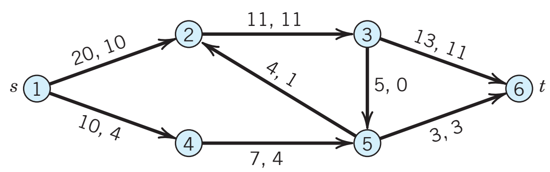


Fig. 501. Maximum flow in Example 1

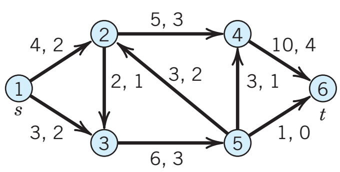
PROBLEM SET 23.7

1. Do the computations indicated near the end of Example 1 in detail.
2. Solve Example 1 by Ford–Fulkerson with initial flow 0. Is it more work than in Example 1?
3. Which are the "bottleneck" edges by which the flow in Example 1 is actually limited? Hence which capacities could be decreased without decreasing the maximum flow?
4. What is the (simple) reason that Kirchhoff's law is preserved in augmenting a flow by the use of a flow augmenting path?
5. How does Ford–Fulkerson prevent the formation of cycles?

6–9 MAXIMUM FLOW

Find the maximum flow by Ford-Fulkerson:

6. In Prob. 12, Sec. 23.6
7. In Prob. 15, Sec. 23.6
8. In Prob. 14, Sec. 23.6
9.



10. **Integer flow theorem.** Prove that, if the capacities in a network G are integers, then a maximum flow exists and is an integer.
11. **CAS PROBLEM. Ford–Fulkerson.** Write a program and apply it to Probs. 6–9.
12. How can you see that Ford–Fulkerson follows a BFS technique?
13. Are the consecutive flow augmenting paths produced by Ford–Fulkerson unique?
14. If the Ford–Fulkerson algorithm stops without reaching t, show that the edges with one end labeled and the other end unlabeled form a cut set (S, T) whose capacity equals the maximum flow.
15. Find a minimum cut set in Fig. 500 and its capacity.
16. Show that in a network G with all $c_{ij} = 1$, the maximum flow equals the number of edge-disjoint paths $s \to t$.
17. In Prob. 15, the cut set contains precisely all forward edges used to capacity by the maximum flow (Fig. 501). Is this just by chance?
18. Show that in a network G with capacities all equal to 1, the capacity of a minimum cut set (S, T) equals the minimum number q of edges whose deletion destroys all directed paths $s \to t$. (A **directed path** $v \to w$ is a path in which each edge has the direction in which it is traversed in going from v to w.)

19. Several sources and sinks. If a network has several sources $s_1, \cdots, s_k$, show that it can be reduced to the case of a single-source network by introducing a new vertex s and connecting s to $s_1, \cdots, s_k$ by k edges of capacity ∞. Similarly if there are several sinks. Illustrate this idea by a network with two sources and two sinks.

20. Find the maximum flow in the network in Fig. 502 with two sources (factories) and two sinks (consumers).

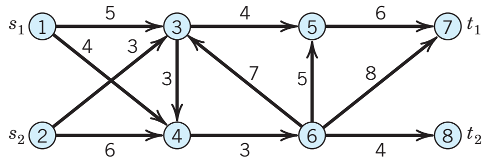


Fig. 502. Problem 20

23.8 Bipartite Graphs. Assignment Problems

From digraphs we return to graphs and discuss another important class of combinatorial optimization problems that arises in **assignment problems** of workers to jobs, jobs to machines, goods to storage, ships to piers, classes to classrooms, exams to time periods, and so on. To explain the problem, we need the following concepts.

A **bipartite graph** $G = (V, E)$ is a graph in which the vertex set V is partitioned into two sets S and T (without common elements, by the definition of a partition) such that every edge of G has one end in S and the other in T. Hence there are no edges in G that have both ends in S or both ends in T. Such a graph $G = (V, E)$ is also written $G = (S, T; E)$.

Figure 503 shows an illustration. V consists of seven elements, three workers a, b, c, making up the set S, and four jobs 1, 2, 3, 4, making up the set T. The edges indicate that worker a can do the jobs 1 and 2, worker b the jobs 1, 2, 3, and worker c the job 4. The problem is to assign one job to each worker so that every worker gets one job to do. This suggests the next concept, as follows.

DEFINITION

Maximum Cardinality Matching

A **matching** in $G = (S, T; E)$ is a set M of edges of G such that no two of them have a vertex in common. If M consists of the greatest possible number of edges, we call it a **maximum cardinality matching** in G.

For instance, a matching in Fig. 503 is $M_1 = \{(a, 2), (b, 1)\}$. Another is $M_2 = \{(a, 1), (b, 3), (c, 4)\}$; obviously, this is of maximum cardinality.

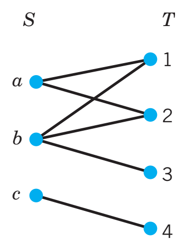


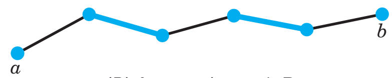
Fig. 503. Bipartite graph in the assignment of a set $S = \{a, b, c\}$ of workers to a set $T = \{1, 2, 3, 4\}$ of jobs

A vertex v is **exposed** (or *not covered*) by a matching M if v is not an endpoint of an edge of M. This concept, which always refers to some matching, will be of interest when we begin to augment given matchings (below). If a matching leaves no vertex exposed, we call it a **complete matching**. Obviously, a complete matching can exist only if S and T consist of the same number of vertices.

We now want to show how one can stepwise increase the cardinality of a matching M until it becomes maximum. Central in this task is the concept of an augmenting path.

An **alternating path** is a path that consists alternately of edges in M and not in M (Fig. 504A). An **augmenting path** is an alternating path both of whose endpoints (a and b in Fig. 504B) are exposed. By dropping from the matching M the edges that are on an augmenting path P (two edges in Fig. 504B) and adding to M the other edges of P (three in the figure), we get a new matching, with one more edge than M. This is how we use an augmenting path in ***augmenting a given matching*** by one edge. We assert that this will always lead, after a number of steps, to a maximum cardinality matching. Indeed, the basic role of augmenting paths is expressed in the following theorem.

(A) Alternating path



(B) Augmenting path P

Fig. 504. Alternating and augmenting paths. Heavy edges are those belonging to a matching M

THEOREM 1

Augmenting Path Theorem for Bipartite Matching

A matching M in a bipartite graph $G = (S, T; E)$ is of maximum cardinality if and only if there does not exist an augmenting path P with respect to M.

PROOF **(a)** We show that if such a path P exists, then M is not of maximum cardinality. Let P have q edges belonging to M. Then P has $q + 1$ edges not belonging to M. (In Fig. 504B we have $q = 2$.) The endpoints a and b of P are exposed, and all the other vertices on P are endpoints of edges in M, by the definition of an alternating path. Hence if an edge of M is not an edge of P, it cannot have an endpoint on P since then M would not be a matching. Consequently, the edges of M not on P, together with the $q + 1$ edges of P not belonging to M form a matching of cardinality one more than the cardinality of M because we omitted q edges from M and added $q + 1$ instead. Hence M cannot be of maximum cardinality.

(b) We now show that if there is no augmenting path for M, then M is of maximum cardinality. Let M^* be a maximum cardinality matching and consider the graph H consisting of all edges that belong either to M or to M^*, but not to both. Then it is possible that two edges of H have a vertex in common, but three edges cannot have a vertex in common since then two of the three would have to belong to M (or to M^*), violating that M and M^* are matchings. So every v in V can be in common with two edges of H or with one or none. Hence we can characterize each "component" (= maximal *connected* subset) of H as follows.

(A) A component of H can be a closed path with an *even* number of edges (in the case of an *odd* number, two edges from M or two from M^* would meet, violating the matching property). See (A) in Fig. 505.

(B) A component of H can be an open path P with the same number of edges from M and edges from M^*, for the following reason. P must be alternating, that is, an edge of M is followed by an edge of M^*, etc. (since M and M^* are matchings). Now if P had an edge more from M^*, then P would be augmenting for M [see (B2) in Fig. 505], contradicting our assumption that there is no augmenting path for M. If P had an edge more from M, it would be augmenting for M^* [see (B3) in Fig. 505], violating the maximum cardinality of M^*, by part (a) of this proof. Hence in each component of H, the two matchings have the same number of edges. Adding to this the number of edges that belong to both M and M^* (which we left aside when we made up H), we conclude that M and M^* must have the same number of edges. Since M^* is of maximum cardinality, this shows that the same holds for M, as we wanted to prove. ■

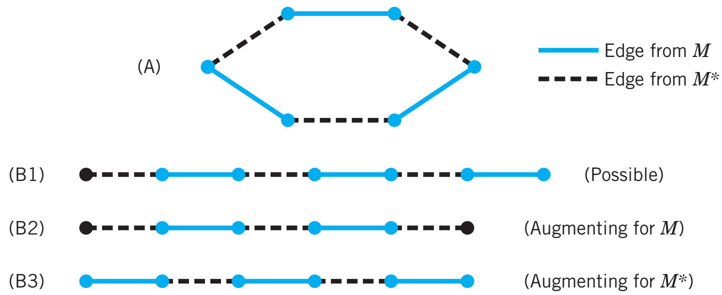


Fig. 505. Proof of the augmenting path theorem for bipartite matching

This theorem suggests the algorithm in Table 23.9 for obtaining augmenting paths, in which vertices are labeled for the purpose of backtracking paths. Such a label is *in addition* to the number of the vertex, which is also retained. Clearly, to get an augmenting path, one must start from an *exposed* vertex, and then trace an alternating path until one arrives at another *exposed* vertex. After Step 3 all vertices in S are labeled. In Step 4, the set T contains at least one exposed vertex, since otherwise we would have stopped at Step 1.

Table 23.9 Bipartite Maximum Cardinality Matching

ALGORITHM MATCHING [$G = (S, T; E), M, n$]

This algorithm determines a maximum cardinality matching M in a bipartite graph G by augmenting a given matching in G.

INPUT: Bipartite graph $G = (S, T; E)$ with vertices $1, \cdots, n$, matching M in G (for instance, $M = \varnothing$)

OUTPUT: Maximum cardinality matching M in G

1. If there is no exposed vertex in S then

OUTPUT M. Stop

[*M is of maximum cardinality in G.*]

Else label all *exposed* vertices *in* S with $\varnothing$.

2. For each i in S and edge (i, j) *not* in M, label j with i, unless already labeled.

3. For each *nonexposed* j in T, label i with j, where i is the other end of the unique edge (i, j) in M.

4. Backtrack the alternating path P ending on an exposed vertex in T by using the labels on the vertices.

5. If no P in Step 4 is augmenting then

OUTPUT M. Stop

[*M is of maximum cardinality in G.*]

Else augment M by using an augmenting path P.

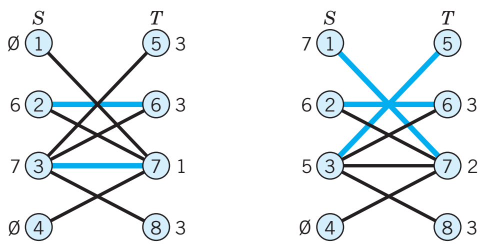
Remove all labels.

Go to Step 1.

End MATCHING

EXAMPLE 1 **Maximum Cardinality Matching**

Is the matching M_1 in Fig. 506a of maximum cardinality? If not, augment it until maximum cardinality is reached.



(a) Given graph and matching M_1

(b) Matching M_2 and new labels

Fig. 506. Example 1

Solution. We apply the algorithm.

1. Label 1 and 4 with $\varnothing$.
2. Label 7 with 1. Label 5, 6, 8 with 3.
3. Label 2 with 6, and 3 with 7.

   [*All vertices are now labeled as shown in Fig.* 506a.]
4. P_1: $1 - 7 - 3 - 5$. [*By backtracking, P_1 is augmenting.*]

   P_2: $1 - 7 - 3 - 8$. [*P_2 is augmenting.*]
5. Augment M_1 by using P_1, dropping (3, 7) from M_1 and including (1, 7) and (3, 5). Remove all labels. Go to Step 1.

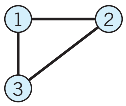
   Figure 506b *shows the resulting matching* $M_2 = \{(1, 7), (2, 6), (3, 5)\}$.
1. Label 4 with $\varnothing$.
2. Label 7 with 2. Label 6 and 8 with 3.
3. Label 1 with 7, and 2 with 6, and 3 with 5.
4. P_3: $5 - 3 - 8$. [*P_3 is alternating but not augmenting.*]
5. Stop. *M_2 is of maximum cardinality* (*namely*, 3).

PROBLEM SET 23.8

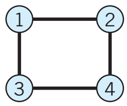
1–7 BIPARTITE OR NOT?

If you answer is yes, find S and T:

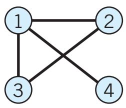
1.



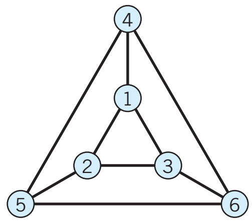
2.



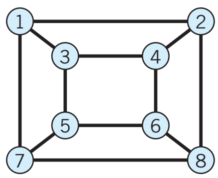
3.



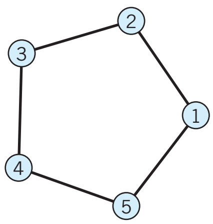
4.



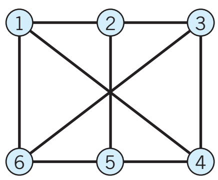
5.



6.



7.



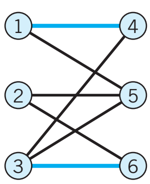
8. Can you obtain the answer to Prob. 3 from that to Prob. 1?

9. Can you obtain a bipartite subgraph in Prob. 4 by omitting two edges? Any two edges? Any two edges without a common vertex?

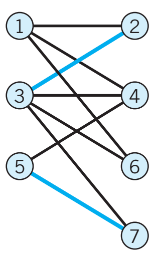
10–12 MATCHING. AUGMENTING PATHS

Find an augmenting path:

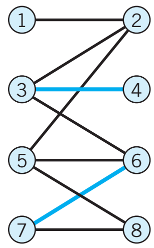
10.



11.



12.



13–15 MAXIMUM CARDINALITY MATCHING

Using augmenting paths, find a maximum cardinality matching:

13. In Prob. 11

14. In Prob. 10

15. In Prob. 12

16. Complete bipartite graphs. A bipartite graph $G = (S, T; E)$ is called *complete* if every vertex in S is joined to every vertex in T by an edge, and is denoted by K_{n_1,n_2}, where n_1 and n_2 are the numbers of vertices in S and T, respectively. How many edges does this graph have?

17. Planar graph. A *planar graph* is a graph that can be drawn on a sheet of paper so that no two edges cross. Show that the complete graph K_4 with four vertices is planar. The complete graph K_5 with five vertices is not planar. Make this plausible by attempting to draw K_5 so that no edges cross. Interpret the result in terms of a net of roads between five cities.

18. Bipartite graph $K_{3,3}$ not planar. Three factories 1, 2, 3 are each supplied underground by water, gas, and electricity, from points A, B, C, respectively. Show that this can be represented by $K_{3,3}$ (the complete bipartite graph $G = (S, T; E)$ with S and T consisting of three vertices each) and that eight of the nine supply lines (edges) can be laid out without crossing. Make it plausible that $K_{3,3}$ is not planar by attempting to draw the ninth line without crossing the others.

19–25 VERTEX COLORING

19. Vertex coloring and exam scheduling. What is the smallest number of exam periods for six subjects a, b, c, d, e, f if some of the students simultaneously take a, b, f, some c, d, e, some a, c, e, and some c, e? Solve this as follows. Sketch a graph with six vertices $a, \cdots, f$ and join vertices if they represent subjects simultaneously taken by some students. Color the vertices so that adjacent vertices receive different colors. (Use numbers $1, 2, \cdots$ instead of actual colors if you want.) What is the minimum number of colors you need? For any graph G, this minimum number is called the (vertex) **chromatic number** $\chi_v(G)$. Why is this the answer to the problem? Write down a possible schedule.

20. Scheduling and matching. Three teachers x_1, x_2, x_3 teach four classes y_1, y_2, y_3, y_4 for these numbers of periods:

	y_1	y_2	y_3	y_4
x_1	1	0	1	1
x_2	1	1	1	1
x_3	0	1	1	1

Show that this arrangement can be represented by a bipartite graph G and that a teaching schedule for one period corresponds to a matching in G. Set up a teaching schedule with the smallest possible number of periods.

21. How many colors do you need for vertex coloring any tree?

22. Harbor management. How many piers does a harbor master need for accommodating six cruise ships $S_1, \cdots, S_6$ with expected dates of arrival A and departure D in July, $(A, D) = (10, 13)$, $(13, 15)$, $(14, 17)$, $(12, 15)$, $(16, 18)$, $(14, 17)$, respectively, if each pier can accommodate only one ship, arrival being at 6 am and departures at 11 pm? *Hint.* Join S_i and S_j by an edge if their intervals overlap. Then color vertices.

23. What would be the answer to Prob. 22 if only the five ships $S_1, \cdots, S_5$ had to be accommodated?

24. Four- (vertex) color theorem. The famous *four-color theorem* states that one can color the vertices of any *planar* graph (so that adjacent vertices get different colors) with at most four colors. It had been conjectured for a long time and was eventually proved in 1976 by Appel and Haken [*Illinois J. Math* **21** (1977), 429–567]. Can you color the complete graph K_5 with four colors? Does the result contradict the four-color theorem? (For more details, see Ref. [F1] in App. 1.)

25. Find a graph, as simple as possible, that cannot be vertex colored with three colors. Why is this of interest in connection with Prob. 24?

26. Edge coloring. The *edge chromatic number* $\chi_e(G)$ of a graph G is the minimum number of colors needed for coloring the edges of G so that incident edges get different colors. Clearly, $\chi_e(G) \geqq \max d(u)$, where $d(u)$ is the degree of vertex u. If $G = (S, T; E)$ is bipartite, the equality sign holds. Prove this for $K_{n,n}$ the complete (cf. Sec. 23.1) bipartite graph $G = (S, T, E)$ with S and T consisting of n vertices each.

CHAPTER 23 REVIEW QUESTIONS AND PROBLEMS

1. What is a graph, a digraph, a cycle, a tree?

2. State some typical problems that can be modeled and solved by graphs or digraphs.

3. State from memory how graphs can be handled on computers.

4. What is a shortest path problem? Give applications.

5. What situations can be handled in terms of the traveling salesman problem?

6. Give typical applications involving spanning trees.

7. What are the basic ideas and concepts in handling flows?

8. What is combinatorial optimization? Which sections of this chapter involved it? Explain details.

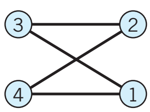
9. Define bipartite graphs and describe some typical applications of them.

10. What is BFS? DFS? In what connection did these concepts occur?

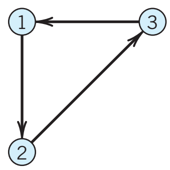
11–16 MATRICES FOR GRAPHS AND DIGRAPHS

Find the adjacency matrix of:

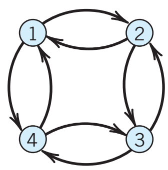
11.



12.



13.



14–16 Sketch the graph whose adjacency matrix is:

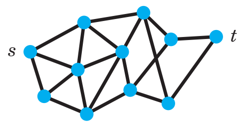
14. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

15. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

16. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

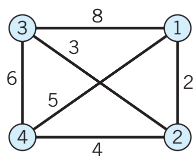
17. Vertex incidence list. Make it for the graph in Prob. 15.

18. Find a shortest path and its length by Moore's BFS algorithm, assuming that all the edges have length 1.



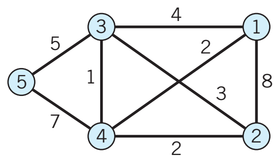
Problem 18

19. Find shortest paths by Dijkstra's algorithm.



Problem 19

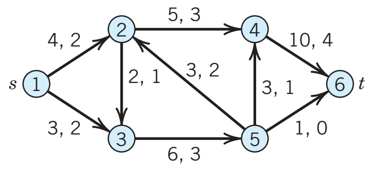
20. Find a shortest spanning tree.



Problem 20

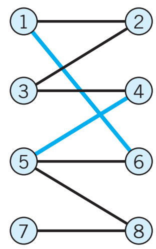
21. Company A has offices in Chicago, Los Angeles, and New York; Company B in Boston and New York; Company C in Chicago, Dallas, and Los Angeles. Represent this by a bipartite graph.

22. Find flow augmenting paths and the maximum flow.



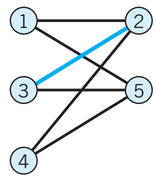
Problem 22

23. Using augmenting paths, find a maximum cardinality matching.



Problem 25

24. Find an augmenting path,



Problem 24

SUMMARY OF CHAPTER 23
Graphs. Combinatorial Optimization

Combinatorial optimization concerns optimization problems of a discrete or combinatorial structure. It uses graphs and digraphs (Sec. 23.1) as basic tools.

A **graph** $G = (V, E)$ consists of a set V of **vertices** $v_1, v_2, \cdots, v_n$ (often simply denoted by $1, 2, \cdots, n$) and a set E of **edges** $e_1, e_2, \cdots, e_m$, each of which connects two vertices. We also write (i, j) for an edge with vertices i and j as endpoints. A **digraph** (= directed graph) is a graph in which each edge has a direction (indicated by an arrow). For handling graphs and digraphs in computers, one can use *matrices* or *lists* (Sec. 23.1).

This chapter is devoted to important classes of optimization problems for graphs and digraphs that all arise from practical applications, and corresponding algorithms, as follows.

In a **shortest path problem** (Sec. 23.2) we determine a path of minimum length (consisting of edges) from a vertex s to a vertex t in a graph whose edges (i, j) have a "length" $l_{ij} > 0$, which may be an actual length or a travel time or cost or an electrical resistance [if (i, j) is a wire in a net], and so on. ***Dijkstra's algorithm*** (Sec. 23.3) or, when all $l_{ij} = 1$, ***Moore's algorithm*** (Sec. 23.2) are suitable for these problems.

A **tree** is a graph that is connected and has no **cycles** (no closed paths). Trees are very important in practice. A *spanning tree* in a graph G is a tree containing *all* the vertices of G. If the edges of G have lengths, we can determine a **shortest spanning tree**, for which the sum of the lengths of all its edges is minimum, by ***Kruskal's algorithm*** or ***Prim's algorithm*** (Secs. 23.4, 23.5).

A **network** (Sec. 23.6) is a digraph in which each edge (i, j) has a *capacity* $c_{ij} > 0$ [= maximum possible flow along (i, j)] and at one vertex, the *source* s, a flow is produced that flows along the edges to a vertex t, the *sink* or *target,* where the flow disappears. The problem is to maximize the flow, for instance, by applying the **Ford–Fulkerson algorithm** (Sec. 23.7), which uses ***flow augmenting paths*** (Sec. 23.6). Another related concept is that of a ***cut set***, as defined in Sec. 23.6.

A **bipartite graph** $G = (V, E)$ (Sec. 23.8) is a graph whose vertex set V consists of two parts S and T such that every edge of G has one end in S and the other in T, so that there are no edges connecting vertices in S or vertices in T. A **matching** in G is a set of edges, no two of which have an endpoint in common. The problem then is to find a **maximum cardinality matching** in G, that is, a matching M that has a maximum number of edges. For an algorithm, see Sec. 23.8.

PART G

Probability, Statistics

Probability theory (Chap. 24) provides models of probability distributions (theoretical models of the observable reality involving chance effects) to be tested by statistical methods, and it will also supply the mathematical foundation of these methods in Chap. 25.

Modern **mathematical statistics** (Chap. 25) has various engineering applications, for instance, in testing materials, control of production processes, quality control of production outputs, performance tests of systems, robotics, and automatization in general, production planning, marketing analysis, and so on.

To this we could add a long list of fields of applications, for instance, in agriculture, biology, computer science, demography, economics, geography, management of natural resources, medicine, meteorology, politics, psychology, sociology, traffic control, urban planning, etc. Although these applications are very heterogeneous, we shall see that most statistical methods are universal in the sense that each of them can be applied in various fields.

Additional Software for Probability and Statistics

See also the list of software at the beginning of Part E on Numerical Analysis.

Data Desk. Data Description, Inc., Ithaca, NY. Phone 1-800-573-5121 or (607) 257-1000, website at www.datadesk.com.

MINITAB. Minitab, Inc., State College, PA. Phone 1-800-448-3555 or (814) 238-3280, website at www.minitab.com.

SAS. SAS Institute, Inc., Cary, NC. Phone 1-800-727-0025 or (919) 677-8000, website at www.sas.com.

R. website at www.r-project.org. Free software, part of the GNU/Free Software Foundation project.

SPSS. SPSS, Inc., Chicago, IL. (part of IBM) Phone 1-800-543-2185 or (312) 651-3000, website at www.spss.com.

STATISTICA. StatSoft, Inc., Tulsa, OK. Phone (918) 749-1119, website at www.statsoft.com.

TIBCO Spotfire S+. TIBCO Software Inc., Palo Alto, CA; Office for this software: Somerville, MA. Phone 1-866-240-0491 (toll-free), (617) 702-1602, website at spotfire.tibco.com/products/s-plus/statistical-analysis-software.aspx

CHAPTER 24

Data Analysis. Probability Theory

We first show how to handle data numerically or in terms of graphs, and how to extract information (average size, spread of data, etc.) from them. If these data are influenced by "chance," by factors whose effect we cannot predict exactly (e.g., weather data, stock prices, life spans of tires, etc.), we have to rely on **probability theory**. This theory originated in games of chance, such as flipping coins, rolling dice, or playing cards. Nowadays it gives mathematical models of chance processes called *random experiments* or, briefly, **experiments**. In such an experiment we observe a **random variable** X, that is, a function whose values in a **trial** (a performance of an experiment) occur "by chance" (Sec. 24.3) according to a **probability distribution** that gives the individual probabilities with which possible values of X may occur in the long run. (Example: Each of the six faces of a die should occur with the same probability, $1/6$.) Or we may simultaneously observe more than one random variable, for instance, height *and* weight of persons or hardness *and* tensile strength of steel. This is discussed in Sec. 24.9, which will also give the basis for the mathematical justification of the statistical methods in Chapter 25.

Prerequisite: Calculus.
References and Answers to Problems: App. 1 Part G, App. 2.

24.1 Data Representation. Average. Spread

Data can be represented numerically or graphically in various ways. For instance, your daily newspaper may contain tables of stock prices and money exchange rates, curves or bar charts illustrating economical or political developments, or pie charts showing how your tax dollar is spent. And there are numerous other representations of data for special purposes.

In this section we discuss the use of standard representations of data in statistics. (For these, software packages, such as DATA DESK, R, and MINITAB, are available, and Maple or Mathematica may also be helpful; see pp. 789 and 1009) We explain corresponding concepts and methods in terms of typical examples.

EXAMPLE 1 **Recording and Sorting**

Sample values (observations, measurements) should be **recorded** in the order in which they occur. **Sorting**, that is, ordering the sample values by size, is done as a first step of investigating properties of the sample and graphing it. Sorting is a standard process on the computer; see Ref. [E35], listed in App. 1.

Super alloys is a collective name for alloys used in jet engines and rocket motors, requiring high temperature (typically 1800°F), high strength, and excellent resistance to oxidation. Thirty specimens of Hastelloy C (nickel-based steel, investment cast) had the tensile strength (in 1000 lb/sq in.), recorded in the order obtained and rounded to integer values,

(1)	89	77	88	91	88	93	99	79	87	84	86	82	88	89	78
	90	91	81	90	83	83	92	87	89	86	89	81	87	84	89

Sorting gives

(2)	77	78	79	81	81	82	83	83	84	84	86	86	87	87	87
	88	88	88	89	89	89	89	89	90	90	91	91	92	93	99

Graphic Representation of Data

We shall now discuss standard graphic representations used in statistics for obtaining information on properties of data.

EXAMPLE 2 Stem-and-Leaf Plot (Fig. 507)

This is one of the simplest but most useful representations of data. For (1) it is shown in Fig. 507. The numbers in (1) range from 78 to 99; see (2). We divide these numbers into 5 groups, 75–79, 80–84, 85–89, 90–94, 95–99. The integers in the tens position of the groups are 7, 8, 8, 9, 9. These form the *stem* in Fig. 507. The first *leaf* is 789, representing 77, 78, 79. The second leaf is 1123344, representing 81, 81, 82, 83, 83, 84, 84. And so on.

The number of times a value occurs is called its **absolute frequency**. Thus 78 has absolute frequency 1, the value 89 has absolute frequency 5, etc. The column to the extreme left in Fig. 507 shows the **cumulative absolute frequencies**, that is, the sum of the absolute frequencies of the values up to the line of the leaf. Thus, the number 10 in the second line on the left shows that (1) has 10 values up to and including 84. The number 23 in the next line shows that there are 23 values not exceeding 89, etc. Dividing the cumulative absolute frequencies by n (= 30 in Fig. 507) gives the **cumulative relative frequencies** 0.1, 0.33, 0.76, 0.93, 1.00.

EXAMPLE 3 Histogram (Fig. 508)

For large sets of data, histograms are better in displaying the distribution of data than stem-and-leaf plots. The principle is explained in Fig. 508. (An application to a larger data set is shown in Sec. 25.7). The bases of the rectangles in Fig. 508 are the x-intervals (known as **class intervals**) 74.5–79.5, 79.5–84.5, 84.5–89.5, 89.5–94.5, 94.5–99.5, whose midpoints (known as **class marks**) are $x = 77, 82, 87, 92, 97$, respectively. The height of a rectangle with class mark x is the **relative class frequency** $f_{\text{rel}}(x)$, defined as the number of data values in that class interval, divided by n (= 30 in our case). Hence the areas of the rectangles are proportional to these relative frequencies, 0.10, 0.23, 0.43, 0.17, 0.07, so that histograms give a good impression of the distribution of data.

Leaf unit = 1.0

3	7	789
10	8	1123344
23	8	6677788899999
29	9	001123
30	9	9

Fig. 507. Stem-and-leaf plot of the data in Example 1

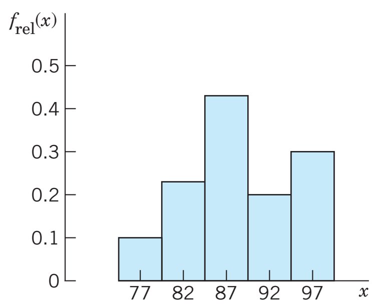


Fig. 508. Histogram of the data in Example 1 (grouped as in Fig. 507)

EXAMPLE 4 Boxplot. Median. Interquartile Range. Outlier

A **boxplot** of a set of data illustrates the average size and the spread of the values, in many cases the two most important quantities characterizing the set, as follows.

The average size is measured by the **median**, or *middle quartile,* q_M. If the number n of values of the set is *odd*, then q_M is the middlemost of the values when ordered as in (2). If n is *even,* then q_M is the average of the two middlemost values of the ordered set. In (2) we have $n = 30$ and thus $q_M = \frac{1}{2}(x_{15} + x_{16}) = \frac{1}{2}(87 + 88) = 87.5$. (In general, q_M will be a fraction if n is even.)

The spread of values can be measured by the **range** $R = x_{\max} - x_{\min}$, the largest value minus the smallest one.

Better information on the spread gives the **interquartile range** $\text{IQR} = q_U - q_L$. Here q_U is the middlemost value (or the average of the two middlemost values) in the data *above* the median; and q_L is the middlemost value (or the average of the two middlemost values) in the data *below* the median. Hence in (2) we have $q_U = x_{23} = 89$, $q_L = x_8 = 83$, and $\text{IQR} = 89 - 83 = 6$.

The box in Fig. 509 extends vertically from q_L to q_U; it has height $\text{IQR} = 6$. The vertical lines below and above the box extend from $x_{\min} = 77$ to $x_{\max} = 99$, so that they show $R = 22$.

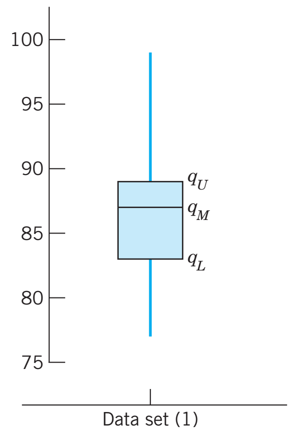


Fig. 509. Boxplot of the data set (1)

The line above the box is suspiciously long. This suggests the concept of an **outlier**, a value that is more than 1.5 times the IQR away from either end of the box; here 1.5 is purely conventional. An outlier indicates that something might have gone wrong in the data collection. In (2) we have $89 + 1.5\ \text{IQR} = 98$, and we regard 99 as an outlier. ■

Mean. Standard Deviation. Variance. Empirical Rule

Medians and quartiles are easily obtained by ordering and counting, practically without calculation. But they do not give full information on data: you can change data values to some extent without changing the median. Similarly for the quartiles.

The average size of the data values can be measured in a more refined way by the **mean**

$$\textbf{(3)} \qquad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

This is the arithmetic mean of the data values, obtained by taking their sum and dividing by the data *size* n. Thus in (1),

$$\bar{x} = \frac{1}{30}(89 + 77 + \cdots + 89) = \frac{260}{3} \approx 86.7.$$

Every data value contributes, and changing one of them will change the mean.

Similarly, the spread (variability) of the data values can be measured in a more refined way by the **standard deviation** s or by its square, the **variance**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]. \tag{4}$$

Thus, to obtain the variance of the data, take the difference $x_j - \bar{x}$ of each data value from the mean, square it, take the sum of these n squares, and divide it by $n - 1$ (not n, as we motivate in Sec. 25.2). To get the standard deviation s, take the square root of s^2.

For example, using $\bar{x} = 260/3$, we get for the data (1) the variance

$$s^2 = \frac{1}{29}[(89 - \tfrac{260}{3})^2 + (77 - \tfrac{260}{3})^2 + \cdots + (89 - \tfrac{260}{3})^2] = \frac{2006}{87} \approx 23.06$$

Hence the standard deviation is $s = \sqrt{2006/87} \approx 4.802$. Note that the standard deviation has the same dimension as the data values (kg/mm^2, see at the beginning), which is an advantage. On the other hand, the variance is preferable to the standard deviation in developing statistical methods, as we shall see in Chap. 25.

CAUTION! Your CAS (Maple, for instance) may use $1/n$ instead of $1/(n-1)$ in (4), but the latter is better when n is small (see Sec. 25.2).

Mean and standard deviation, introduced to give center and spread, actually give much more information according to this rule.

Empirical Rule. For any mound-shaped, nearly symmetric distribution of data the intervals

$$\bar{x} \pm s, \quad \bar{x} \pm 2s, \quad \bar{x} \pm 3s \qquad \text{contain about} \qquad 68\%, \quad 95\%, \quad 99.7\%,$$

respectively, of the data points.

EXAMPLE 5 Empirical Rule and Outliers. *z*-Score

For (1), with $\bar{x} = 86.7$ and $s = 4.8$, the three intervals in the Rule are $81.9 \leqq x \leqq 91.5$, $77.1 \leqq x \leqq 96.3$, $72.3 \leqq x \leqq 101.1$ and contain 73% (22 values remain, 5 are too small, and 5 too large), 93% (28 values, 1 too small, and 1 too large), and 100%, respectively.

If we reduce the sample by omitting the outlier 99, mean and standard deviation reduce to $\bar{x}_{\text{red}} = 86.2$, $s_{\text{red}} = 4.3$, approximately, and the percentage values become 67% (5 and 5 values outside), 93% (1 and 1 outside), and 100%.

Finally, the relative position of a value x in a set of mean $\bar{x}$ and standard deviation s can be measured by the **z-score**

$$z(s) = \frac{x - \bar{x}}{s}.$$

This is the distance of x from the mean $\bar{x}$ measured in multiples of s. For instance, $z(83) = (83 - 86.7)/4.8 = -0.77$. This is negative because 83 lies below the mean. By the Empirical Rule, the extreme z-values are about -3 and 3. ■

PROBLEM SET 24.1

1–10 DATA REPRESENTATIONS

Represent the data by a stem-and-leaf plot, a histogram, and a boxplot:

1. Length of nails [mm]

19 21 19 20 19 20 21 20

2. Phone calls per minute in an office between 9:00 A.M. and 9:10 A.M.

6 6 4 2 1 7 0 4 6 7

3. Systolic blood pressure of 15 female patients of ages 20–22

156 158 154 133 141 130 144 137
151 146 156 138 138 149 139

4. Iron content [%] of 15 specimens of hermatite (Fe_2O_3)

72.8 70.4 71.2 69.2 70.3 68.9 71.1 69.8
71.5 69.7 70.5 71.3 69.1 70.9 70.6

5. Weight of filled bags [g] in an automatic filling

203 199 198 201 200 201 201

6. Gasoline consumption [miles per gallon, rounded] of six cars of the same model under similar conditions

15.0 15.5 14.5 15.0 15.5 15.0

7. Release time [sec] of a relay

1.3 1.2 1.4 1.5 1.3 1.3 1.4 1.1 1.5 1.4
1.6 1.3 1.5 1.1 1.4 1.2 1.3 1.5 1.4 1.4

8. Foundrax test of Brinell hardness (2.5 mm steel ball, 62.5 kg load, 30 sec) of 20 copper plates (values in kg/mm^2)

86 86 87 89 76 85 82 86 87 85
90 88 89 90 88 80 84 89 90 89

9. Efficiency [%] of seven Voith Francis turbines of runner diameter 2.3 m under a head range of 185 m

91.8 89.1 89.9 92.5 90.7 91.2 91.0

10. -0.51 0.12 -0.47 0.95 0.25 -0.18 -0.54

11–16 AVERAGE AND SPREAD

Find the mean and compare it with the median. Find the standard deviation and compare it with the interquartile range.

11. For the data in Prob. 1

12. For the phone call data in Prob. 2

13. For the medical data in Prob. 3

14. For the iron contents in Prob. 4

15. For the release times in Prob. 7

16. For the Brinell hardness data in Prob. 8

17. Outlier, reduced data. Calculate s for the data 4 1 3 10 2. Then reduce the data by deleting the outlier and calculate s. Comment.

18. Outlier, reduction. Do the same tasks as in Prob. 17 for the hardness data in Prob. 8.

19. Construct the simplest possible data with $\bar{x} = 100$ but $q_M = 0$. What is the point of this problem?

20. Mean. Prove that $\bar{x}$ must always lie between the smallest and the largest data values.

24.2 Experiments, Outcomes, Events

We now turn to **probability theory**. This theory has the purpose of providing mathematical models of situations affected or even governed by “chance effects,” for instance, in weather forecasting, life insurance, quality of technical products (computers, batteries, steel sheets, etc.), traffic problems, and, of course, games of chance with cards or dice. And the accuracy of these models can be tested by suitable observations or experiments—this is a main purpose of **statistics** to be explained in Chap. 25.

We begin by defining some standard terms. An **experiment** is a process of measurement or observation, in a laboratory, in a factory, on the street, in nature, or wherever; so “experiment” is used in a rather general sense. Our interest is in experiments that involve **randomness**, chance effects, so that we cannot predict a result exactly. A **trial** is a single performance of an experiment. Its result is called an **outcome** or a **sample point**. n trials then give a **sample** of **size** n consisting of n sample points. The **sample space** S of an experiment is the set of all possible outcomes.

EXAMPLES 1–6 Random Experiments. Sample Spaces

(1) Inspecting a lightbulb. $S = \{\text{Defective, Nondefective}\}$.

(2) Rolling a die. $S = \{1, 2, 3, 4, 5, 6\}$.

(3) Measuring tensile strength of wire. S the numbers in some interval.

(4) Measuring copper content of brass. S: 50% to 90%, say.

(5) Counting daily traffic accidents in New York. S the integers in some interval.

(6) Asking for opinion about a new car model. $S = \{\text{Like, Dislike, Undecided}\}$. ■

The subsets of S are called **events** and the outcomes **simple events**.

EXAMPLE 7 Events

In (2), events are $A = \{1, 3, 5\}$ (*"Odd number"*), $B = \{2, 4, 6\}$ (*"Even number"*), $C = \{5, 6\}$. etc. Simple events are $\{1\}, \{2\}, \cdots, \{6\}$. ■

If, in a trial, an outcome a happens and $a \in A$ (a *is an element of* A), we say that A happens. For instance, if a die turns up a 3, the event A*: Odd number* happens. Similarly, if C in Example 7 happens (meaning 5 or 6 turns up), then, say, $D = \{4, 5, 6\}$ happens. Also note that S happens in each trial, meaning that *some* event of S always happens. All this is quite natural.

Unions, Intersections, Complements of Events

In connection with basic probability laws we shall need the following concepts and facts about events (subsets) $A, B, C, \cdots$ of a given sample space S.

The **union** $A \cup B$ of A and B consists of all points in A or B or both.

The **intersection** $A \cap B$ of A and B consists of all points that are in both A and B.

If A and B have no points in common, we write

$$A \cap B = \varnothing$$

where $\varnothing$ is the *empty set* (set with no elements) and we call A and B **mutually exclusive** (or **disjoint**) because, in a trial, the occurrence of A *excludes* that of B (and conversely)—if your die turns up an odd number, it cannot turn up an even number in the same trial. Similarly, a coin cannot turn up *Head* and *Tail* at the same time.

Complement A^c of A. This is the set of all the points of S ***not*** in A. Thus,

$$A \cap A^c = \varnothing, \qquad A \cup A^c = S.$$

In Example 7 we have $A^c = B$, hence $A \cup A^c = \{1, 2, 3, 4, 5, 6\} = S$.

Another notation for the complement of A is $\bar{A}$ (instead of A^c), but we shall not use this because in set theory $\bar{A}$ is used to denote the *closure* of A (not needed in our work).

Unions and intersections of more events are defined similarly. The **union**

$$\bigcup_{j=1}^{m} A_j = A_1 \cup A_2 \cup \cdots \cup A_m$$

of events $A_1, \cdots, A_m$ consists of all points that are in at least one A_j. Similarly for the union $A_1 \cup A_2 \cup \cdots$ of infinitely many subsets $A_1, A_2, \cdots$ of an *infinite* sample space S (that is, S consists of infinitely many points). The **intersection**

$$\bigcap_{j=1}^{m} A_j = A_1 \cap A_2 \cap \cdots \cap A_m$$

of $A_1, \cdots, A_m$ consists of the points of S that are in each of these events. Similarly for the intersection $A_1 \cap A_2 \cap \cdots$ of infinitely many subsets of S.

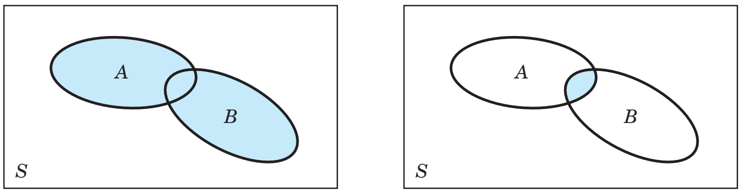
Working with events can be illustrated and facilitated by **Venn diagrams**[1] for showing unions, intersections, and complements, as in Figs. 510 and 511, which are typical examples that give the idea.

EXAMPLE 8 Unions and Intersections of 3 Events

In rolling a die, consider the events

A: *Number greater than* 3, B: *Number less than* 6, C: *Even number*.

Then $A \cap B = \{4, 5\}$, $B \cap C = \{2, 4\}$, $C \cap A = \{4, 6\}$, $A \cap B \cap C = \{4\}$. Can you sketch a Venn diagram of this? Furthermore, $A \cup B = S$, hence $A \cup B \cup C = S$ (why?). ■



Union $A \cup B$ Intersection $A \cap B$

Fig. 510. Venn diagrams showing two events A and B in a sample space S and their union $A \cup B$ (colored) and intersection $A \cap B$ (colored)

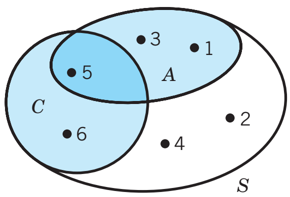


Fig. 511. Venn diagram for the experiment of rolling a die, showing S, $A = \{1, 3, 5\}$, $C = \{5, 6\}$, $A \cup C = \{1, 3, 5, 6\}$, $A \cap C = \{5\}$

PROBLEM SET 24.2

1–12 SAMPLE SPACES, EVENTS

Graph a sample space for the experiments:

1. Drawing 3 screws from a lot of right-handed and left-handed screws
2. Tossing 2 coins
3. Rolling 2 dice
4. Rolling a die until the first *Six* appears
5. Tossing a coin until the first *Head* appears
6. Recording the lifetime of each of 3 lightbulbs

[1] JOHN VENN (1834–1923), English mathematician.

7. Recording the daily maximum temperature X and the daily maximum air pressure Y at Times Square in New York

8. Choosing a committee of 2 from a group of 5 people

9. Drawing gaskets from a lot of 10, containing one defective D, unitil D is drawn, one at a time and assuming **sampling without replacement**, that is, gaskets drawn are *not* returned to the lot. (More about this in Sec. 24.6)

10. In rolling 3 dice, are the events *A: Sum divisible by* 3 and *B: Sum divisible by* 5 mutually exclusive?

11. Answer the questions in Prob. 10 for rolling 2 dice.

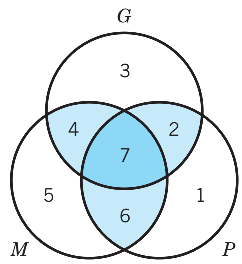
12. List all 8 subsets of the sample space $S = \{a, b, c\}$.

13. In Prob. 3 circle and mark the events *A: Faces are equal, B: Sum of faces less than* 5, $A \cup B$, $A \cap B$, A^c, B^c.

14. In drawing 2 screws from a lot of right-handed and left-handed screws, let A, B, C, D mean at a least 1 right-handed, at least 1 left-handed, 2 right-handed, 2 left-handed, respectively. Are A and B mutually exclusive? C and D?

15–20 VENN DIAGRAMS

15. In connection with a trip to Europe by some students, consider the events P that they see Paris, G that they have a good time, and M that they run out of money, and describe in words the events $1, \cdots, 7$ in the diagram.



Problem 15

16. Show that, by the definition of complement, for any subset A of a sample space S.

$$(A^c)^c = A, \qquad S^c = \varnothing, \qquad \varnothing^c = S,$$
$$A \cup A^c = S, \qquad A \cap A^c = \varnothing.$$

17. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cup B = B$.

18. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.

19. (**De Morgan's laws**) Using Venn diagrams, graph and check *De Morgan's laws*

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c.$$

20. Using Venn diagrams, graph and check the rules

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

24.3 Probability

The "probability" of an event A in an experiment is supposed to measure how frequently A is *about* to occur if we make many trials. If we flip a coin, then heads H and tails T will appear *about* equally often—we say that H and T are "**equally likely**." Similarly, for a regularly shaped die of homogeneous material ("**fair die**") each of the six outcomes $1, \cdots, 6$ will be equally likely. These are examples of experiments in which the sample space S consists of finitely many outcomes (points) that for reasons of some symmetry can be regarded as equally likely. This suggests the following definition.

DEFINITION 1

First Definition of Probability

If the sample space S of an experiment consists of finitely many outcomes (points) that are equally likely, then the probability $P(A)$ of an event A is

$$P(A) = \frac{\text{Number of points in } A}{\text{Number of points in } S}. \tag{1}$$

From this definition it follows immediately that, in particular,

(2) $$P(S) = 1.$$

EXAMPLE 1 **Fair Die**

In rolling a fair die once, what is the probability $P(A)$ of A of obtaining a 5 or a 6? The probability of B: "*Even number*"?

Solution. The six outcomes are equally likely, so that each has probability $1/6$. Thus $P(A) = 2/6 = 1/3$ because $A = \{5, 6\}$ has 2 points, and $P(B) = 3/6 = 1/2$. ■

Definition 1 takes care of many games as well as some practical applications, as we shall see, but certainly not of all experiments, simply because in many problems we do not have finitely many equally likely outcomes. To arrive at a more general definition of probability, we regard ***probability as the counterpart of relative frequency***. Recall from Sec. 24.1 that the **absolute frequency** $f(A)$ of an event A in n trials is the number of times A occurs, and the **relative frequency** of A in these trials is $f(A)/n$; thus

(3) $$f_{\text{rel}}(A) = \frac{f(A)}{n} = \frac{\text{Number of times } A \text{ occurs}}{\text{Number of trials}}.$$

Now if A did not occur, then $f(A) = 0$. If A always occurred, then $f(A) = n$. These are the extreme cases. Division by n gives

(4*) $$0 \leqq f_{\text{rel}}(A) \leqq 1.$$

In particular, for $A = S$ we have $f(S) = n$ because S always occurs (meaning that some event always occurs; if necessary, see Sec. 24.2, after Example 7). Division by n gives

(5*) $$f_{\text{rel}}(S) = 1.$$

Finally, if A and B are mutually exclusive, they cannot occur together. Hence the absolute frequency of their union $A \cup B$ must equal the sum of the absolute frequencies of A and B. Division by n gives the same relation for the relative frequencies,

(6*) $$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B) \qquad (A \cap B = \varnothing).$$

We are now ready to extend the definition of probability to experiments in which equally likely outcomes are not available. Of course, the extended definition should include Definition 1. Since probabilities are supposed to be the theoretical counterpart of relative frequencies, we choose the properties in (4*), (5*), (6*) as axioms. (Historically, such a choice is the result of a long process of gaining experience on what might be best and most practical.)

DEFINITION 2

General Definition of Probability

Given a sample space S, with each event A of S (subset of S) there is associated a number $P(A)$, called the **probability** of A, such that the following **axioms of probability** are satisfied.

1. For every A in S,

$$0 \leqq P(A) \leqq 1. \tag{4}$$

2. The entire sample space S has the probability

$$P(S) = 1. \tag{5}$$

3. For mutually exclusive events A and B ($A \cap B = \varnothing$; see Sec. 24.2),

$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \varnothing). \tag{6}$$

If S is infinite (has infinitely many points), Axiom 3 has to be replaced by

3′. For mutually exclusive events $A_1, A_2, \cdots$,

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots. \tag{6'}$$

In the infinite case the subsets of S on which $P(A)$ is defined are restricted to form a so-called σ-*algebra*, as explained in Ref. [GenRef6] (not [G6]!) in App. 1. This is of no practical consequence to us.

Basic Theorems of Probability

We shall see that the axioms of probability will enable us to build up probability theory and its application to statistics. We begin with three basic theorems. The first of them is useful if we can get the probability of the complement A^c more easily than $P(A)$ itself.

THEOREM 1

Complementation Rule

For an event A and its complement A^c in a sample space S,

$$P(A^c) = 1 - P(A). \tag{7}$$

PROOF By the definition of complement (Sec. 24.2), we have $S = A \cup A^c$ and $A \cap A^c = \varnothing$. Hence by Axioms 2 and 3,

$$1 = P(S) = P(A) + P(A^c), \qquad \text{thus} \qquad P(A^c) = 1 - P(A). \qquad \blacksquare$$

EXAMPLE 2 **Coin Tossing**

Five coins are tossed simultaneously. Find the probability of the event A: *At least one head turns up.* Assume that the coins are fair.

Solution. Since each coin can turn up heads or tails, the sample space consists of $2^5 = 32$ outcomes. Since the coins are fair, we may assign the same probability (1/32) to each outcome. Then the event A^c (*No heads turn up*) consists of only 1 outcome. Hence $P(A^c) = 1/32$, and the answer is $P(A) = 1 - P(A^c) = 31/32$. ■

The next theorem is a simple extension of Axiom 3, which you can readily prove by induction.

THEOREM 2 **Addition Rule for Mutually Exclusive Events**

For mutually exclusive events $A_1, \cdots, A_m$ in a sample space S,

$$P(A_1 \cup A_2 \cup \cdots A_m) = P(A_1) + P(A_2) + \cdots + P(A_m). \tag{8}$$

EXAMPLE 3 **Mutually Exclusive Events**

If the probability that on any workday a garage will get 10–20, 21–30, 31–40, over 40 cars to service is 0.20, 0.35, 0.25, 0.12, respectively, what is the probability that on a given workday the garage gets at least 21 cars to service?

Solution. Since these are mutually exclusive events, Theorem 2 gives the answer $0.35 + 0.25 + 0.12 = 0.72$. Check this by the complementation rule. ■

In many cases, events will not be mutually exclusive. Then we have

THEOREM 3 **Addition Rule for Arbitrary Events**

For events A and B in a sample space,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{9}$$

PROOF C, D, E in Fig. 512 make up $A \cup B$ and are mutually exclusive (disjoint). Hence by Theorem 2,

$$P(A \cup B) = P(C) + P(D) + P(E).$$

This gives (9) because on the right $P(C) + P(D) = P(A)$ by Axiom 3 and disjointness; and $P(E) = P(B) - P(D) = P(B) - P(A \cap B)$, also by Axiom 3 and disjointness. ■

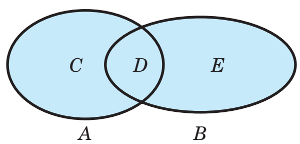


Fig. 512. Proof of Theorem 3

Note that for mutually exclusive events A and B we have $A \cap B = \varnothing$ by definition and, by comparing (9) and (6),

$$P(\varnothing) = 0. \tag{10}$$

(Can you also prove this by (5) and (7)?)

EXAMPLE 4 Union of Arbitrary Events

In tossing a fair die, what is the probability of getting an odd number or a number less than 4?

Solution. Let A be the event "*Odd number*" and B the event "*Number less than* 4." Then Theorem 3 gives the answer

$$P(A \cup B) = \tfrac{3}{6} + \tfrac{3}{6} - \tfrac{2}{6} = \tfrac{2}{3}$$

because $A \cap B =$ "*Odd number less than* 4" $= \{1, 3\}$. ■

Conditional Probability. Independent Events

Often it is required to find the probability of an event B under the condition that an event A occurs. This probability is called the **conditional probability** *of B given A* and is denoted by $P(B|A)$. In this case A serves as a new (reduced) sample space, and that probability is the fraction of $P(A)$ which corresponds to $A \cap B$. Thus

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) \neq 0]. \tag{11}$$

Similarly, the *conditional probability of A given B* is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad [P(B) \neq 0]. \tag{12}$$

Solving (11) and (12) for $P(A \cap B)$, we obtain

THEOREM 4

Multiplication Rule

If A and B are events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B). \tag{13}$$

EXAMPLE 5 Multiplication Rule

In producing screws, let A mean "screw too slim" and B "screw too short." Let $P(A) = 0.1$ and let the conditional probability that a slim screw is also too short be $P(B|A) = 0.2$. What is the probability that a screw that we pick randomly from the lot produced will be both too slim and too short?

Solution. $P(A \cap B) = P(A)P(B|A) = 0.1 \cdot 0.2 = 0.02 = 2\%$, by Theorem 4. ■

Independent Events. If events A and B are such that

$$P(A \cap B) = P(A)P(B), \tag{14}$$

they are called **independent events**. Assuming $P(A) \neq 0$, $P(B) \neq 0$, we see from (11)–(13) that in this case

$$P(A|B) = P(A), \qquad P(B|A) = P(B).$$

This means that the probability of A does not depend on the occurrence or nonoccurrence of B, and conversely. This justifies the term "independent."

Independence of m Events. Similarly, m events $A_1, \cdots, A_m$ are called **independent** if

$$P(A_1 \cap \cdots \cap A_m) = P(A_1) \cdots P(A_m) \tag{15a}$$

as well as for every k different events $A_{j_1}, A_{j_2}, \cdots, A_{j_k}$.

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}) \tag{15b}$$

where $k = 2, 3, \cdots, m - 1$.

Accordingly, three events A, B, C are independent if and only if

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(B \cap C) &= P(B)P(C), \\ P(C \cap A) &= P(C)P(A), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned} \tag{16}$$

Sampling. Our next example has to do with randomly drawing objects, *one at a time,* from a given set of objects. This is called **sampling from a population**, and there are two ways of sampling, as follows.

1. In **sampling with replacement**, the object that was drawn at random is placed back to the given set and the set is mixed thoroughly. Then we draw the next object at random.
2. In **sampling without replacement** the object that was drawn is put aside.

EXAMPLE 6 Sampling With and Without Replacement

A box contains 10 screws, three of which are defective. Two screws are drawn at random. Find the probability that neither of the two screws is defective.

Solution. We consider the events

A: First drawn screw nondefective.

B: Second drawn screw nondefective.

Clearly, $P(A) = \frac{7}{10}$ because 7 of the 10 screws are nondefective and we sample at random, so that each screw has the same probability $(\frac{1}{10})$ of being picked. If we sample with replacement, the situation before the second drawing is the same as at the beginning, and $P(B) = \frac{7}{10}$. The events are independent, and the answer is

$$P(A \cap B) = P(A)P(B) = 0.7 \cdot 0.7 = 0.49 = 49\%.$$

If we sample without replacement, then $P(A) = \frac{7}{10}$, as before. If A has occurred, then there are 9 screws left in the box, 3 of which are defective. Thus $P(B|A) = \frac{6}{9} = \frac{2}{3}$, and Theorem 4 yields the answer

$$P(A \cap B) = \frac{7}{10} \cdot \frac{2}{3} = 47\%.$$

Is it intuitively clear that this value must be smaller than the preceding one? ■

PROBLEM SET 24.3

1. In rolling 3 fair dice, what is the probability of obtaining a sum not greater than 16?
2. In rolling 2 fair dice, what is the probability of a sum greater than 3 but not exceeding 6?
3. Three screws are drawn at random from a lot of 100 screws, 10 of which are defective. Find the probability of the event that all 3 screws drawn are nondefective, assuming that we draw **(a)** with replacement, **(b)** without replacement.
4. In Prob. 3 find the probability of *E: At least* 1 *defective* (i) directly, (ii) by using complements; in both cases **(a)** and **(b)**.
5. If a box contains 10 left-handed and 20 right-handed screws, what is the probability of obtaining at least one right-handed screw in drawing 2 screws with replacement?
6. Will the probability in Prob. 5 increase or decrease if we draw without replacement. First guess, then calculate.
7. Under what conditions will it make *practically* no difference whether we sample with or without replacement?
8. If a certain kind of tire has a life exceeding 40,000 miles with probability 0.90, what is the probability that a set of these tires on a car will last longer than 40,000 miles?
9. If we inspect photocopy paper by randomly drawing 5 sheets without replacement from every pack of 500, what is the probability of getting 5 clean sheets although 0.4% of the sheets contain spots?
10. Suppose that we draw cards repeatedly and with replacement from a file of 100 cards, 50 of which refer to male and 50 to female persons. What is the probability of obtaining the second "female" card before the third "male" card?
11. A batch of 200 iron rods consists of 50 oversized rods, 50 undersized rods, and 100 rods of the desired length. If two rods are drawn at random without replacement, what is the probability of obtaining **(a)** two rods of the desired length, **(b)** exactly one of the desired length, **(c)** none of the desired length?
12. If a circuit contains four automatic switches and we want that, with a probability of 99%, during a given time interval the switches to be all working, what probability of failure per time interval can we admit for a single switch?
13. A pressure control apparatus contains 3 electronic tubes. The apparatus will not work unless all tubes are operative. If the probability of failure of each tube during some interval of time is 0.04, what is the corresponding probability of failure of the apparatus?
14. Suppose that in a production of spark plugs the fraction of defective plugs has been constant at 2% over a long time and that this process is controlled every half hour by drawing and inspecting two just produced. Find the probabilities of getting **(a)** no defectives, **(b)** 1 defective, **(c)** 2 defectives. What is the sum of these probabilities?
15. What gives the greater probability of hitting at least once: **(a)** hitting with probability $1/2$ and firing 1 shot, **(b)** hitting with probability $1/4$ and firing 2 shots, **(c)** hitting with probability $1/8$ and firing 4 shots? First guess.
16. You may wonder whether in (16) the last relation follows from the others, but the answer is no. To see this, imagine that a chip is drawn from a box containing 4 chips numbered 000, 011, 101, 110, and let A, B, C be the events that the first, second, and third digit, respectively, on the drawn chip is 1. Show that then the first three formulas in (16) hold but the last one does not hold.
17. Show that if B is a subset of A, then $P(B) \leqq P(A)$.
18. Extending Theorem 4, show that $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$.
19. Make up an example similar to Prob. 16, for instance, in terms of divisibility of numbers.

24.4 Permutations and Combinations

Permutations and combinations help in finding probabilities $P(A) = a/k$ by ***systematically counting*** the number a of points of which an event A consists; here, k is the number of points of the sample space S. The practical difficulty is that a may often be surprisingly large, so that actual counting becomes hopeless. For example, if in assembling some instrument you need 10 different screws in a certain order and you want to draw them randomly from a box (which contains nothing else) the probability of obtaining them in the required order is only 1/3,628,800 because there are

$$10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3{,}628{,}800$$

orders in which they can be drawn. Similarly, in many other situations the numbers of orders, arrangements, etc. are often incredibly large. (If you are unimpressed, take 20 screws—how much bigger will the number be?)

Permutations

A **permutation** of given things (*elements* or *objects*) is an arrangement of these things in a row in some order. For example, for three letters a, b, c there are $3! = 1 \cdot 2 \cdot 3 = 6$ permutations: *abc*, *acb*, *bac*, *bca*, *cab*, *cba*. This illustrates (a) in the following theorem.

THEOREM 1

Permutations

(a) ***Different things.*** *The number of permutations of n different things taken all at a time is*

$$n! = 1 \cdot 2 \cdot 3 \cdots n \qquad \text{(read "}n\text{ factorial")}. \tag{1}$$

(b) ***Classes of equal things.*** *If n given things can be divided into c classes of alike things differing from class to class, then the number of permutations of these things taken all at a time is*

$$\frac{n!}{n_1!n_2!\cdots n_c!} \qquad (n_1 + n_2 + \cdots + n_c = n) \tag{2}$$

Where n_j is the number of things in the jth class.

PROOF **(a)** There are n choices for filling the first place in the row. Then $n - 1$ things are still available for filling the second place, etc.

(b) n_1 alike things in class 1 make $n_1!$ permutations collapse into a single permutation (those in which class 1 things occupy the same n_1 positions), etc., so that (2) follows from (1). ∎

EXAMPLE 1 **Illustration of Theorem 1(b)**

If a box contains 6 red and 4 blue balls, the probability of drawing first the red and then the blue balls is

$$P = 6!4!/10! = 1/210 \approx 0.5\%.$$ ∎

A **permutation of *n* things taken *k* at a time** is a permutation containing only k of the n given things. Two such permutations consisting of the same k elements, in a different order, are different, by definition. For example, there are 6 different permutations of the three letters a, b, c, taken two letters at a time, *ab*, *ac*, *bc*, *ba*, *ca*, *cb*.

A **permutation of *n* things taken *k* at a time with repetitions** is an arrangement obtained by putting any given thing in the first position, any given thing, including a repetition of the one just used, in the second, and continuing until k positions are filled. For example, there are $3^2 = 9$ different such permutations of a, b, c taken 2 letters at a time, namely, the preceding 6 permutations and aa, bb, cc. You may prove (see Team Project 14):

THEOREM 2

Permutations

The number of different permutations of n different things taken k at a time without repetitions is

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} \tag{3a}$$

and with repetitions is

$$n^k. \tag{3b}$$

EXAMPLE 2 Illustration of Theorem 2

In an encrypted message the letters are arranged in groups of five letters, called *words*. From (3b) we see that the number of different such words is

$$26^5 = 11{,}881{,}376.$$

From (3a) it follows that the number of different such words containing each letter no more than once is

$$26!/(26-5)! = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7{,}893{,}600.$$

Combinations

In a permutation, the order of the selected things is essential. In contrast, a **combination** of given things means any selection of one or more things *without regard to order*. There are two kinds of combinations, as follows.

The number of **combinations of n different things, taken k at a time, without repetitions** is the number of sets that can be made up from the n given things, each set containing k different things and no two sets containing exactly the same k things.

The number of **combinations of n different things, taken k at a time, with repetitions** is the number of sets that can be made up of k things chosen from the given n things, each being used as often as desired.

For example, there are three combinations of the three letters a, b, c, taken two letters at a time, without repetitions, namely, ab, ac, bc, and six such combinations with repetitions, namely, ab, ac, bc, aa, bb, cc.

THEOREM 3

Combinations

The number of different combinations of n different things taken, k at a time, without repetitions, is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdots k}, \tag{4a}$$

and the number of those combinations with repetitions is

$$\binom{n+k-1}{k}. \tag{4b}$$

PROOF The statement involving (4a) follows from the first part of Theorem 2 by noting that there are $k!$ *permutations* of k things from the given n things that differ by the order of the elements (see Theorem 1), but there is only a single *combination* of those k things of the type characterized in the first statement of Theorem 3. The last statement of Theorem 3 can be proved by induction (see Team Project 14). ∎

EXAMPLE 3 **Illustration of Theorem 3**

The number of samples of five lightbulbs that can be selected from a lot of 500 bulbs is [see (4a)]

$$\binom{500}{5} = \frac{500!}{5!495!} = \frac{500 \cdot 499 \cdot 498 \cdot 497 \cdot 496}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 255{,}244{,}687{,}600. \qquad ∎$$

Factorial Function

In (1)–(4) the **factorial function** is basic. By definition,

$$0! = 1. \tag{5}$$

Values may be computed recursively from given values by

$$(n+1)! = (n+1)n!. \tag{6}$$

For large n the function is very large (see Table A3 in App. 5). A convenient approximation for large n is the **Stirling formula**[2]

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (e = 2.718 \cdots) \tag{7}$$

where $\sim$ is read "**asymptotically equal**" and means that the ratio of the two sides of (7) approaches 1 as n approaches infinity.

EXAMPLE 4 **Stirling Formula**

$n!$	By (7)	Exact Value	Relative Error
4!	23.5	24	2.1%
10!	3,598,696	3,628,800	0.8%
20!	$2.42279 \cdot 10^{18}$	2,432,902,008,176,640,000	0.4%

∎

Binomial Coefficients

The **binomial coefficients** are defined by the formula

$$\binom{a}{k} = \frac{a(a-1)(a-2)\cdots(a-k+1)}{k!} \qquad (k \geqq 0, \text{ integer}). \tag{8}$$

[2]JAMES STIRLING (1692–1770), Scots mathematician.

The numerator has k factors. Furthermore, we define

$$\binom{a}{0} = 1, \qquad \text{in particular,} \qquad \binom{0}{0} = 1. \tag{9}$$

For integer $a = n$ we obtain from (8)

$$\binom{n}{k} = \binom{n}{n-k} \qquad (n \geqq 0,\ 0 \leqq k \leqq n). \tag{10}$$

Binomial coefficients may be computed recursively, because

$$\binom{a}{k} + \binom{a}{k+1} = \binom{a+1}{k+1} \qquad (k \geqq 0, \text{ integer}). \tag{11}$$

Formula (8) also yields

$$\binom{-m}{k} = (-1)^k \binom{m+k-1}{k} \qquad \begin{array}{r}(k \geqq 0, \text{ integer})\\ (m > 0).\end{array} \tag{12}$$

There are numerous further relations; we mention two important ones,

$$\sum_{s=0}^{n-1} \binom{k+s}{k} = \binom{n+k}{k+1} \qquad \begin{array}{r}(k \geqq 0,\ n \geqq 1,\\ \text{both integer)}\end{array} \tag{13}$$

and

$$\sum_{k=0}^{r} \binom{p}{k}\binom{q}{r-k} = \binom{p+q}{r} \qquad (r \geqq 0, \text{ integer}). \tag{14}$$

PROBLEM SET 24.4

Note the large numbers in the answers to some of these problems, which would make ***counting cases hopeless!***

1. In how many ways can a company assign 10 drivers to n buses, one driver to each bus and conversely?

2. List **(a)** all permutations, **(b)** all combinations without repetitions, **(c)** all combinations with repetitions, of 5 letters a, e, i, o, u taken 2 at a time.

3. If a box contains 4 rubber gaskets and 2 plastic gaskets, what is the probability of drawing **(a)** first the plastic and then the rubber gaskets, **(b)** first the rubber and then the plastic ones? Do this by using a theorem and checking it by multiplying probabilities.

4. An urn contains 2 green, 3 yellow, and 5 red balls. We draw 1 ball at random and put it aside. Then we draw the next ball, and so on. Find the probability of drawing at first the 2 green balls, then the 3 yellow ones, and finally the red ones.

5. In how many different ways can we select a committee consisting of 3 engineers, 2 physicists, and 2 computer scientists from 10 engineers, 5 physicists, and 6 computer scientists? First guess.

6. How many different samples of 4 objects can we draw from a lot of 50?

7. Of a lot of 10 items, 2 are defective. **(a)** Find the number of different samples of 4. Find the number of samples of 4 containing **(b)** no defectives, **(c)** 1 defective, **(d)** 2 defectives.

8. Determine the number of different bridge hands. (A bridge hand consists of 13 cards selected from a full deck of 52 cards.)

9. In how many different ways can 6 people be seated at a round table?

10. If a cage contains 100 mice, 3 of which are male, what is the probability that the 3 male mice will be included if 10 mice are randomly selected?

11. How many automobile registrations may the police have to check in a hit-and-run accident if a witness reports KDP7 and cannot remember the last two digits on the license plate but is certain that all three digits were different?

12. If 3 suspects who committed a burglary and 6 innocent persons are lined up, what is the probability that a witness who is not sure and has to pick three persons will pick the three suspects by chance? That the witness picks 3 innocent persons by chance?

13. CAS PROJECT. Stirling formula. (a) Using (7), compute approximate values of $n!$ for $n = 1, \cdots, 20$.

(b) Determine the relative error in (a). Find an empirical formula for that relative error.

(c) An upper bound for that relative error is $e^{1/12n} - 1$. Try to relate your empirical formula to this.

(d) Search through the literature for further information on Stirling's formula. Write a short eassy about your findings, arranged in logical order and illustrated with numeric examples.

14. TEAM PROJECT. Permutations, Combinations.

(a) Prove Theorem 2.

(b) Prove the last statement of Theorem 3.

(c) Derive (11) from (8).

(d) By the **binomial theorem**,

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

so that $a^k b^{n-k}$ has the coefficient $\binom{n}{k}$. Can you conclude this from Theorem 3 or is this a mere coincidence?

(e) Prove (14) by using the binomial theorem.

(f) Collect further formulas for binomial coefficients from the literature and illustrate them numerically.

15. Birthday problem. What is the probability that in a group of 20 people (that includes no twins) at least two have the same birthday, if we assume that the probability of having birthday on a given day is $1/365$ for every day. First guess. *Hint.* Consider the complementary event.

24.5 Random Variables. Probability Distributions

In Sec. 24.1 we considered frequency distributions of data. These distributions show the absolute or relative frequency of the data values. Similarly, a **probability distribution** or, briefly, a **distribution**, shows the probabilities of events in an experiment. The quantity that we observe in an experiment will be denoted by X and called a **random variable** (or **stochastic variable**) because the value it will assume in the next trial depends on chance, on **randomness**—if you roll a die, you get one of the numbers from 1 to 6, but you don't know which one will show up next. Thus X = *Number a die turns up* is a random variable. So is X = *Elasticity of rubber* (elongation at break). ("Stochastic" means related to chance.)

If we ***count*** (cars on a road, defective screws in a production, tosses until a die shows the first Six), we have a **discrete random variable and distribution**. If we ***measure*** (electric voltage, rainfall, hardness of steel), we have a **continuous random variable and distribution**. Precise definitions follow. In both cases the distribution of X is determined by the **distribution function**

$$F(x) = P(X \leqq x); \tag{1}$$

this is the probability that in a trial, X will assume any value not exceeding x.

CAUTION! The terminology is not uniform. $F(x)$ is sometimes also called the **cumulative distribution function**.

For (1) to make sense in both the discrete and the continuous case we formulate conditions as follows.

DEFINITION

Random Variable

A **random variable** X is a function defined on the sample space S of an experiment. Its values are real numbers. For every number a the probability

$$P(X = a)$$

with which X assumes a is defined. Similarly, for any interval I the probability

$$P(X \in I)$$

with which X assumes any value in I is defined.

Although this definition is very general, in practice only a very small number of distributions will occur over and over again in applications.

From (1) we obtain the fundamental formula for the probability corresponding to an interval $a < x \leqq b$,

$$P(a < X \leqq b) = F(b) - F(a). \tag{2}$$

This follows because $X \leqq a$ (*"X assumes any value not exceeding a"*) and $a < X \leqq b$ (*"X assumes any value in the interval $a < x \leqq b$"*) are mutually exclusive events, so that by (1) and Axiom 3 of Definition 2 in Sec. 24.3

$$\begin{aligned} F(b) = P(X \leqq b) &= P(X \leqq a) + P(a < X \leqq b) \\ &= F(a) + P(a < X \leqq b) \end{aligned}$$

and subtraction of $F(a)$ on both sides gives (2).

Discrete Random Variables and Distributions

By definition, a random variable X and its distribution are **discrete** if X assumes only finitely many or at most countably many values $x_1, x_2, x_3, \cdots$, called the **possible values** of X, with positive probabilities $p_1 = P(X = x_1), p_2 = P(X = x_2), \quad p_3 = P(X = x_3), \cdots$, whereas the probability $P(X \in I)$ is zero for any interval I containing no possible value.

Clearly, the discrete distribution of X is also determined by the **probability function** $f(x)$ of X, defined by

$$f(x) = \begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases} \qquad (j = 1, 2, \cdots), \tag{3}$$

From this we get the values of the **distribution function** $F(x)$ by taking sums,

$$F(x) = \sum_{x_j \leqq x} f(x_j) = \sum_{x_j \leqq x} p_j \tag{4}$$

where for any given x we sum all the probabilities p_j for which x_j is smaller than or equal to that of x. This is a **step function** with upward jumps of size p_j at the possible values x_j of X and constant in between.

EXAMPLE 1 Probability Function and Distribution Function

Figure 513 shows the probability function $f(x)$ and the distribution function $F(x)$ of the discrete random variable

$$X = \textit{Number a fair die turns up}.$$

X has the possible values $x = 1, 2, 3, 4, 5, 6$ with probability $1/6$ each. At these x the distribution function has upward jumps of magnitude $1/6$. Hence from the graph of $f(x)$ we can construct the graph of $F(x)$ and conversely.

In Figure 513 (and the next one) at each jump the ***fat dot*** indicates the ***function value at the jump!*** ■

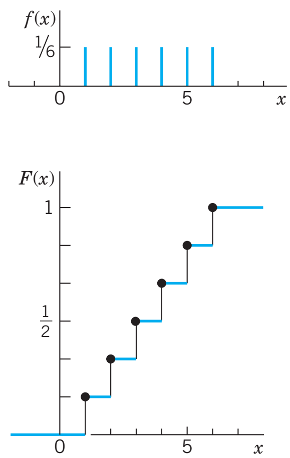


Fig. 513. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Number obtained in tossing a fair die once*

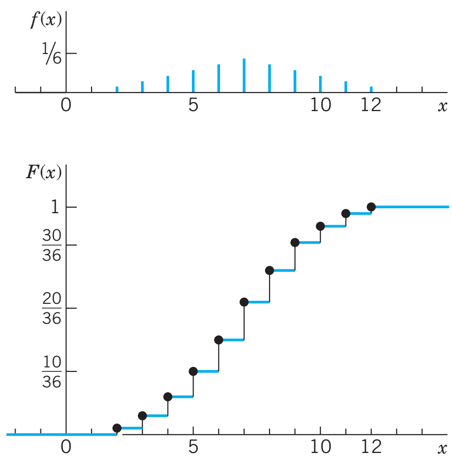


Fig. 514. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Sum of the two numbers obtained in tossing two fair dice once*

EXAMPLE 2 Probability Function and Distribution Function

The random variable X = *Sum of the two numbers two fair dice turn up* is discrete and has the possible values $2\,(= 1 + 1), 3, 4, \cdots, 12\,(= 6 + 6)$. There are $6 \cdot 6 = 36$ equally likely outcomes $(1, 1)\,(1, 2), \cdots, (6, 6)$, where the first number is that shown on the first die and the second number that on the other die. Each such outcome has probability $1/36$. Now $X = 2$ occurs in the case of the outcome $(1, 1)$; $X = 3$ in the case of the two outcomes $(1, 2)$ and $(2, 1)$; $X = 4$ in the case of the three outcomes $(1, 3), (2, 2), (3, 1)$; and so on. Hence $f(x) = P(X = x)$ and $F(x) = P(X \leqq x)$ have the values

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Figure 514 shows a bar chart of this function and the graph of the distribution function, which is again a step function, with jumps (of different height!) at the possible values of X. ■

Two useful formulas for discrete distributions are readily obtained as follows. For the probability corresponding to intervals we have from (2) and (4)

$$P(a < X \leqq b) = F(b) - F(a) = \sum_{a < x_j \leqq b} p_j \qquad (X \text{ discrete}). \tag{5}$$

This is the sum of all probabilities p_j for which x_j satisfies $a < x_j \leqq b$. (Be careful about $<$ and $\leqq$!) From this and $P(S) = 1$ (Sec. 24.3) we obtain the following formula.

$$\sum_j p_j = 1 \qquad \text{(sum of all probabilities).} \tag{6}$$

EXAMPLE 3 Illustration of Formula (5)

In Example 2, compute the probability of a sum of at least 4 and at most 8.

Solution. $P(3 < X \leqq 8) = F(8) - F(3) = \frac{26}{36} - \frac{3}{36} = \frac{23}{36}$. ■

EXAMPLE 4 Waiting Time Problem. Countably Infinite Sample Space

In tossing a fair coin, let X = *Number of trials until the first head appears.* Then, by independence of events (Sec. 24.3),

$$\begin{aligned}
P(X = 1) &= P(H) &&= \tfrac{1}{2} && (H = \text{Head})\\
P(X = 2) &= P(TH) &&= \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4} && (T = \text{Tail})\\
P(X = 3) &= P(TTH) &&= \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{8}, && \text{etc.}
\end{aligned}$$

and in general $P(X = n) = (\frac{1}{2})^n$, $n = 1, 2, \cdots$. Also, (6) can be confirmed by the sum formula for the geometric series,

$$\begin{aligned}
\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots &= -1 + \frac{1}{1 - \frac{1}{2}}\\
&= -1 + 2 = 1.
\end{aligned}$$ ■

Continuous Random Variables and Distributions

Discrete random variables appear in experiments in which we ***count*** (defectives in a production, days of sunshine in Chicago, customers standing in a line, etc.). Continuous random variables appear in experiments in which we ***measure*** (lengths of screws, voltage in a power line, Brinell hardness of steel, etc.). By definition, a random variable X and its distribution are *of continuous type* or, briefly, **continuous**, if its distribution function $F(x)$ [defined in (1)] can be given by an integral

$$F(x) = \int_{-\infty}^{x} f(v)\, dv \tag{7}$$

(we write v because x is needed as the upper limit of the integral) whose integrand $f(x)$, called the **density** of the distribution, is nonnegative, and is continuous, perhaps except for finitely many x-values. Differentiation gives the relation of f to F as

$$f(x) = F'(x) \tag{8}$$

for every x at which $f(x)$ is continuous.

From (2) and (7) we obtain the very important formula for the probability corresponding to an interval:

$$P(a < X \leqq b) = F(b) - F(a) = \int_a^b f(v)\,dv. \tag{9}$$

This is the analog of (5).

From (7) and $P(S) = 1$ (Sec. 24.3) we also have the analog of (6):

$$\int_{-\infty}^{\infty} f(v)\,dv = 1. \tag{10}$$

Continuous random variables are ***simpler than discrete ones*** with respect to intervals. Indeed, in the continuous case the four probabilities corresponding to $a < X \leqq b$, $a < X < b$, $a \leqq X < b$, and $a \leqq X \leqq b$ with any fixed a and b $(> a)$ are all the same. Can you see why? (*Answer.* This probability is the area under the density curve, as in Fig. 515, and does not change by adding or subtracting a single point in the interval of integration.) This is different from the discrete case! (Explain.)

The next example illustrates notations and typical applications of our present formulas.

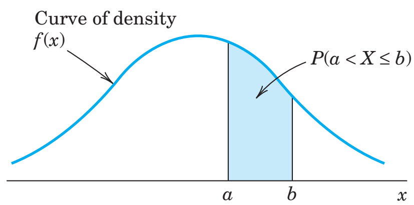


Fig. 515. Example illustrating formula (9)

EXAMPLE 5 Continuous Distribution

Let X have the density function $f(x) = 0.75(1 - x^2)$ if $-1 \leqq x \leqq 1$ and zero otherwise. Find the distribution function. Find the probabilities $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2})$ and $P(\frac{1}{4} \leqq X \leqq 2)$. Find x such that $P(X \leqq x) = 0.95$.

Solution. From (7) we obtain $F(x) = 0$ if $x \leqq -1$,

$$F(x) = 0.75 \int_{-1}^{x} (1 - v^2)\,dv = 0.5 + 0.75x - 0.25x^3 \qquad \text{if } -1 < x \leqq 1,$$

and $F(x) = 1$ if $x > 1$. From this and (9) we get

$$P(-\tfrac{1}{2} \leqq X \leqq \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-\tfrac{1}{2}) = 0.75 \int_{-1/2}^{1/2} (1 - v^2)\,dv = 68.75\%$$

(because $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2}) = P(-\frac{1}{2} < X \leqq \frac{1}{2})$ for a continuous distribution) and

$$P(\tfrac{1}{4} \leqq X \leqq 2) = F(2) - F(\tfrac{1}{4}) = 0.75 \int_{1/4}^{1} (1 - v^2)\, dv = 31.64\%.$$

(Note that the upper limit of integration is 1, not 2. Why?) Finally,

$$P(X \leqq x) = F(x) = 0.5 + 0.75x - 0.25x^3 = 0.95.$$

Algebraic simplification gives $3x - x^3 = 1.8$. A solution is $x = 0.73$, approximately.

Sketch $f(x)$ and mark $x = -\frac{1}{2}, \frac{1}{2}, \frac{1}{4}$, and 0.73, so that you can see the results (the probabilities) as areas under the curve. Sketch also $F(x)$. ■

Further examples of continuous distributions are included in the next problem set and in later sections.

PROBLEM SET 24.5

1. Graph the probability function $f(x) = kx^2$ ($x = 1, 2, 3, 4, 5$; k suitable) and the distribution function.

2. Graph the density function $f(x) = kx^2$ ($0 \leqq x \leqq 5$; k suitable) and the distribution function.

3. Uniform distribution. Graph f and F when the density of X is $f(x) = k = \text{const}$ if $-2 \leqq x \leqq 2$ and 0 elsewhere. Find $P(0 \leqq X \leqq 2)$.

4. In Prob. 3 find c and $\tilde{c}$ such that $P(-c < X < c) = 95\%$ and $P(0 < X < \tilde{c}) = 95\%$.

5. Graph f and F when $f(-2) = f(2) = \frac{1}{8}$, $f(-1) = f(1) = \frac{3}{8}$. Can f have further positive values?

6. A box contains 4 right-handed and 6 left-handed screws. Two screws are drawn at random without replacement. Let X be the number of left-handed screws drawn. Find the probabilities $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, $P(1 < X < 2)$, $P(X \leqq 1)$, $P(X \geqq 1)$, $P(X > 1)$, and $P(0.5 < X < 10)$.

7. Let X be the number of years before a certain kind of pump needs replacement. Let X have the probability function $f(x) = kx^3$, $x = 0, 1, 2, 3, 4$, Find k. Sketch f and F.

8. Graph the distribution function $F(x) = 1 - e^{-3x}$ if $x > 0$, $F(x) = 0$ if $x \leqq 0$, and the density $f(x)$. Find x such that $F(x) = 0.9$.

9. Let X [millimeters] be the thickness of washers. Assume that X has the density $f(x) = kx$ if $0.9 < x < 1.1$ and 0 otherwise. Find k. What is the probability that a washer will have thickness between 0.95 mm and 1.05 mm?

10. If the diameter X of axles has the density $f(x) = k$ if $119.9 \leqq x \leqq 120.1$ and 0 otherwise, how many defectives will a lot of 500 axles approximately contain if defectives are axles slimmer than 119.91 or thicker than 120.09?

11. Find the probability that none of three bulbs in a traffic signal will have to be replaced during the first 1500 hours of operation if the lifetime X of a bulb is a random variable with the density $f(x) = 6[0.25 - (x - 1.5)^2]$ when $1 \leqq x \leqq 2$ and $f(x) = 0$ otherwise, where x is measured in multiples of 1000 hours.

12 Let X be the ratio of sales to profits of some company. Assume that X has the distribution function $F(x) = 0$ if $x < 2$, $F(x) = (x^2 - 4)/5$ if $2 \leqq x < 3$, $F(x) = 1$ if $x \geqq 3$. Find and sketch the density. What is the probability that X is between 2.5 (40% profit) and 5 (20% profit)?

13. Suppose that in an automatic process of filling oil cans, the content of a can (in gallons) is $Y = 100 + X$, where X is a random variable with density $f(x) = 1 - |x|$ when $|x| \leqq 1$ and 0 when $|x| > 1$. Sketch $f(x)$ and $F(x)$. In a lot of 1000 cans, about how many will contain 100 gallons or more? What is the probability that a can will contain less than 99.5 gallons? Less than 99 gallons?

14. Find the probability function of X = *Number of times a fair die is rolled until the first Six appears* and show that it satisfies (6).

15. Let X be a random variable that can assume every real value. What are the complements of the events $X \leqq b$, $X < b$, $X \geqq c$, $X > c$, $b \leqq X \leqq c$, $b < X \leqq c$?

24.6 Mean and Variance of a Distribution

The mean μ and variance σ^2 of a random variable X and of its distribution are the theoretical counterparts of the mean $\bar{x}$ and variance s^2 of a frequency distribution in Sec. 24.1 and serve a similar purpose. Indeed, the mean characterizes the central location and the variance the spread (the variability) of the distribution. The **mean** μ (mu) is defined by

$$\textbf{(1)} \quad \begin{aligned} &\textbf{(a)} \quad \mu = \sum_j x_j f(x_j) && \text{(Discrete distribution)} \\ &\textbf{(b)} \quad \mu = \int_{-\infty}^{\infty} x f(x)\,dx && \text{(Continuous distribution)} \end{aligned}$$

and the **variance** σ^2 (sigma square) by

$$\textbf{(2)} \quad \begin{aligned} &\textbf{(a)} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) && \text{(Discrete distribution)} \\ &\textbf{(b)} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx && \text{(Continuous distribution).} \end{aligned}$$

σ (the positive square root of σ^2) is called the **standard deviation** of X and its distribution. f is the probability function or the density, respectively, in (a) and (b).

The mean μ is also denoted by $E(X)$ and is called the **expectation** *of* X because it gives the average value of X to be expected in many trials. Quantities such as μ and σ^2 that measure certain properties of a distribution are called **parameters**. μ and σ^2 are the two most important ones. From (2) we see that

$$(3) \qquad \sigma^2 > 0$$

(except for a discrete "distribution" with only one possible value, so that $\sigma^2 = 0$). We assume that μ and σ^2 exist (are finite), as is the case for practically all distributions that are useful in applications.

EXAMPLE 1 Mean and Variance

The random variable X = *Number of heads in a single toss of a fair coin* has the possible values $X = 0$ and $X = 1$ with probabilities $P(X = 0) = \frac{1}{2}$ and $P(X = 1) = \frac{1}{2}$. From (la) we thus obtain the mean $\mu = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$, and (2a) yields the variance

$$\sigma^2 = (0 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} + (1 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} = \tfrac{1}{4}. \qquad \blacksquare$$

EXAMPLE 2 Uniform Distribution. Variance Measures Spread

The distribution with the density

$$f(x) = \frac{1}{b - a} \qquad \text{if} \qquad a < x < b$$

and $f = 0$ otherwise is called the **uniform distribution** on the interval $a < x < b$. From (1b) (or from Theorem 1, below) we find that $\mu = (a + b)/2$, and (2b) yields the variance

$$\sigma^2 = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12}.$$

Figure 516 illustrates that the spread is large if and only if σ^2 is large. ■

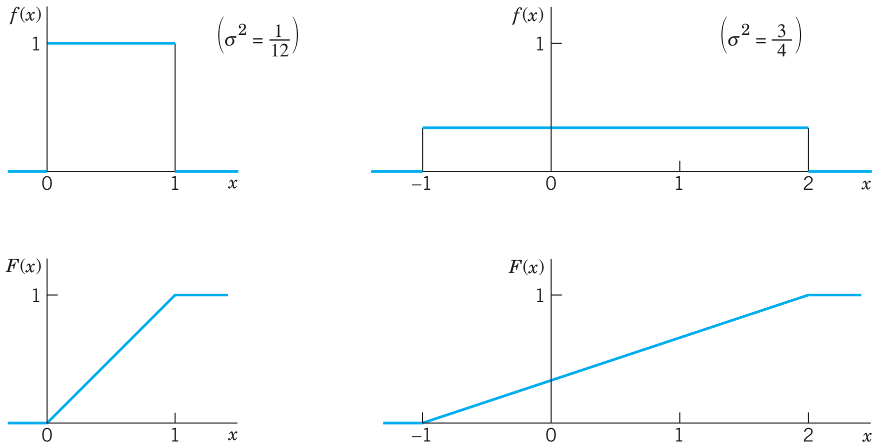


Fig. 516. Uniform distributions having the same mean (0.5) but different variances σ^2

Symmetry. We can obtain the mean μ without calculation if a distribution is symmetric. Indeed, you may prove

THEOREM 1

Mean of a Symmetric Distribution

If a distribution is **symmetric** *with respect to* $x = c$, *that is,* $f(c - x) = f(c + x)$, *then* $\mu = c$. (Examples 1 and 2 illustrate this.)

Transformation of Mean and Variance

Given a random variable X with mean μ and variance σ^2, we want to calculate the mean and variance of $X^* = a_1 + a_2X$, where a_1 and a_2 are given constants. This problem is important in statistics, where it often appears.

THEOREM 2

Transformation of Mean and Variance

(a) *If a random variable X has mean* μ *and variance* σ^2, *then the random variable*

(4) $$X^* = a_1 + a_2X \qquad (a_2 > 0)$$

has the mean μ^* *and variance* σ^{*2}, *where*

(5) $$\mu^* = a_1 + a_2\mu \qquad \text{and} \qquad \sigma^{*2} = a_2^2\sigma^2.$$

(b) *In particular, the* **standardized random variable** Z *corresponding to* X, *given by*

(6)
$$Z = \frac{X - \mu}{\sigma}$$

has the mean 0 *and the variance* 1.

PROOF We prove (5) for a continuous distribution. To a small interval I of length Δx on the x-axis there corresponds the probability $f(x)\Delta x$ [approximately; the area of a rectangle of base Δx and height $f(x)$]. Then the probability $f(x)\Delta x$ must equal that for the corresponding interval on the x^*-axis, that is, $f^*(x^*)\Delta x^*$, where f^* is the density of X^* and Δx^* is the length of the interval on the x^*-axis corresponding to I. Hence for differentials we have $f^*(x^*)\,dx^* = f(x)\,dx$. Also, $x^* = a_1 + a_2 x$ by (4), so that (1b) applied to X^* gives

$$\begin{aligned}
\mu^* &= \int_{-\infty}^{\infty} x^* f^*(x^*)\,dx^* \\
&= \int_{-\infty}^{\infty} (a_1 + a_2 x) f(x)\,dx \\
&= a_1 \int_{-\infty}^{\infty} f(x)\,dx + a_2 \int_{-\infty}^{\infty} x f(x)\,dx.
\end{aligned}$$

On the right the first integral equals 1, by (10) in Sec. 24.5. The second intergral is μ. This proves (5) for μ^*. It implies

$$x^* - \mu^* = (a_1 + a_2 x) - (a_1 + a_2 \mu) = a_2(x - \mu).$$

From this and (2) applied to X^*, again using $f^*(x^*)\,dx^* = f(x)\,dx$, we obtain the second formula in (5),

$$\sigma^{*2} = \int_{-\infty}^{\infty} (x^* - \mu^*)^2 f^*(x^*)\,dx^* = a_2^2 \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx = a_2^2 \sigma^2.$$

For a discrete distribution the proof of (5) is similar.

Choosing $a_1 = -\mu/\sigma$ and $a_2 = 1/\sigma$ we obtain (6) from (4), writing $X^* = Z$. For these a_1, a_2 formula (5) gives $\mu^* = 0$ and $\sigma^{*2} = 1$, as claimed in (b). ■

Expectation, Moments

Recall that (1) defines the expectation (the mean) of X, the value of X to be expected on the average, written $\mu = E(X)$. More generally, if $g(x)$ is nonconstant and continuous for all x, then $g(X)$ is a random variable. Hence its *mathematical expectation* or, briefly, its

expectation $E(g(X))$ is the value of $g(X)$ to be expected on the average, defined [similarly to (1)] by

$$E(g(X)) = \sum_j g(x_j)f(x_j) \quad \text{or} \quad E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx. \tag{7}$$

In the first formula, f is the probability function of the discrete random variable X. In the second formula, f is the density of the continuous random variable X. Important special cases are the ***k*th moment** of X (where $k = 1, 2, \cdots$)

$$E(X^k) = \sum_j x_j^k f(x_j) \quad \text{or} \quad \int_{-\infty}^{\infty} x^k f(x)\,dx \tag{8}$$

and the ***k*th central moment** of X ($k = 1, 2, \cdots$)

$$E([X-\mu]^k) = \sum_j (x_j - \mu)^k f(x_j) \quad \text{or} \quad \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,dx. \tag{9}$$

This includes the first moment, the **mean** of X

$$\mu = E(X) \qquad [(8) \text{ with } k = 1]. \tag{10}$$

It also includes the second central moment, the **variance** of X

$$\sigma^2 = E([X-\mu]^2) \qquad [(9) \text{ with } k = 2]. \tag{11}$$

For later use you may prove

$$E(1) = 1. \tag{12}$$

PROBLEM SET 24.6

1–8 MEAN, VARIANCE

Find the mean and variance of the random variable X with probability function or density $f(x)$.

1. $f(x) = kx$ ($0 \leqq x \leqq 2$, k suitable)
2. X = Number a fair die turns up
3. Uniform distribution on $[0, 2\pi]$
4. $Y = \sqrt{3}(X - \mu)/\pi$ with X as in Prob. 3
5. $f(x) = 4e^{-4x}$ $(x \geqq 0)$
6. $f(x) = k(1 - x^2)$ if $-1 \leqq x \leqq 1$ and 0 otherwise
7. $f(x) = Ce^{-x/2}$ $\quad (x = 0)$
8. X = *Number of times a fair coin is flipped until the first Head appears.* (Calculate μ only.)
9. If the diameter X [cm] of certain bolts has the density $f(x) = k(x - 0.9)(1.1 - x)$ for $0.9 < x < 1.1$ and 0 for other x, what are k, μ, and σ^2? Sketch $f(x)$.
10. If, in Prob. 9, a defective bolt is one that deviates from 1.00 cm by more than 0.06 cm, what percentage of defectives should we expect?
11. For what choice of the maximum possible deviation from 1.00 cm shall we obtain 10% defectives in Probs. 9 and 10?
12. What total sum can you expect in rolling a fair die 20 times? Do the experiment. Repeat it a number of times and record how the sum varies.
13. What is the expected daily profit if a store sells X air conditioners per day with probability $f(10) = 0.1$, $f(11) = 0.3$, $f(12) = 0.4$, $f(13) = 0.2$ and the profit per conditioner is \$55?
14. Find the expectation of $g(X) = X^2$, where X is uniformly distributed on the interval $-1 \leqq x \leqq 1$.

15. A small filling station is supplied with gasoline every **Saturday** afternoon. Assume that its volume X of sales in ten thousands of gallons has the probability density $f(x) = 6x(1 - x)$ if $0 \leqq x \leqq 1$ and 0 otherwise. Determine the mean, the variance, and the standardized variable.

16. What capacity must the tank in Prob. 15 have in order that the probability that the tank will be emptied in a given week be 5%?

17. James rolls 2 fair dice, and Harry pays k cents to James, where k is the product of the two faces that show on the dice. How much should James pay to Harry for each game to make the game fair?

18. What is the mean life of a lightbulb whose life X [hours] has the density $f(x) = 0.001e^{-0.001x}$ $(x \geqq 0)$?

19. Let X be discrete with probability function $f(0) = f(3) = \frac{1}{8}$, $f(1) = f(2) = \frac{3}{8}$. Find the expectation of X^3.

20. TEAM PROJECT. Means, Variances, Expectations.

(a) Show that $E(X - \mu) = 0$, $\sigma^2 = E(X^2) - \mu^2$.

(b) Prove (10)–(12).

(c) Find all the moments of the uniform distribution on an interval $a \leqq x \leqq b$.

(d) The **skewness** γ of a random variable X is defined by

$$\gamma = \frac{1}{\sigma^3} E([X - \mu]^3). \tag{13}$$

Show that for a symmetric distribution (whose third central moment exists) the skewness is zero.

(e) Find the skewness of the distribution with density $f(x) = xe^{-x}$ when $x > 0$ and $f(x) = 0$ otherwise. Sketch $f(x)$.

(f) Calculate the skewness of a few simple discrete distributions of your own choice.

(g) Find a *nonsymmetric* discrete distribution with 3 possible values, mean 0, and skewness 0.

24.7 Binomial, Poisson, and Hypergeometric Distributions

These are the three most important ***discrete*** distributions, with numerous applications.

Binomial Distribution

The **binomial distribution** occurs in games of chance (rolling a die, see below, etc.), quality inspection (e.g., counting of the number of defectives), opinion polls (counting number of employees favoring certain schedule changes, etc.), medicine (e.g., recording the number of patients who recovered on a new medication), and so on. The conditions of its occurrence are as follows.

We are interested in the number of times an event A occurs in n independent trials. In each trial the event A has the same probability $P(A) = p$. Then in a trial, A will *not* occur with probability $q = 1 - p$. In n trials the random variable that interests us is

$$X = \textit{Number of times the event A occurs in n trials.}$$

X can assume the values $0, 1, \cdots, n$, and we want to determine the corresponding probabilities. Now $X = x$ means that A occurs in x trials and in $n - x$ trials it does not occur. This may look as follows.

$$\underbrace{A \quad A \cdots A}_{x \text{ times}} \qquad \underbrace{B \quad B \cdots B}_{n - x \text{ times}}. \tag{1}$$

Here $B = A^c$ is the complement of A, meaning that A does not occur (Sec. 24.2). We now use the assumption that the trials are independent, that is, they do not influence each other. Hence (1) has the probability (see Sec. 24.3 on independent events)

(1*) $$\underbrace{pp\cdots p}_{x \text{ times}} \cdot \underbrace{qq\cdots q}_{n-x \text{ times}} = p^x q^{n-x}.$$

Now (1) is just one order of arranging x A's and $n - x$ B's. We now use Theorem 1(b) in Sec. 24.4, which gives the number of permutations of n things (the n outcomes of the n trials) consisting of 2 classes, class 1 containing the $n_1 = x$ A's and class 2 containing the $n - n_1 = n - x$ B's. This number is

$$\frac{n!}{x!(n-x)!} = \binom{n}{x}.$$

Accordingly, (1*), multiplied by this binomial coefficient, gives the probability $P(X = x)$ of $X = x$, that is, of obtaining A precisely x times in n trials. Hence X has the probability function

(2) $$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n)$$

and $f(x) = 0$ otherwise. The distribution of X with probability function (2) is called the **binomial distribution** or *Bernoulli distribution*. The occurrence of A is called *success* (regardless of what it actually is; it may mean that you miss your plane or lose your watch) and the nonoccurrence of A is called *failure*. Figure 517 shows typical examples. Numeric values can be obtained from Table A5 in App. 5 or from your CAS.

The mean of the binomial distribution is (see Team Project 16)

(3) $$\mu = np$$

and the variance is (see Team Project 16)

(4) $$\sigma^2 = npq.$$

For the ***symmetric case*** of equal chance of success and failure ($p = q = \frac{1}{2}$) this gives the mean $n/2$, the variance $n/4$, and the probability function

(2*) $$f(x) = \binom{n}{x}\left(\frac{1}{2}\right)^n \qquad (x = 0, 1, \cdots, n).$$

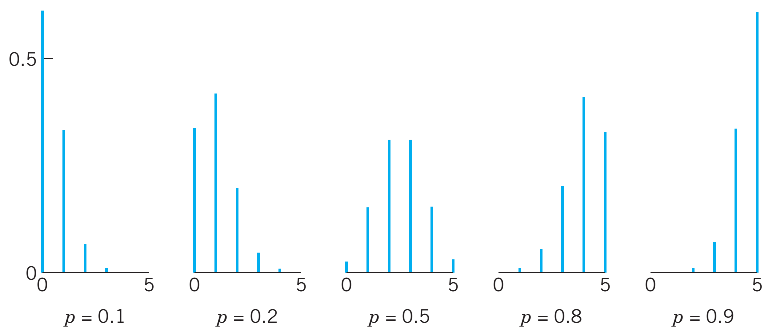


Fig. 517. Probability function (2) of the binomial distribution for $n = 5$ and various values of p

EXAMPLE 1 Binomial Distribution

Compute the probability of obtaining at least two *"Six"* in rolling a fair die 4 times.

Solution. $p = P(A) = P(\text{"}Six\text{"}) = \frac{1}{6}$, $q = \frac{5}{6}$, $n = 4$. The event *"At least two 'Six'"* occurs if we obtain 2 or 3 or 4 "Six." Hence the answer is

$$P = f(2) + f(3) + f(4) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right) + \binom{4}{4}\left(\frac{1}{6}\right)^4$$

$$= \frac{1}{6^4}(6 \cdot 25 + 4 \cdot 5 + 1) = \frac{171}{1296} = 13.2\%.$$

Poisson Distribution

The discrete distribution with infinitely many possible values and probability function

$$f(x) = \frac{\mu^x}{x!}e^{-\mu} \qquad (x = 0, 1, \cdots) \tag{5}$$

is called the **Poisson distribution**, named after S. D. Poisson (Sec. 18.5). Figure 518 shows (5) for some values of μ. It can be proved that this distribution is obtained as a limiting case of the binomial distribution, if we let $p \to 0$ and $n \to \infty$ so that the mean $\mu = np$ approaches a finite value. (For instance, $\mu = np$ may be kept constant.) The Poisson distribution has the mean μ and the variance (see Team Project 16)

$$\sigma^2 = \mu. \tag{6}$$

Figure 518 gives the impression that, with increasing mean, the spread of the distribution increases, thereby illustrating formula (6), and that the distribution becomes more and more (approximately) symmetric.

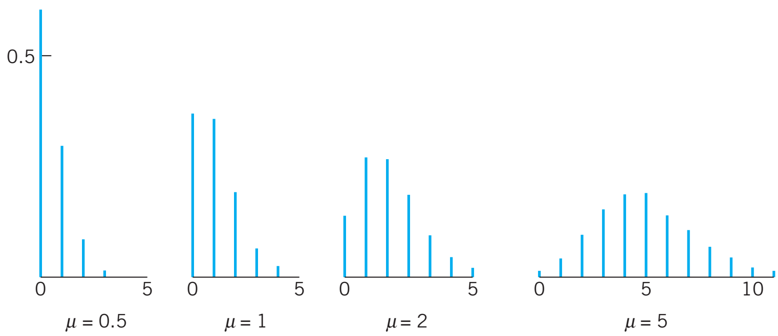


Fig. 518. Probability function (5) of the Poisson distribution for various values of μ

EXAMPLE 2 Poisson Distribution

If the probability of producing a defective screw is $p = 0.01$, what is the probability that a lot of 100 screws will contain more than 2 defectives?

Solution. The complementary event is A^c: *Not more than* 2 *defectives.* For its probability we get, from the binomial distribution with mean $\mu = np = 1$, the value [see (2)]

$$P(A^c) = \binom{100}{0}0.99^{100} + \binom{100}{1}0.01 \cdot 0.99^{99} + \binom{100}{2}0.01^2 \cdot 0.99^{98}.$$

Since p is very small, we can approximate this by the much more convenient Poisson distribution with mean $\mu = np = 100 \cdot 0.01 = 1$, obtaining [see (5)]

$$P(A^c) \approx e^{-1}\,(1 + 1 + \tfrac{1}{2})$$
$$= 91.97\%.$$

Thus $P(A) = 8.03\%$. Show that the binomial distribution gives $P(A) = 7.94\%$, so that the Poisson approximation is quite good. ■

EXAMPLE 3 Parking Problems. Poisson Distribution

If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute 4 or more cars will enter the lot?

Solution. To understand that the Poisson distribution is a model of the situation, we imagine the minute to be divided into very many short time intervals, let p be the (constant) probability that a car will enter the lot during any such short interval, and assume independence of the events that happen during those intervals. Then we are dealing with a binomial distribution with very large n and very small p, which we can approximate by the Poisson distribution with

$$\mu = np = 2,$$

because 2 cars enter on the average. The complementary event of the event "4 cars or more during a given minute" is "3 *cars or fewer enter the lot*" and has the probability

$$f(0) + f(1) + f(2) + f(3) = e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right)$$
$$= 0.857.$$

Answer: 14.3%. (Why did we consider that complement?) ■

Sampling with Replacement

This means that we draw things from a given set one by one, and after each trial we replace the thing drawn (put it back to the given set and mix) before we draw the next thing. This guarantees independence of trials and leads to the **binomial distribution**. Indeed, if a box contains N things, for example, screws, M of which are defective, the probability of drawing a defective screw in a trial is $p = M/N$. Hence the probability of drawing a nondefective screw is $q = 1 - p = 1 - M/N$, and (2) gives the probability of drawing x defectives in n trials in the form

$$f(x) = \binom{n}{x}\left(\frac{M}{N}\right)^x\left(1 - \frac{M}{N}\right)^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{7}$$

Sampling without Replacement. Hypergeometric Distribution

Sampling without replacement means that we return no screw to the box. Then we no longer have independence of trials (why?), and instead of (7) the probability of drawing x defectives in n trials is

(8)
$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (x = 0, 1, \cdots, n).$$

The distribution with this probability function is called the **hypergeometric distribution** (because its moment generating function (see Team Project 16) can be expressed by the hypergeometric function defined in Sec. 5.4, a fact that we shall not use).

Derivation of (8). By (4a) in Sec. 24.4 there are

(a) $\binom{N}{n}$ different ways of picking n things from N,

(b) $\binom{M}{x}$ different ways of picking x defectives from M,

(c) $\binom{N-M}{n-x}$ different ways of picking $n - x$ nondefectives from $N - M$,

and each way in (b) combined with each way in (c) gives the total number of mutually exclusive ways of obtaining x defectives in n drawings without replacement. Since (a) is the total number of outcomes and we draw at random, each such way has the probability $1\Big/\binom{N}{n}$. From this, (8) follows. ■

The hypergeometric distribution has the mean (Team Project 16)

(9)
$$\mu = n\frac{M}{N}$$

and the variance

(10)
$$\sigma^2 = \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

EXAMPLE 4 Sampling with and without Replacement

We want to draw random samples of two gaskets from a box containing 10 gaskets, three of which are defective. Find the probability function of the random variable X = *Number of defectives in the sample*.

Solution. We have $N = 10$, $M = 3$, $N - M = 7$, $n = 2$. For sampling with replacement, (7) yields

$$f(x) = \binom{2}{x}\left(\frac{3}{10}\right)^x\left(\frac{7}{10}\right)^{2-x}, \qquad f(0) = 0.49, \quad f(1) = 0.42, \quad f(2) = 0.09.$$

For sampling without replacement we have to use (8), finding

$$f(x) = \binom{3}{x}\binom{7}{2-x}\Big/\binom{10}{2}, \qquad f(0) = f(1) = \frac{21}{45} \approx 0.47, \quad f(2) = \frac{3}{45} \approx 0.07.$$ ■

If N, M, and $N - M$ are large compared with n, then it does not matter too much whether we sample with or without replacement, and in this case the hypergeometric distribution may be approximated by the binomial distribution (with $p = M/N$), *which is somewhat simpler.*

Hence, in sampling from an indefinitely large population ("**infinite population**"), *we may use the binomial distribution, regardless of whether we sample with or without replacement.*

PROBLEM SET 24.7

1. Mark the positions of μ in Fig. 517. Comment.
2. Graph (2) for $n = 8$ as in Fig. 517 and compare with Fig. 517.
3. In Example 3, if 5 cars enter the lot on the average, what is the probability that during any given minute 6 or more cars will enter? First guess. Compare with Example 3.
4. How do the probabilities in Example 4 of the text change if you double the numbers: drawing 4 gaskets from 20, 6 of which are defective? First guess.
5. Five fair coins are tossed simultaneously. Find the probability function of the random variable $X =$ *Number of heads* and compute the probabilities of obtaining no heads, precisely 1 head, at least 1 head, not more than 4 heads.
6. Suppose that 4% of steel rods made by a machine are defective, the defectives occurring at random during production. If the rods are packaged 100 per box, what is the Poisson approximation of the probability that a given box will contain $x = 0, 1, \cdots, 5$ defectives?
7. Let X be the number of cars per minute passing a certain point of some road between 8 A.M. and 10 A.M. on a Sunday. Assume that X has a Poisson distribution with mean 5. Find the probability of observing 4 or fewer cars during any given minute.
8. Suppose that a telephone switchboard of some company on the average handles 300 calls per hour, and that the board can make at most 10 connections per minute. Using the Poisson distribution, estimate the probability that the board will be overtaxed during a given minute. (Use Table A6 in App. 5 or your CAS.)
9. **Rutherford–Geiger experiments.** In 1910, E. Rutherford and H. Geiger showed experimentally that the number of alpha particles emitted per second in a radioactive process is a random variable X having a Poisson distribution. If X has mean 0.5, what is the probability of observing two or more particles during any given second?
10. Let $p = 2\%$ be the probability that a certain type of lightbulb will fail in a 24-hour test. Find the probability that a sign consisting of 15 such bulbs will burn 24 hours with no bulb failures.
11. Guess how much less the probability in Prob. 10 would be if the sign consisted of 100 bulbs. Then calculate.
12. Suppose that a certain type of magnetic tape contains, on the average, 2 defects per 100 meters. What is the probability that a roll of tape 300 meters long will contain **(a)** x defects, **(b)** no defects?
13. Suppose that a test for extrasensory perception consists of naming (in any order) 3 cards randomly drawn from a deck of 13 cards. Find the probability that by chance alone, the person will correctly name **(a)** no cards, **(b)** 1 card, **(c)** 2 cards, **(d)** 3 cards.
14. If a ticket office can serve at most 4 customers per minute and the average number of customers is 120 per hour, what is the probability that during a given minute customers will have to wait? (Use the Poisson distribution, Table 6 in Appendix 5.)
15. Suppose that in the production of 60-ohm radio resistors, nondefective items are those that have a resistance between 58 and 62 ohms and the probability of a resistor's being defective is 0.1%. The resistors are sold in lots of 200, with the guarantee that all resistors are nondefective. What is the probability that a given lot will violate this guarantee? (Use the Poisson distribution.)
16. **TEAM PROJECT. Moment Generating Function.** The moment generating function $G(t)$ is defined by

$$G(t) = E(e^{tX_j}) = \sum_j e^{tx_j} f(x_j)$$

or

$$G(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

where X is a discrete or continuous random variable, respectively.

(a) Assuming that termwise differentiation and differentiation under the integral sign are permissible, show that $E(X^k) = G^{(k)}(0)$, where $G^{(k)} = d^kG/dt^k$, in particular, $\mu = G'(0)$.

(b) Show that the binomial distribution has the moment generating function

$$G(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x} = \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x q^{n-x}$$
$$= (pe^t + q)^n.$$

(c) Using (b), prove (3).

(d) Prove (4).

(e) Show that the Poisson distribution has the moment generating function $G(t) = e^{-\mu}e^{\mu e^t}$ and prove (6).

(f) Prove $x\binom{M}{x} = M\binom{M-1}{x-1}$.

Using this, prove (9).

17. Multinomial distribution. Suppose a trial can result in precisely one of k mutually exclusive events $A_1, \cdots, A_k$ with probabilities $p_1, \cdots, p_k$, respectively, where $p_1 + \cdots + p_k = 1$. Suppose that n independent trials are performed. Show that the probability of getting x_1 A_1's, $\cdots$, x_k A_k's is

$$f(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!}p_1^{x_1} \cdots p_k^{x_k}$$

where $0 \leqq x_j \leqq n$, $j = 1, \cdots, k$, and $x_1 + \cdots + x_k = n$. The distribution having this probability function is called the *multinomial distribution.*

18. A process of manufacturing screws is checked every hour by inspecting n screws selected at random from that hour's production. If one or more screws are defective, the process is halted and carefully examined. How large should n be if the manufacturer wants the probability to be about 95% that the process will be halted when 10% of the screws being produced are defective? (Assume independence of the quality of any screw from that of the other screws.)

24.8 Normal Distribution

Turning from discrete to continuous distributions, in this section we discuss the normal distribution. This is the most important continuous distribution because in applications many random variables are **normal random variables** (that is, they have a normal distribution) or they are approximately normal or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions, and it also occurs in the proofs of various statistical tests.

The **normal distribution** or *Gauss distribution* is defined as the distribution with the density

$$\textbf{(1)} \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0)$$

where exp is the exponential function with base $e = 2.718 \cdots$. This is simpler than it may at first look. $f(x)$ has these features (see also Fig. 519).

1. μ is the mean and σ the standard deviation.

2. $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be by (10), Sec. 24.5.

3. The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$ (Fig. 519, *"bell-shaped curves"*).

4. The exponential function in (1) goes to zero very fast—the faster the smaller the standard deviation σ is, as it should be (Fig. 519).

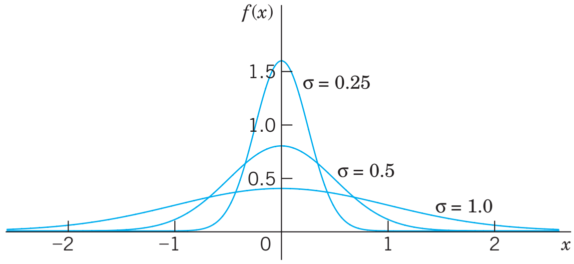


Fig. 519. Density (1) of the normal distribution with $\mu = 0$ for various values of σ

Distribution Function *F*(x)

From (7) in Sec. 24.5 and (1) we see that the normal distribution has the **distribution function**

(2)
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv.$$

Here we needed x as the upper limit of integration and wrote v (instead of x) in the integrand.

For the corresponding **standardized normal distribution** with mean 0 and standard deviation 1 we denote $F(x)$ by $\Phi(z)$. Then we simply have from (2)

(3)
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du.$$

This integral cannot be integrated by one of the methods of calculus. But this is no serious handicap because its values can be obtained from Table A7 in App. 5 or from your CAS. These values are needed in working with the normal distribution. The curve of $\Phi(z)$ is *S*-shaped. It increases monotone (why?) from 0 to 1 and intersects the vertical axis at $\frac{1}{2}$ (why?), as shown in Fig. 520.

Relation Between $F(x)$ and $\Phi(z)$. Although your CAS will give you values of $F(x)$ in (2) with any μ and σ directly, it is important to comprehend that and why any such an $F(x)$ can be expressed in terms of the tabulated standard $\Phi(z)$, as follows.

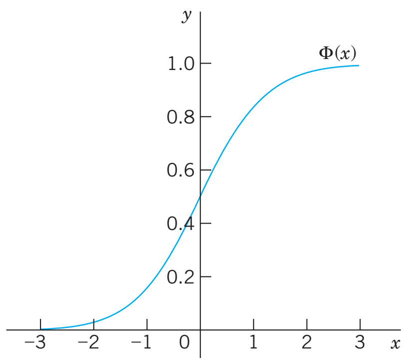


Fig. 520. Distribution function $\Phi(z)$ of the normal distribution with mean 0 and variance 1

THEOREM 1

Use of the Normal Table A7 in App. 5

The distribution function $F(x)$ of the normal distribution with any μ and σ [see (2)] *is related to the standardized distribution function $\Phi(z)$ in* (3) *by the formula*

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{4}$$

PROOF Comparing (2) and (3) we see that we should set

$$u = \frac{v-\mu}{\sigma}. \qquad \text{Then } v = x \text{ gives} \qquad u = \frac{x-\mu}{\sigma}$$

as the new upper limit of integration. Also $v - \mu = \sigma u$, thus $dv = \sigma\, du$. Together, since σ drops out,

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-u^2/2} \sigma\, du = \Phi\left(\frac{x-\mu}{\sigma}\right).$$ ■

Probabilities corresponding to intervals will be needed quite frequently in statistics in Chap. 25. These are obtained as follows.

THEOREM 2

Normal Probabilities for Intervals

The probability that a normal random variable X with mean μ and standard deviation σ assume any value in an interval $a < x \leqq b$ is

$$P(a < X \leqq b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{5}$$

PROOF Formula (2) in Sec. 24.5 gives the first equality in (5), and (4) in this section gives the second equality. ■

Numeric Values

In practical work with the normal distribution it is good to remember that about $\frac{2}{3}$ of all values of X to be observed will lie between $\mu \pm \sigma$, about 95% between $\mu \pm 2\sigma$, and practically all between the **three-sigma limits** $\mu \pm 3\sigma$. More precisely, by Table A7 in App. 5,

$$\text{(6)} \quad \begin{aligned} &\text{(a)} \quad P(\mu - \sigma < X \leqq \mu + \sigma) \approx 68\% \\ &\text{(b)} \quad P(\mu - 2\sigma < X \leqq \mu + 2\sigma) \approx 95.5\% \\ &\text{(c)} \quad P(\mu - 3\sigma < X \leqq \mu + 3\sigma) \approx 99.7\%. \end{aligned}$$

Formulas (6a) and (6b) are illustrated in Fig. 521.

The formulas in (6) show that a value deviating from μ by more than σ, 2σ, or 3σ will occur in one of about 3, 20, and 300 trials, respectively.

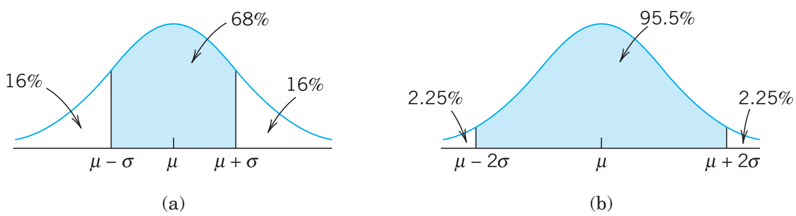


Fig. 521. Illustration of formula (6)

In tests (Chap. 25) we shall ask, conversely, for the intervals that correspond to certain given probabilities; practically most important are the probabilities of 95%, 99%, and 99.9%. For these, Table A8 in App. 5 gives the answers $\mu \pm 2\sigma$, $\mu \pm 2.6\sigma$, and $\mu \pm 3.3\sigma$, respectively. More precisely,

$$
\textbf{(7)} \qquad
\begin{array}{ll}
\textbf{(a)} & P(\mu - 1.96\sigma < X \leqq \mu + 1.96\sigma) = 95\% \\
\textbf{(b)} & P(\mu - 2.58\sigma < X \leqq \mu + 2.58\sigma) = 99\% \\
\textbf{(c)} & P(\mu - 3.29\sigma < X \leqq \mu + 3.29\sigma) = 99.9\%.
\end{array}
$$

Working with the Normal Tables A7 and A8 in App. 5

There are two normal tables in App. 5, Tables A7 and A8. If you want probabilities, use Table A7. If probabilities are given and corresponding intervals or x-values are wanted, use Table A8. The following examples are typical. Do them with care, verifying all values, and don't just regard them as dull exercises for your software. Make sketches of the density to see whether the results look reasonable.

EXAMPLE 1 Reading Entries from Table A7

If X is standardized normal (so that $\mu = 0$, $\sigma = 1$), then

$$P(X \leqq 2.44) = 0.9927 \approx 99\tfrac{1}{4}\,\%$$

$$P(X \leqq -1.16) = 1 - \Phi(1.16) = 1 - 0.8770 = 0.1230 = 12.3\%$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - 0.8413 = 0.1587) \text{ by (7), Sec. 24.3}$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(1.8) - \Phi(1.0) = 0.9641 - 0.8413 = 0.1228. \qquad \blacksquare$$

EXAMPLE 2 Probabilities for Given Intervals, Table A7

Let X be normal with mean 0.8 and variance 4 (so that $\sigma = 2$). Then by (4) and (5)

$$P(X \leqq 2.44) = F(2.44) = \Phi\left(\frac{2.44 - 0.80}{2}\right) = \Phi(0.82) = 0.7939 \approx 80\%$$

or, if you like it better, (similarly in the other cases)

$$P(X \leqq 2.44) = P\left(\frac{X - 0.80}{2} \leqq \frac{2.44 - 0.80}{2}\right) = P(Z \leqq 0.82) = 0.7939$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - \Phi\left(\frac{1 - 0.8}{2}\right) = 1 - 0.5398 = 0.4602$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(0.5) - \Phi(0.1) = 0.6915 - 0.5398 = 0.1517. \qquad \blacksquare$$

EXAMPLE 3 Unknown Values c for Given Probabilities, Table A8

Let X be normal with mean 5 and variance 0.04 (hence standard deviation 0.2). Find c or k corresponding to the given probability

$$P(X \leqq c) = 95\%, \qquad \Phi\left(\frac{c-5}{0.2}\right) = 95\%, \qquad \frac{c-5}{0.2} = 1.645, \qquad c = 5.329$$

$$P(5 - k \leqq X \leqq 5 + k) = 90\%, \qquad 5 + k = 5.329 \qquad \text{(as before; why?)}$$

$$P(X \geqq c) = 1\%, \qquad \text{thus } P(X \leqq c) = 99\%, \qquad \frac{c-5}{0.2} = 2.326, \qquad c = 5.465. \qquad \blacksquare$$

EXAMPLE 4 Defectives

In a production of iron rods let the diameter X be normally distributed with mean 2 in. and standard deviation 0.008 in.

(a) What percentage of defectives can we expect if we set the tolerance limits at 2 ± 0.02 in.?

(b) How should we set the tolerance limits to allow for 4% defectives?

Solution. **(a)** $1\frac{1}{4}\%$ because from (5) and Table A7 we obtain for the complementary event the probability

$$\begin{aligned} P(1.98 \leqq X \leqq 2.02) &= \Phi\left(\frac{2.02 - 2.00}{0.008}\right) - \Phi\left(\frac{1.98 - 2.00}{0.008}\right) \\ &= \Phi(2.5) - \Phi(-2.5) \\ &= 0.9938 - (1 - 0.9938) \\ &= 0.9876 \\ &= 98\tfrac{3}{4}\%. \end{aligned}$$

(b) 2 ± 0.0164 because, for the complementary event, we have

$$0.96 = P(2 - c \leqq X \leqq 2 + c)$$

or

$$0.98 = P(X \leqq 2 + c)$$

so that Table A8 gives

$$0.98 = \Phi\left(\frac{2 + c - 2}{0.008}\right),$$

$$\frac{2 + c - 2}{0.008} = 2.054, \qquad c = 0.0164. \qquad \blacksquare$$

Normal Approximation of the Binomial Distribution

The probability function of the binomial distribution is (Sec. 24.7)

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{8}$$

If n is large, the binomial coefficients and powers become very inconvenient. It is of great practical (and theoretical) importance that, in this case, the normal distribution provides a good approximation of the binomial distribution, according to the following theorem, one of the most important theorems in all probability theory.

THEOREM 3

Limit Theorem of De Moivre and Laplace

For large n,

$$f(x) \sim f^*(x) \qquad (x = 0, 1, \cdots, n). \tag{9}$$

Here f is given by (8). *The function*

$$f^*(x) = \frac{1}{\sqrt{2\pi}\sqrt{npq}}\, e^{-z^2/2}, \qquad z = \frac{x - np}{\sqrt{npq}} \tag{10}$$

is the density of the normal distribution with mean $\mu = np$ *and variance* $\sigma^2 = npq$ (*the mean and variance of the binomial distribution*). *The symbol* $\sim$ (read **asymptotically equal**) *means that the ratio of both sides approaches* 1 *as n approaches* ∞. *Furthermore, for any nonnegative integers a and b* ($> a$),

$$\begin{gathered} P(a \leqq X \leqq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \sim \Phi(\beta) - \Phi(\alpha), \\ \alpha = \frac{a - np - 0.5}{\sqrt{npq}}, \qquad \beta = \frac{b - np + 0.5}{\sqrt{npq}}. \end{gathered} \tag{11}$$

A proof of this theorem can be found in [G3] listed in App. 1. The proof shows that the term 0.5 in α and β is a correction caused by the change from a discrete to a continuous distribution.

PROBLEM SET 24.8

1. Let X be normal with mean 10 and variance 4. Find $P(X > 12)$, $P(X < 10)$, $P(X < 11)$, $P(9 < X < 13)$.

2. Let X be normal with mean 105 and variance 25. Find $P(X \leqq 112.5)$, $P(x > 100)$, $P(110.5 < X < 111.25)$.

3. Let X be normal with mean 50 and variance 9. Determine c such that $P(X < c) = 5\%$, $P(X > c) = 1\%$, $P(50 - c < X < 50 + c) = 50\%$.

4. Let X be normal with mean 3.6 and variance 0.01. Find c such that $P(X \leqq c) = 50\%$, $P(X > c) = 10\%$, $P(-c < X - 3.6 \leqq c) = 99.9\%$.

5. If the lifetime X of a certain kind of automobile battery is normally distributed with a mean of 5 years and a standard deviation of 1 year, and the manufacturer wishes to guarantee the battery for 4 years, what percentage of the batteries will he have to replace under the guarantee?

6. If the standard deviation in Prob. 5 were smaller, would that percentage be larger or smaller?

7. A manufacturer knows from experience that the resistance of resistors he produces is normal with mean $\mu = 150\ \Omega$ and standard deviation $\sigma = 5\ \Omega$. What percentage of the resistors will have resistance between $148\ \Omega$ and $152\ \Omega$? Between $140\ \Omega$ and $160\ \Omega$?

8. The breaking strength X [kg] of a certain type of plastic block is normally distributed with a mean of 1500 kg and a standard deviation of 50 kg. What is the maximum load such that we can expect no more than 5% of the blocks to break?

9. If the mathematics scores of the SAT college entrance exams are normal with mean 480 and standard deviation 100 (these are about the actual values over the past years) and if some college sets 500 as the minimum score for new students, what percent of students would not reach that score?

10. A producer sells electric bulbs in cartons of 1000 bulbs. Using (11), find the probability that any given carton contains not more than 1% defective bulbs, assuming the production process to be a Bernoulli experiment with $p = 1\%$ (= probability that any given bulb will be defective). First guess. Then calculate.

11. If sick-leave time X used by employees of a company in one month is (very roughly) normal with mean 1000 hours and standard deviation 100 hours, how much time t should be budgeted for sick leave during the next month if t is to be exceeded with probability of only 20%?

12. If the monthly machine repair and maintenance cost X in a certain factory is known to be normal with mean \$12,000 and standard deviation \$2000, what is the probability that the repair cost for the next month will exceed the budgeted amount of \$15,000?

13. If the resistance X of certain wires in an electrical network is normal with mean $0.01\ \Omega$ and standard deviation $0.001\ \Omega$, how many of 1000 wires will meet the specification that they have resistance between 0.009 and $0.011\ \Omega$?

14. TEAM PROJECT. Normal Distribution. (a) Derive the formulas in (6) and (7) from the appropriate normal table.

(b) Show that $\Phi(-z) = 1 - \Phi(z)$. Give an example.

(c) Find the points of inflection of the curve of (1).

(d) Considering $\Phi^2(\infty)$ and introducing polar coordinates in the double integral (a standard trick worth remembering), prove

$$\Phi(\infty) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du = 1. \tag{12}$$

(e) Show that σ in (1) is indeed the standard deviation of the normal distribution. [Use (12).]

(f) Bernoulli's law of large numbers. In an experiment let an event A have probability p $(0 < p < 1)$, and let X be the number of times A happens in n independent trials. Show that for any given $\epsilon > 0$,

$$P\left(\left|\frac{X}{n} - p\right| \leqq \epsilon\right) \to 1 \qquad \text{as } n \to \infty.$$

(g) Transformation. If X is normal with mean μ and variance σ^2, show that $X^* = c_1X + c_2$ $(c_1 > 0)$ is normal with mean $\mu^* = c_1\mu + c_2$ and variance $\sigma^{*2} = c_1^2\sigma^2$.

15. WRITING PROJECT. Use of Tables, Use of CAS. Give a systematic discussion of the use of Tables A7 and A8 for obtaining $P(X < b)$, $P(X > a)$, $P(a < X < b)$, $P(X < c) = k$, $P(X > c) = k$, as well as $P(\mu - c < X < \mu + c) = k$; include simple examples. If you have a CAS, describe to what extent it makes the use of those tables superfluous; give examples.

24.9 Distributions of Several Random Variables

Distributions of two or more random variables are of interest for two reasons:

1. They occur in experiments in which we observe several random variables, for example, carbon content X and hardness Y of steel, amount of fertilizer X and yield of corn Y, height X_1, weight X_2, and blood pressure X_3 of persons, and so on.

2. They will be needed in the mathematical justification of the methods of statistics in Chap. 25.

In this section we consider two random variables X and Y or, as we also say, a **two-dimensional random variable** (X, Y). For (X, Y) the outcome of a trial is a pair of numbers $X = x$, $Y = y$, briefly $(X, Y) = (x, y)$, which we can plot as a point in the XY-plane.

The **two-dimensional probability distribution** of the random variable (X, Y) is given by the **distribution function**

$$F(x, y) = P(X \leqq x, Y \leqq y). \tag{1}$$

This is the probability that in a trial, X will assume any value not greater than x and in the same trial, Y will assume any value not greater than y. This corresponds to the blue region in Fig. 522, which extends to $-\infty$ to the left and below. $F(x, y)$ determines the

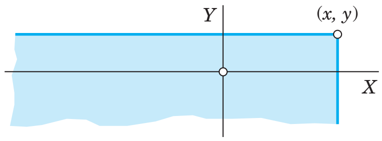


Fig. 522. Formula (1)

probability distribution uniquely, because in analogy to formula (2) in Sec. 24.5, that is, $P(a < X \leqq b) = F(b) - F(a)$, we now have for a rectangle (see Prob. 16)

$$(2) \quad P(a_1 < X \leqq b_1, \quad a_2 < Y \leqq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2).$$

As before, in the two-dimensional case we shall also have discrete and continuous random variables and distributions.

Discrete Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5), we call (X, Y) and its distribution **discrete** if (X, Y) can assume only finitely many or at most countably infinitely many pairs of values $(x_1, y_1), (x_2, y_2), \cdots$ with positive probabilities, whereas the probability for any domain containing none of those values of (X, Y) is zero.

Let (x_i, y_j) be any of those pairs and let $P(X = x_i, Y = y_j) = p_{ij}$ (where we admit that p_{ij} may be 0 for certain pairs of subscripts i, j). Then we define the **probability function** $f(x, y)$ of (X, Y) by

$$(3) \qquad f(x, y) = p_{ij} \quad \text{if} \quad x = x_i, y = y_j \qquad \text{and} \qquad f(x, y) = 0 \quad \text{otherwise;}$$

here, $i = 1, 2, \cdots$ and $j = 1, 2, \cdots$ independently. In analogy to (4), Sec. 24.5, we now have for the distribution function the formula

$$\textbf{(4)} \qquad F(x, y) = \sum_{x_i \leqq x} \sum_{y_j \leqq y} f(x_i, y_j).$$

Instead of (6) in Sec. 24.5 we now have the condition

$$(5) \qquad \sum_i \sum_j f(x_i, y_j) = 1.$$

EXAMPLE 1 Two-Dimensional Discrete Distribution

If we simultaneously toss a dime and a nickel and consider

$$X = \textit{Number of heads the dime turns up},$$

$$Y = \textit{Number of heads the nickel turns up},$$

then X and Y can have the values 0 or 1, and the probability function is

$$f(0, 0) = f(1, 0) = f(0, 1) = f(1, 1) = \tfrac{1}{4}, \quad f(x, y) = 0 \text{ otherwise.}$$ ■

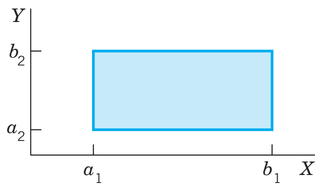


Fig. 523. Notion of a two-dimensional distribution

Continuous Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5) we call (X, Y) and its distribution **continuous** if the corresponding distribution function $F(x, y)$ can be given by a double integral

(6)
$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x^*, y^*)\, dx^*\, dy^*$$

whose integrand f, called the **density** of (X, Y), is nonnegative everywhere, and is continuous, possibly except on finitely many curves.

From (6) we obtain the probability that (X, Y) assume any value in a rectangle (Fig. 523) given by the formula

(7)
$$P(a_1 < X \leqq b_1, \quad a_2 < Y \leqq b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x, y)\, dx\, dy.$$

EXAMPLE 2 Two-Dimensional Uniform Distribution in a Rectangle

Let R be the rectangle $\alpha_1 < x \leqq \beta_1$, $\alpha_2 < y \leqq \beta_2$. The density (see Fig. 524)

(8)
$$f(x, y) = 1/k \quad \text{if } (x, y) \text{ is in } R, \qquad f(x, y) = 0 \text{ otherwise}$$

defines the so-called **uniform distribution** *in the rectangle R;* here $k = (\beta_1 - \alpha_1)(\beta_2 - \alpha_2)$ is the area of R. The distribution function is shown in Fig. 525. ■

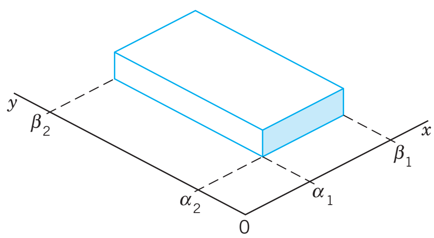


Fig. 524. Density function (8) of the uniform distribution

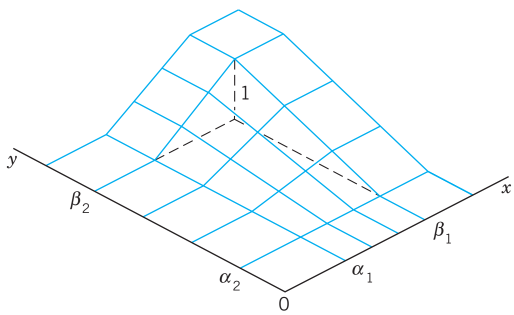


Fig. 525. Distribution function of the uniform distribution defined by (8)

Marginal Distributions of a Discrete Distribution

This is a rather natural idea, without counterpart for a single random variable. It amounts to being interested only in one of the two variables in (X, Y), say, X, and asking for its distribution, called the **marginal distribution** of X in (X, Y). So we ask for the probability

$P(X = x, Y \text{ arbitrary})$. Since (X, Y) is discrete, so is X. We get its probability function, call it $f_1(x)$, from the probability function $f(x, y)$ of (X, Y) by summing over y:

$$f_1(x) = P(X = x, Y \text{ arbitrary}) = \sum_y f(x, y) \tag{9}$$

where we sum all the values of $f(x, y)$ that are not 0 for that x.

From (9) we see that the distribution function of the marginal distribution of X is

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) = \sum_{x^* \leqq x} f_1(x^*). \tag{10}$$

Similarly, the probability function

$$f_2(y) = P(X \text{ arbitrary}, Y \leqq y) = \sum_x f(x, y) \tag{11}$$

determines the **marginal distribution** of Y in (X, Y). Here we sum all the values of $f(x, y)$ that are not zero for the corresponding y. The distribution function of this marginal distribution is

$$F_2(y) = P(X \text{ arbitrary}, Y \leqq y) = \sum_{y^* \leqq y} f_2(y^*). \tag{12}$$

EXAMPLE 3 Marginal Distributions of a Discrete Two-Dimensional Random Variable

In drawing 3 cards with replacement from a bridge deck let us consider

$$(X, Y), \qquad X = \textit{Number of queens}, \qquad Y = \textit{Number of kings or aces}.$$

The deck has 52 cards. These include 4 queens, 4 kings, and 4 aces. Hence in a single trial a queen has probability $\frac{4}{52} = \frac{1}{13}$ and a king or ace $\frac{8}{52} = \frac{2}{13}$. This gives the probability function of (X, Y),

$$f(x, y) = \frac{3!}{x!y!(3 - x - y)!}\left(\frac{1}{13}\right)^x\left(\frac{2}{13}\right)^y\left(\frac{10}{13}\right)^{3-x-y} \qquad (x + y \leqq 3)$$

and $f(x, y) = 0$ otherwise. Table 24.1 shows in the center the values of $f(x, y)$ and on the right and lower margins the values of the probability functions $f_1(x)$ and $f_2(y)$ of the marginal distributions of X and Y, respectively. ■

Table 24.1 Values of the Probability Functions $f(x, y)$, $f_1(x)$, $f_2(y)$ in Drawing Three Cards with Replacement from a Bridge Deck, where X is the Number of Queens Drawn and Y is the Number of Kings or Aces Drawn

x \ y	0	1	2	3	$f_1(x)$
0	$\frac{1000}{2197}$	$\frac{600}{2197}$	$\frac{120}{2197}$	$\frac{8}{2197}$	$\frac{1728}{2197}$
1	$\frac{300}{2197}$	$\frac{120}{2197}$	$\frac{12}{2197}$	0	$\frac{432}{2197}$
2	$\frac{30}{2197}$	$\frac{6}{2197}$	0	0	$\frac{36}{2197}$
3	$\frac{1}{2197}$	0	0	0	$\frac{1}{2197}$
$f_2(y)$	$\frac{1331}{2197}$	$\frac{726}{2197}$	$\frac{132}{2197}$	$\frac{8}{2197}$	

Marginal Distributions of a Continuous Distribution

This is conceptually the same as for discrete distributions, with probability functions and sums replaced by densities and integrals. For a continuous random variable (X, Y) with density $f(x, y)$ we now have the **marginal distribution** of X in (X, Y), defined by the distribution function

$$F_1(x) = P(X \leqq x, -\infty < Y < \infty) = \int_{-\infty}^{x} f_1(x^*)\, dx^* \tag{13}$$

with the density f_1 of X obtained from $f(x, y)$ by integration over y,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy. \tag{14}$$

Interchanging the roles of X and Y, we obtain the **marginal distribution** of Y in (X, Y) with the distribution function

$$F_2(y) = P(-\infty < X < \infty, Y \leqq y) = \int_{-\infty}^{y} f_2(y^*)\, dy^* \tag{15}$$

and density

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx. \tag{16}$$

Independence of Random Variables

X and Y in a (discrete or continuous) random variable (X, Y) are said to be **independent** if

$$F(x, y) = F_1(x)F_2(y) \tag{17}$$

holds for all (x, y). Otherwise these random variables are said to be **dependent**. These definitions are suggested by the corresponding definitions for events in Sec. 24.3.

Necessary and sufficient for independence is

$$f(x, y) = f_1(x)f_2(y) \tag{18}$$

for all x and y. Here the f's are the above probability functions if (X, Y) is discrete or those densities if (X, Y) is continuous. (See Prob. 20.)

EXAMPLE 4 Independence and Dependence

In tossing a dime and a nickel, X = *Number of heads on the dime*, Y = *Number of heads on the nickel* may assume the values 0 or 1 and are independent. The random variables in Table 24.1 are dependent. ■

Extension of Independence to *n*-Dimensional Random Variables. This will be needed throughout Chap. 25. The distribution of such a random variable $\mathbf{X} = (X_1, \cdots, X_n)$ is determined by a **distribution function** of the form

$$F(x_1, \cdots, x_n) = P(X_1 \leqq x_1, \cdots, X_n \leqq x_n).$$

The random variables $X_1, \cdots, X_n$ are said to be **independent** if

$$F(x_1, \cdots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n) \tag{19}$$

for all $(x_1, \cdots, x_n)$. Here $F_j(x_j)$ is the distribution function of the marginal distribution of X_j in $\mathbf{X}$, that is,

$$F_j(x_j) = P(X_j \leqq x_j, X_k \text{ arbitrary}, k = 1, \cdots, n, k \neq j).$$

Otherwise these random variables are said to be **dependent**.

Functions of Random Variables

When $n = 2$, we write $X_1 = X$, $X_2 = Y$, $x_1 = x$, $x_2 = y$. Taking a nonconstant continuous function $g(x, y)$ defined for all x, y, we obtain a random variable $Z = g(X, Y)$. For example, if we roll two dice and X and Y are the numbers the dice turn up in a trial, then $Z = X + Y$ is the sum of those two numbers (see Fig. 514 in Sec. 24.5).

In the case of a ***discrete*** random variable (X, Y) we may obtain the probability function $f(z)$ of $Z = g(X, Y)$ by summing all $f(x, y)$ for which $g(x, y)$ equals the value of z considered; thus

$$f(z) = P(Z = z) = \sum_{g(x,y)=z}\sum f(x, y). \tag{20}$$

Hence the distribution function of Z is

$$F(z) = P(Z \leqq z) = \sum_{g(x,y)\leqq z}\sum f(x, y) \tag{21}$$

where we sum all values of $f(x, y)$ for which $g(x, y) \leqq z$.

In the case of a ***continuous*** random variable (X, Y) we similarly have

$$F(z) = P(Z \leqq z) = \iint_{g(x,y)\leqq z} f(x, y)\, dx\, dy \tag{22}$$

where for each z we integrate the density $f(x, y)$ of (X, Y) over the region $g(x, y) \leqq z$ in the xy-plane, the boundary curve of this region being $g(x, y) = z$.

Addition of Means

The number

$$(23) \qquad E(g(X, Y)) = \begin{cases} \displaystyle\sum_x \sum_y g(x, y) f(x, y) & [(X, Y) \text{ discrete}] \\[2ex] \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & [(X, Y) \text{ continuous}] \end{cases}$$

is called the *mathematical expectation* or, briefly, the **expectation of** $g(X, Y)$. Here it is assumed that the double series converges absolutely and the integral of $|g(x, y)| f(x, y)$ over the xy-plane exists (is finite). Since summation and integration are linear processes, we have from (23)

$$(24) \qquad E(ag(X, Y) + bh(X, Y)) = aE(g(X, Y)) + bE(h(X, Y)).$$

An important special case is

$$E(X + Y) = E(X) + E(Y),$$

and by induction we have the following result.

THEOREM 1

Addition of Means

The mean (expectation) of a sum of random variables equals the sum of the means (expectations), that is,

$$\textbf{(25)} \qquad E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

Furthermore, we readily obtain

THEROEM 2

Multiplication of Means

The mean (expectation) of the product of ***independent*** *random variables equals the product of the means (expectations), that is,*

$$(26) \qquad E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

PROOF If X and Y are independent random variables (both discrete or both continuous), then $E(XY) = E(X)E(Y)$. In fact, in the discrete case we have

$$E(XY) = \sum_x \sum_y xyf(x, y) = \sum_x xf_1(x) \sum_y yf_2(y) = E(X)E(Y),$$

and in the continuous case the proof of the relation is similar. Extension to n independent random variables gives (26), and Theorem 2 is proved. ■

Addition of Variances

This is another matter of practical importance that we shall need. As before, let $Z = X + Y$ and denote the mean and variance of Z by μ and σ^2. Then we first have (see Team Project 20(a) in Problem Set 24.6)

$$\sigma^2 = E([Z - \mu]^2) = E(Z^2) - [E(Z)]^2.$$

From (24) we see that the first term on the right equals

$$E(Z^2) = E(X^2 + 2XY + Y^2) = E(X^2) + 2E(XY) + E(Y^2).$$

For the second term on the right we obtain from Theorem 1

$$[E(Z)]^2 = [E(X) + E(Y)]^2 = [E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2.$$

By substituting these expressions into the formula for σ^2 we have

$$\begin{aligned}\sigma^2 = E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 \\ + 2[E(XY) - E(X)E(Y)].\end{aligned}$$

From Team Project 20, Sec. 24.6, we see that the expression in the first line on the right is the sum of the variances of X and Y, which we denote by σ_1^2 and σ_2^2, respectively. The quantity in the second line (except for the factor 2) is

$$\sigma_{XY} = E(XY) - E(X)E(Y) \tag{27}$$

and is called the **covariance** of X and Y. Consequently, our result is

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{XY}. \tag{28}$$

If X and Y are independent, then

$$E(XY) = E(X)E(Y);$$

hence $\sigma_{XY} = 0$, and

$$\sigma^2 = \sigma_1^2 + \sigma_2^2. \tag{29}$$

Extension to more than two variables gives the basic

THEOREM 3

Addition of Variances

The variance of the sum of ***independent*** *random variables equals the sum of the variances of these variables.*

CAUTION! In the numerous applications of Theorems 1 and 3 we must always remember that Theorem 3 holds only for ***independent*** variables.

This is the end of Chap. 24 on probability theory. Most of the concepts, methods, and special distributions discussed in this chapter will play a fundamental role in the next chapter, which deals with methods of **statistical inference**, that is, conclusions from samples to populations, whose unknown properties we want to know and try to discover by looking at suitable properties of samples that we have obtained.

PROBLEM SET 24.9

1. Let $f(x, y) = k$ when $8 \leqq x \leqq 12$ and $0 \leqq y \leqq 2$ and zero elsewhere. Find k. Find $P(X \leqq 11, 1 \leqq Y \leqq 1.5)$ and $P(9 \leqq X \leqq 13, Y \leqq 1)$.

2. Find $P(X > 4, Y > 4)$ and $P(X \leqq 1, Y \leqq 1)$ if (X, Y) has the density $f(x, y) = \frac{1}{32}$ if $x \geqq 0, y \geqq 0, x + y \leqq 8$.

3. Let $f(x, y) = k$ if $x > 0, y > 0, x + y < 3$ and 0 otherwise. Find k. Sketch $f(x, y)$. Find $P(X + Y \leqq 1)$, $P(Y > X)$.

4. Find the density of the marginal distribution of X in Prob. 2.

5. Find the density of the marginal distribution of Y in Fig. 524.

6. If certain sheets of wrapping paper have a mean weight of 10 g each, with a standard deviation of 0.05 g, what are the mean weight and standard deviation of a pack of 10,000 sheets?

7. What are the mean thickness and the standard deviation of transformer cores each consisting of 50 layers of sheet metal and 49 insulating paper layers if the metal sheets have mean thickness 0.5 mm each with a standard deviation of 0.05 mm and the paper layers have mean 0.05 mm each with a standard deviation of 0.02 mm?

8. Let X [cm] and Y [cm] be the diameters of a pin and hole, respectively. Suppose that (X, Y) has the density

$$f(x, y) = 625 \quad \text{if} \quad 0.98 < x < 1.02, \quad 1.00 < y < 1.04$$

and 0 otherwise. **(a)** Find the marginal distributions. **(b)** What is the probability that a pin chosen at random will fit a hole whose diameter is 1.00?

9. Using Theorems 1 and 3, obtain the formulas for the mean and the variance of the binomial distribution.

10. Using Theorem 1, obtain the formula for the mean of the hypergeometric distribution. Can you use Theorem 3 to obtain the variance of that distribution?

11. A 5-gear assembly is put together with spacers between the gears. The mean thickness of the gears is 5.020 cm with a standard deviation of 0.003 cm. The mean thickness of the spacers is 0.040 cm with a standard deviation of 0.002 cm. Find the mean and standard deviation of the assembled units consisting of 5 randomly selected gears and 4 randomly selected spacers.

12. If the mean weight of certain (empty) containers is 5 lb the standard deviation is 0.2 lb, and if the filling of the containers has mean weight 100 lb and standard deviation 0.5 lb, what are the mean weight and the standard deviation of filled containers?

13. Find $P(X > Y)$ when (X, Y) has the density

$$f(x, y) = 0.25e^{-0.5(x+y)} \quad \text{if} \quad x \geqq 0, y \geqq 0$$

and 0 otherwise.

14. An electronic device consists of two components. Let X and Y [years] be the times to failure of the first and second components, respectively. Assume that (X, Y) has the density $f(x, y) = 4e^{-2(x+y)}$ if $x > 0$ and $y > 0$ and 0 otherwise. **(a)** Are X and Y dependent or independent? **(b)** Find the densities of the marginal distributions. **(c)** What is the probability that the first component will have a lifetime of 2 years or longer?

15. Give an example of two different discrete distributions that have the same marginal distributions.

16. Prove (2).

17. Let (X, Y) have the probability function

$$f(0, 0) = f(1, 1) = \tfrac{1}{8},$$
$$f(0, 1) = f(1, 0) = \tfrac{3}{8}.$$

Are X and Y independent?

18. Let (X, Y) have the density

$$f(x, y) = k \text{ if } x^2 + y^2 < 1$$

and 0 otherwise. Determine k. Find the densities of the marginal distributions. Find the probability

$$P(X^2 + Y^2 < \tfrac{1}{4}).$$

19. Show that the random variables with the densities

$$f(x, y) = x + y$$

and

$$g(x, y) = (x + \tfrac{1}{2})(y + \tfrac{1}{2})$$

if $0 \leqq x \leqq 1, 0 \leqq y \leqq 1$ and $f(x, y) = 0$ and $g(x, y) = 0$ elsewhere, have the same marginal distribution.

20. Prove the statement involving (18).

CHAPTER 24 REVIEW QUESTIONS AND PROBLEMS

1. What are stem-and-leaf plots? Boxplots? Histograms? Compare their advantages.
2. What properties of data are measured by the mean? The median? The standard deviation? The variance?
3. What do we mean by an experiment? An outcome? An event? Give examples.
4. What is a random variable? Its distribution function? Its probability function or density?
5. State the definition of probability from memory. Give simple examples.
6. What is sampling with and without replacement? What distributions are involved?
7. When is the Poisson distribution a good approximation of the binomial distribution? The normal distribution?
8. Explain the use of the tables of the normal distribution. If you have a CAS, how would you proceed without the tables?
9. State the main theorems on probability. Illustrate them by simple examples.
10. State the most important facts about distributions of two random variables and their marginal distributions.
11. Make a stem-and-leaf plot, histogram, and boxplot of the data 110, 113, 109, 118, 110, 115, 104, 111, 116, 113.
12. Same task as in Prob. 11. for the data 13.5, 13.2, 12.1, 13.6, 13.3.
13. Find the mean, standard deviation, and variance in Prob. 11.
14. Find the mean, standard deviation, and variance in Prob. 12.
15. Show that the mean always lies between the smallest and the largest data value.
16. What are the outcomes in the sample space of the experiment of simultaneously tossing three coins?
17. Plot a histogram of the data 8, 2, 4, 10 and guess $\bar{x}$ and s by inspecting the histogram. Then calculate $\bar{x}$, s^2, and s.
18. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.
19. Suppose that 3% of bolts made by a machine are defective, the defectives occurring at random during production. If the bolts are packaged 50 per box, what is the binomial approximation of the probability that a given box will contain $x = 0, 1, \cdots, 5$ defectives?
20. Of a lot of 12 items, 3 are defective. **(a)** Find the number of different samples of 3 items. Find the number of samples of 3 items containing **(b)** no defectives, **(c)** 1 defective, **(d)** 2 defectives, **(e)** 3 defectives.
21. Find the probability function of X = *Number of times of tossing a fair coin until the first head appears.*
22. If the life of ball bearings has the density $f(x) = ke^{-x}$ if $0 \leqq x \leqq 2$ and 0 otherwise, what is k? What is the probability $P(X \geqq 1)$?
23. Find the mean and variance of a discrete random variable X having the probability function $f(0) = \frac{1}{4}$, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$.
24. Let X be normal with mean 14 and variance 4. Determine c such that $P(X \leqq c) = 95\%$, $P(X \leqq c) = 5\%$, $P(X \leqq c) = 99.5\%$.
25. Let X be normal with mean 80 and variance 9. Find $P(X > 83)$, $P(X < 81)$, $P(X < 80)$, and $P(78 < X < 82)$.

SUMMARY OF CHAPTER 24
Data Analysis. Probability Theory

A *random experiment,* briefly called **experiment**, is a process in which the result ("**outcome**") depends on "chance" (effects of factors unknown to us). Examples are games of chance with dice or cards, measuring the hardness of steel, observing weather conditions, or recording the number of accidents in a city. (Thus the word "experiment" is used here in a much wider sense than in common language.) The outcomes are regarded as points (elements) of a set S, called the **sample space**, whose subsets are called **events**. For events E we define a **probability** $P(E)$ by the axioms (Sec. 24.3)

$$
\begin{aligned}
&0 \leqq P(E) \leqq 1 \\
(1)\qquad &P(S) = 1 \\
&P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots \qquad (E_j \cap E_k = \varnothing).
\end{aligned}
$$

These axioms are motivated by properties of frequency distributions of data (Sec. 24.1).

The complement E^c of E has the probability

$$P(E^c) = 1 - P(E). \tag{2}$$

The **conditional probability** of an event B under the condition that an event A happens is (Sec. 24.3)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) > 0]. \tag{3}$$

Two events A and B are called **independent** if the probability of their simultaneous appearance in a trial equals the product of their probabilities, that is, if

$$P(A \cap B) = P(A)P(B). \tag{4}$$

With an experiment we associate a **random variable** X. This is a function defined on S whose values are real numbers; furthermore, X is such that the probability $P(X = a)$ with which X assumes any value a, and the probability $P(a < X \leqq b)$ with which X assumes any value in an interval $a < X \leqq b$ are defined (Sec. 24.5). The **probability distribution** of X is determined by the distribution function

$$F(x) = P(X \leqq x). \tag{5}$$

In applications there are two important kinds of random variables: those of the **discrete** type, which appear if we count (defective items, customers in a bank, etc.) and those of the **continuous** type, which appear if we measure (length, speed, temperature, weight, etc.).

A discrete random variable has a **probability function**

$$f(x) = P(X = x). \tag{6}$$

Its **mean** μ and **variance** σ^2 are (Sec. 24.6)

$$\mu = \sum_j x_j f(x_j) \qquad \text{and} \qquad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \tag{7}$$

where the x_j are the values for which X has a positive probability. Important discrete random variables and distributions are the **binomial, Poisson,** and **hypergeometric distributions** discussed in Sec. 24.7.

A continuous random variable has a **density**

$$f(x) = F'(x) \qquad [\text{see } (5)]. \tag{8}$$

Its mean and variance are (Sec. 24.6)

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx. \tag{9}$$

Very important is the **normal distribution** (Sec. 24.8), whose density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{10}$$

and whose distribution function is (Sec. 24.8; Tables A7, A8 in App. 5)

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{11}$$

A **two-dimensional random variable** (X, Y) occurs if we simultaneously observe two quantities (for example, height X and weight Y of adults). Its distribution function is (Sec. 24.9)

$$F(x, y) = P(X \leqq x, Y \leqq y). \tag{12}$$

X and Y have the distribution functions (Sec. 24.9)

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) \qquad \text{and} \qquad F_2(y) = P(x \text{ arbitrary}, Y \leqq y) \tag{13}$$

respectively; their distributions are called **marginal distributions**. If both X and Y are discrete, then (X, Y) has a probability function

$$f(x, y) = P(X = x, Y = y).$$

If both X and Y are continuous, then (X, Y) has a density $f(x, y)$.

CHAPTER 25

Mathematical Statistics

In probability theory we set up mathematical models of processes that are affected by "chance." In mathematical statistics or, briefly, **statistics**, we check these models against the observable reality. This is called **statistical inference**. It is done by **sampling**, that is, by drawing random samples, briefly called **samples**. These are sets of values from a much larger set of values that could be studied, called the **population**. An example is 10 diameters of screws drawn from a large lot of screws. Sampling is done in order to see whether a model of the population is accurate enough for practical purposes. If this is the case, the model can be used for predictions, decisions, and actions, for instance, in planning productions, buying equipment, investing in business projects, and so on.

Most important methods of statistical inference are **estimation of parameters** (Secs. 25.2), determination of **confidence intervals** (Sec. 25.3), and **hypothesis testing** (Sec. 25.4, 25.7, 25.8), with application to *quality control* (Sec. 25.5) and *acceptance sampling* (Sec. 25.6).

In the last section (25.9) we give an introduction to **regression** and **correlation analysis**, which concern experiments involving two variables.

Prerequisite: Chap. 24.
Sections that may be omitted in a shorter course: 25.5, 25.6, 25.8.
References, Answers to Problems, and Statistical Tables: App. 1 Part G, App. 2, App. 5.

25.1 Introduction. Random Sampling

Mathematical statistics consists of methods for designing and evaluating random experiments to obtain information about practical problems, such as exploring the relation between iron content and density of iron ore, the quality of raw material or manufactured products, the efficiency of air-conditioning systems, the performance of certain cars, the effect of advertising, the reactions of consumers to a new product, etc.

Random variables occur more frequently in engineering (and elsewhere) than one would think. For example, properties of mass-produced articles (screws, lightbulbs, etc.) always show **random variation**, due to small (uncontrollable!) differences in raw material or manufacturing processes. Thus the diameter of screws is a random variable X and we have *nondefective screws,* with diameter between given tolerance limits, and *defective screws,* with diameter outside those limits. We can ask for the distribution of X, for the percentage of defective screws to be expected, and for necessary improvements of the production process.

Samples are selected from populations—20 screws from a lot of 1000, 100 of 5000 voters, 8 beavers in a wildlife conservation project—because inspecting the entire population would be too expensive, time-consuming, impossible or even senseless (think of destructive testing of lightbulbs or dynamite). To obtain meaningful conclusions, samples must be **random selections**. Each of the 1000 screws must have the same chance of being sampled (of being drawn when we sample), at least approximately. Only then will the sample mean $\bar{x} = (x_1 + \cdots + x_{20})/20$ (Sec. 24.1) of a sample of size $n = 20$ (or any other n) be a good approximation of the population mean μ (Sec. 24.6); and the accuracy of the approximation will generally improve with increasing n, as we shall see. Similarly for other parameters (standard deviation, variance, etc.).

Independent sample values will be obtained in experiments with an infinite sample space S (Sec. 24.2), certainly for the normal distribution. This is also true in sampling with replacement. It is approximately true in drawing *small* samples from a large finite population (for instance, 5 or 10 of 1000 items). However, if we sample without replacement from a small population, the effect of dependence of sample values may be considerable.

Random numbers help in obtaining samples that are in fact random selections. This is sometimes not easy to accomplish because there are many subtle factors that can bias sampling (by personal interviews, by poorly working machines, by the choice of nontypical observation conditions, etc.). Random numbers can be obtained from a **random number generator** in Maple, Mathematica, or other systems listed on p. 789. (The numbers are not truly random, as they would be produced in flipping coins or rolling dice, but are calculated by a tricky formula that produces numbers that do have practically all the essential features of true randomness. Because these numbers eventually repeat, they must not be used in cryptography, for example, where true randomness is required.)

EXAMPLE 1 **Random Numbers from a Random Number Generator**

To select a sample of size $n = 10$ from 80 given ball bearings, we number the bearings from 1 to 80. We then let the generator randomly produce 10 of the integers from 1 to 80 and include the bearings with the numbers obtained in our sample, for example.

44 55 53 03 52 61 67 78 39 54

or whatever.

Random numbers are also contained in (older) statistical tables. ■

Representing and processing data were considered in Sec. 24.1 in connection with frequency distributions. These are the empirical counterparts of probability distributions and helped motivating axioms and properties in probability theory. The new aspect in this chapter is **randomness**: the data are samples selected **randomly** from a population. Accordingly, we can immediately make the connection to Sec. 24.1, using stem-and-leaf plots, box plots, and histograms for representing samples graphically.

Also, we now call the mean $\bar{x}$ in (5), Sec. 24.1, the **sample mean**

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n). \tag{1}$$

We call n the **sample size**, the variance s^2 in (6), Sec. 24.1, the **sample variance**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2], \tag{2}$$

and its positive square root s the **sample standard deviation**. $\bar{x}$, s^2, and s are called **parameters** *of a sample;* they will be needed throughout this chapter.

25.2 Point Estimation of Parameters

Beginning in this section, we shall discuss the most basic practical tasks in statistics and corresponding statistical methods to accomplish them. The first of them is point estimation of **parameters**, that is, of quantities appearing in distributions, such as p in the binomial distribution and μ and σ in the normal distribution.

A **point estimate** of a parameter is a number (point on the real line), which is computed from a given sample and serves as an approximation of the unknown exact value of the parameter of the population. An **interval estimate** is an interval (*"confidence interval"*) obtained from a sample; such estimates will be considered in the next section. Estimation of parameters is of great practical importance in many applications.

As an approximation of the mean μ of a population we may take the mean $\bar{x}$ of a corresponding sample. This gives the estimate $\hat{\mu} = \bar{x}$ for μ, that is,

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \tag{1}$$

where n is the sample size. Similarly, an estimate $\hat{\sigma}^2$ for the variance of a population is the variance s^2 of a corresponding sample, that is,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2. \tag{2}$$

Clearly, (1) and (2) are estimates of parameters for distributions in which μ or σ^2 appear explicity as parameters, such as the normal and Poisson distributions. For the binomial distribution, $p = \mu/n$ [see (3) in Sec. 24.7]. From (1) we thus obtain for p the estimate

$$\hat{p} = \frac{\bar{x}}{n}. \tag{3}$$

We mention that (1) is a special case of the so-called **method of moments**. In this method the parameters to be estimated are expressed in terms of the moments of the distribution (see Sec. 24.6). In the resulting formulas, those moments of the distribution are replaced by the corresponding moments of the sample. This gives the estimates. Here the ***k*th moment of a sample** $x_1, \cdots, x_n$ is

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k.$$

Maximum Likelihood Method

Another method for obtaining estimates is the so-called **maximum likelihood method** of R. A. Fisher [*Messenger Math.* **41** (1912), 155–160]. To explain it, we consider a discrete (or continuous) random variable X whose probability function (or density) $f(x)$ depends on a single parameter θ. We take a corresponding sample of n ***independent*** values $x_1, \cdots, x_n$. Then in the discrete case the probability that a sample of size n consists precisely of those n values is

$$l = f(x_1)f(x_2) \cdots f(x_n). \tag{4}$$

In the continuous case the probability that the sample consists of values in the small intervals $x_j \leqq x \leqq x_j + \Delta x \ (j = 1, 2, \cdots, n)$ is

$$f(x_1)\Delta x\, f(x_2)\Delta x \cdots f(x_n)\Delta x = l(\Delta x)^n. \tag{5}$$

Since $f(x_j)$ depends on θ, the function l in (5) given by (4) depends on $x_1, \cdots, x_n$ and θ. We imagine $x_1, \cdots, x_n$ to be given and fixed. Then l is a function of θ, which is called the **likelihood function**. The basic idea of the maximum likelihood method is quite simple, as follows. We choose *that* approximation for the unknown value of θ for which l is as large as possible. If l is a differentiable function of θ, a necessary condition for l to have a maximum in an interval (not at the boundary) is

$$\frac{\partial l}{\partial \theta} = 0. \tag{6}$$

(We write a *partial* derivative, because l depends also on $x_1, \cdots, x_n$.) A solution of (6) depending on $x_1, \cdots, x_n$ is called a **maximum likelihood estimate** for θ. We may replace (6) by

$$\frac{\partial \ln l}{\partial \theta} = 0, \tag{7}$$

because $f(x_j) > 0$, a maximum of l is in general positive, and $\ln l$ is a monotone increasing function of l. This often simplifies calculations.

Several Parameters. If the distribution of X involves r parameters $\theta_1, \cdots, \theta_r$, then instead of (6) we have the r conditions $\partial l/\partial\theta_1 = 0, \cdots, \partial l/\partial\theta_r = 0$, and instead of (7) we have

$$\frac{\partial \ln l}{\partial \theta_1} = 0, \qquad \cdots, \qquad \frac{\partial \ln l}{\partial \theta_r} = 0. \tag{8}$$

EXAMPLE 1 Normal Distribution

Find maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma$ in the case of the normal distribution.

Solution. From (1), Sec. 24.8, and (4) we obtain the likelihood function

$$l = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-h} \qquad \text{where} \qquad h = \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2.$$

Taking logarithms, we have

$$\ln l = -n \ln \sqrt{2\pi} - n \ln \sigma - h.$$

The first equation in (8) is $\partial(\ln l)/\partial\mu = 0$, written out

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial h}{\partial \mu} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j - \mu) = 0, \qquad \text{hence} \qquad \sum_{j=1}^{n} x_j - n\mu = 0.$$

The solution is the desired estimate $\hat{\mu}$ for μ: we find

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j = \bar{x}.$$

The second equation in (8) is $\partial(\ln l)/\partial\sigma = 0$, written out

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{\partial h}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j - \mu)^2 = 0.$$

Replacing μ by $\hat{\mu}$ and solving for σ^2, we obtain the estimate

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

which we shall use in Sec. 25.7. Note that this differs from (2). We cannot discuss criteria for the goodness of estimates but want to mention that for small n, formula (2) is preferable. ■

PROBLEM SET 25.2

1. **Normal distribution.** Apply the maximum likelihood method to the normal distribution with $\mu = 0$.
2. Find the maximum likelihood estimate for the parameter μ of a normal distribution with known variance $\sigma^2 = \sigma_0^2 = 16$.
3. **Poisson distribution.** Derive the maximum likelihood estimator for μ. Apply it to the sample (10, 25, 26, 17, 10, 4), giving numbers of minutes with 0–10, 11–20, 21–30, 31–40, 41–50, more than 50 fliers per minute, respectively, checking in at some airport check-in.
4. **Uniform distribution.** Show that, in the case of the parameters a and b of the uniform distribution (see Sec. 24.6), the maximum likelihood estimate cannot be obtained by equating the first derivative to zero. How can we obtain maximum likelihood estimates in this case, more or less by using common sense?
5. **Binomial distribution.** Derive a maximum likelihood estimate for p.
6. Extend Prob. 5 as follows. Suppose that m times n trials were made and in the first n trials A happened k_1 times, in the second n trials A happened k_2 times, $\cdots$, in the mth n trials A happened k_m times. Find a maximum likelihood estimate of p based on this information.
7. Suppose that in Prob. 6 we made 3 times 4 trials and A happened 2, 3, 2 times, respectively. Estimate p.
8. **Geometric distribution.** Let X = *Number of independent trials until an event A occurs*. Show that X has a geometric distribution, defined by the probability function $f(x) = pq^{x-1}$, $x = 1, 2, \cdots$, where p is the probability of A in a single trial and $q = 1 - p$. Find the maximum likelihood estimate of p corresponding to a sample $x_1, x_2, \cdots, x_n$ of observed values of X.
9. In Prob. 8, show that $f(1) + f(2) + \cdots = 1$ (as it should be!). Calculate independently of Prob. 8 the maximum likelihood of p in Prob. 8 corresponding to a single observed value of X.
10. In rolling a die, suppose that we get the first "*Six*" in the 7th trial and in doing it again we get it in the 6th trial. Estimate the probability p of getting a "*Six*" in rolling that die once.
11. Find the maximum likelihood estimate of θ in the density $f(x) = \theta e^{-\theta x}$ if $x \geqq 0$ and $f(x) = 0$ if $x < 0$.
12. In Prob. 11, find the mean μ, substitute it in $f(x)$, find the maximum likelihood estimate of μ, and show that it is identical with the estimate for μ which can be obtained from that for θ in Prob. 11.

13. Compute $\hat{\theta}$ in Prob. 11 from the sample 1.9, 0.4, 0.7, 0.6, 1.4. Graph the sample distribution function $\hat{F}(x)$ and the distribution function $F(x)$ of the random variable, with $\theta = \hat{\theta}$, on the same axes. Do they agree reasonably well? (We consider goodness of fit systematically in Sec. 25.7.)

14. Do the same task as in Prob. 13 if the given sample is 0.4, 0.7, 0.2, 1.1, 0.1.

15. CAS EXPERIMENT. Maximum Likelihood Estimates. (MLEs). Find experimentally how much MLEs can differ depending on the sample size. *Hint.* Generate many samples of the same size n, e.g., of the standardized normal distribution, and record $\bar{x}$ and s^2. Then increase n.

25.3 Confidence Intervals

Confidence intervals[1] for an unknown parameter θ of some distribution (e.g., $\theta = \mu$) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ that contain θ, not with certainty but with a high probability γ, which we can choose (95% and 99% are popular). Such an interval is calculated from a sample. $\gamma = 95\%$ means probability $1 - \gamma = 5\% = \frac{1}{20}$ of being wrong—one of about 20 such intervals will not contain θ. Instead of writing $\theta_1 \leqq \theta \leqq \theta_2$, we denote this more distinctly by writing

$$\mathrm{CONF}_{\gamma}\,\{\theta_1 \leqq \theta \leqq \theta_2\}. \tag{1}$$

Such a special symbol, CONF, seems worthwhile in order to avoid the misunderstanding that θ *must* lie between θ_1 and θ_2.

γ is called the **confidence level**, and θ_1 and θ_2 are called the **lower** and **upper confidence limits**. They depend on γ. The larger we choose γ, the smaller is the error probability $1 - \gamma$, but the longer is the confidence interval. If $\gamma \to 1$, then its length goes to infinity. The choice of γ depends on the kind of application. In taking no umbrella, a 5% chance of getting wet is not tragic. In a medical decision of life or death, a 5% chance of being wrong may be too large and a 1% chance of being wrong ($\gamma = 99\%$) may be more desirable.

Confidence intervals are more valuable than point estimates (Sec. 25.2). Indeed, we can take the midpoint of (1) as an approximation of θ and half the length of (1) as an "error bound" (not in the strict sense of numerics, but except for an error whose probability we know).

θ_1 and θ_2 in (1) are calculated from a sample $x_1, \cdots, x_n$. These are n observations of a random variable X. Now comes a **standard trick**. *We regard $x_1, \cdots, x_n$ as* ***single*** *observations of n random variables $X_1, \cdots, X_n$ (with the same distribution, namely, that of X).* Then $\theta_1 = \theta_1(x_1, \cdots, x_n)$ and $\theta_2 = \theta_2(x_1, \cdots, x_n)$ in (1) are observed values of two random variables $\Theta_1 = \Theta_1(X_1, \cdots, X_n)$ and $\Theta_2 = \Theta_2(X_1, \cdots, X_n)$. The condition (1) involving γ can now be written

$$P(\Theta_1 \leqq \theta \leqq \Theta_2) = \gamma. \tag{2}$$

Let us see what all this means in concrete practical cases.

In each case in this section we shall first state the steps of obtaining a confidence interval in the form of a table, then consider a typical example, and finally justify those steps theoretically.

---

[1]JERZY NEYMAN (1894–1981), American statistician, developed the theory of confidence intervals (*Annals of Mathematical Statistics* **6** (1935), 111–116).

Confidence Interval for μ of the Normal Distribution with Known σ^2

Table 25.1 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Known Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the corresponding c:

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

Step 3. Compute the mean $\bar{x}$ of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\text{CONF}_\gamma\,\{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{3}$$

EXAMPLE 1 Confidence Interval for μ of the Normal Distribution with Known σ^2

Determine a 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of $n = 100$ values with mean $\bar{x} = 5$.

Solution. ***Step 1.*** $\gamma = 0.95$ is required. ***Step 2.*** The corresponding c equals 1.960; see Table 25.1. ***Step 3.*** $\bar{x} = 5$ is given. ***Step 4.*** We need $k = 1.960 \cdot 3/\sqrt{100} = 0.588$. Hence $\bar{x} - k = 4.412$, $\bar{x} + k = 5.588$ and the confidence interval is $\text{CONF}_{0.95}\,\{4.412 \leqq \mu \leqq 5.588\}$.

This is sometimes written $\mu = 5 \pm 0.588$, but we shall not use this notation, which can be misleading. With your CAS you can determine this interval more directly. Similarly for the other examples in this section. ■

Theory for Table 25.1. The method in Table 25.1 follows from the basic

THEOREM 1 Sum of Independent Normal Random Variables

*Let $X_1, \cdots, X_n$ be **independent** normal random variables each of which has mean μ and variance σ^2. Then the following holds.*

(a) *The sum $X_1 + \cdots + X_n$ is normal with mean $n\mu$ and variance $n\sigma^2$.*

(b) The following random variable $\bar{X}$ *is normal with mean μ and variance σ^2/n.*

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \tag{4}$$

(c) *The following random variable Z is normal with mean 0 and variance 1.*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{5}$$

PROOF The statements about the mean and variance in (a) follow from Theorems 1 and 3 in Sec. 24.9. From this, and Theorem 2 in Sec. 24.6, we see that $\bar{X}$ has the mean $(1/n)n\mu = \mu$ and the variance $(1/n)^2 n\sigma^2 = \sigma^2/n$. This implies that Z has the mean 0 and variance 1, by Theorem 2(b) in Sec. 24.6. The normality of $X_1 + \cdots + X_n$ is proved in Ref. [G3] listed in App. 1. This implies the normality of (4) and (5). ■

Derivation of (3) in Table 25.1. Sampling from a normal distribution gives independent sample values (see Sec. 25.1), so that Theorem 1 applies. Hence we can choose γ and then determine c such that

$$P(-c \leqq Z \leqq c) = P\left(-c \leqq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqq c\right) = \Phi(c) - \Phi(-c) = \gamma. \tag{6}$$

For the value $\gamma = 0.95$ we obtain $z(D) = 1.960$ from Table A8 in App. 5, as used in Example 1. For $\gamma = 0.9, 0.99, 0.999$ we get the other values of c listed in Table 25.1. Finally, all we have to do is to convert the inequality in (6) into one for μ and insert observed values obtained from the sample. We multiply $-c \leqq Z \leqq c$ by -1 and then by $\sigma/\sqrt{n}$, writing $c\sigma/\sqrt{n} = k$ (as in Table 25.1),

$$\begin{aligned} P(-c \leqq Z \leqq c) = P(c \geqq -Z \geqq -c) &= P\left(c \geqq \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \geqq -c\right) \\ &= P(k \geqq \mu - \bar{X} \geqq -k) = \gamma. \end{aligned}$$

Adding $\bar{X}$ gives $P(\bar{X} + k \geqq \mu \geqq \bar{X} - k) = \gamma$ or

$$P(\bar{X} - k \leqq \mu \leqq \bar{X} + k) = \gamma. \tag{7}$$

Inserting the observed value $\bar{x}$ of $\bar{X}$ gives (3). Here we have regarded $x_1, \cdots, x_n$ as single observations of $X_1, \cdots, X_n$ (the standard trick!), so that $x_1 + \cdots + x_n$ is an observed value of $X_1 + \cdots + X_n$ and $\bar{x}$ is an observed value of $\bar{X}$. Note further that (7) is of the form (2) with $\Theta_1 = \bar{X} - k$ and $\Theta_2 = \bar{X} + k$. ■

EXAMPLE 2 Sample Size Needed for a Confidence Interval of Prescribed Length

How large must n be in Example 1 if we want to obtain a 95% confidence interval of length $L = 0.4$?

Solution. The interval (3) has the length $L = 2k = 2c\sigma/\sqrt{n}$. Solving for n, we obtain

$$n = (2c\sigma/L)^2.$$

In the present case the answer is $n = (2 \cdot 1.960 \cdot 3/0.4)^2 \approx 870$.

Figure 526 shows how L decreases as n increases and that for $\gamma = 99\%$ the confidence interval is substantially longer than for $\gamma = 95\%$ (and the same sample size n). ■

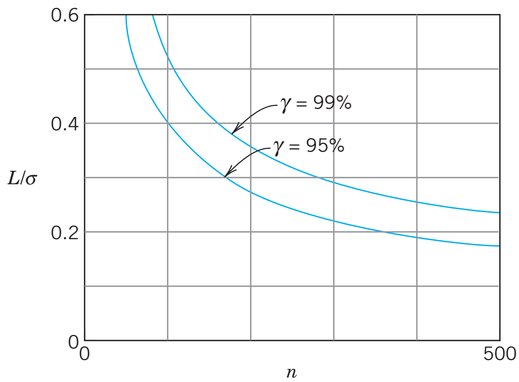


Fig. 526. Length of the confidence interval (3) (measured in multiples of σ) as a function of the sample size n for $\gamma = 95\%$ and $\gamma = 99\%$

Confidence Interval for μ of the Normal Distribution with Unknown σ^2

In practice σ^2 is frequently unknown. Then the method in Table 25.1 does not help and the whole theory changes, although the steps of determining a confidence interval for μ remain quite similar. They are shown in Table 25.2. We see that k differs from that in Table 25.1, namely, the sample standard deviation s has taken the place of the unknown standard deviation σ of the population. And c now depends on the sample size n and must be determined from Table A9 in App. 5 or from your CAS. That table lists values z for given values of the distribution function (Fig. 527)

$$F(z) = K_m \int_{-\infty}^{x} \left(1 + \frac{u^2}{m}\right)^{-(m+1)/2} du \tag{8}$$

of the ***t*-distribution**. Here, $m\ (= 1, 2, \cdots)$ is a parameter, called the **number of degrees of freedom** of the distribution (***abbreviated* d.f.**). In the present case, $m = n - 1$; see Table 25.2. The constant K_m is such that $F(\infty) = 1$. By integration it turns out that $K_m = \Gamma(\frac{1}{2}m + \frac{1}{2})/[\sqrt{m\pi}\,\Gamma(\frac{1}{2}m)]$, where Γ is the gamma function (see (24) in App. A3.1).

Table 25.2 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Unknown Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{9}$$

from the table of the t-distribution with $n - 1$ degrees of freedom (Table A9 in App. 5; or use a CAS; n = sample size).

Step 3. Compute the mean $\bar{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = cs/\sqrt{n}$. The confidence interval is

$$\mathrm{CONF}_\gamma\ \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{10}$$

Figure 528 compares the curve of the density of the t-distribution with that of the normal distribution. The latter is steeper. This illustrates that Table 25.1 (which uses more information, namely, the known value of σ^2) yields shorter confidence intervals than Table 25.2. This is confirmed in Fig. 529, which also gives an idea of the gain by increasing the sample size.

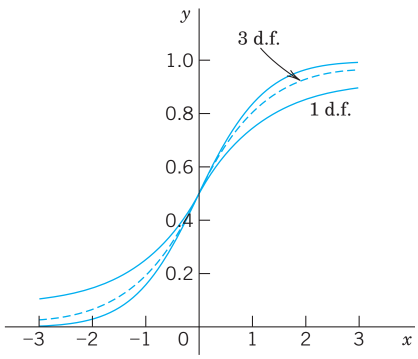


Fig. 527. Distribution functions of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution (steepest curve)

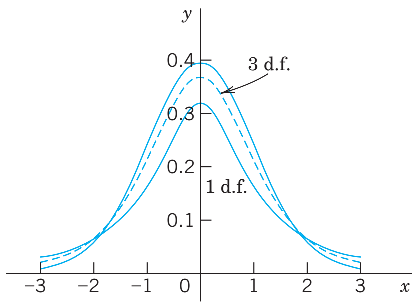


Fig. 528. Densities of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution

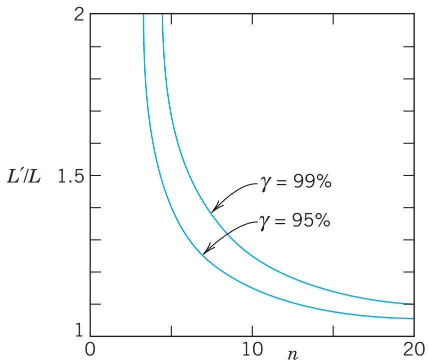


Fig. 529. Ratio of the lengths L' and L of the confidence intervals (10) and (3) with $\gamma = 95\%$ and $\gamma = 99\%$ as a function of the sample size n for equal s and σ

EXAMPLE 3 Confidence Interval for μ of the Normal Distribution with Unknown σ^2

Five independent measurements of the point of inflammation (flash point) of Diesel oil (D-2) gave the values (in °F) 144 147 146 142 144. Assuming normality, determine a 99% confidence interval for the mean.

Solution. ***Step 1.*** $\gamma = 0.99$ is required.

Step 2. $F(c) = \frac{1}{2}(1 + \gamma) = 0.995$, and Table A9 in App. 5 with $n - 1 = 4$ d.f. gives $c = 4.60$.

Step 3. $\bar{x} = 144.6, \quad s^2 = 3.8$.

Step 4. $k = \sqrt{3.8} \cdot 4.60/\sqrt{5} = 4.01$. The confidence interval is $\text{CONF}_{0.99}\{140.5 \leqq \mu \leqq 148.7\}$.

If the variance σ^2 were known and equal to the sample variance s^2, thus $\sigma^2 = 3.8$, then Table 25.1 would give $k = c\sigma/\sqrt{n} = 2.576\sqrt{3.8}/\sqrt{5} = 2.25$ and $\text{CONF}_{0.99}\{142.35 \leqq \mu \leqq 146.85\}$. We see that the present interval is almost twice as long as that obtained from Table 25.1 (with $\sigma^2 = 3.8$). Hence for small samples the difference is considerable! See also Fig. 529. ■

Theory for Table 25.2. For deriving (10) in Table 25.2 we need from Ref. [G3]

THEOREM 2

Student's *t*-Distribution

Let $X_1, \cdots, X_n$ *be independent normal random variables with the same mean* μ *and the same variance* σ^2*. Then the random variable*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{11}$$

has a t-distribution [see (8)] *with* $n - 1$ *degrees of freedom* (d.f.); *here* $\overline{X}$ *is given by* (4) *and*

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2. \tag{12}$$

Derivation of (10). This is similar to the derivation of (3). We choose a number γ between 0 and 1 and determine a number c from Table A9 in App. 5 with $n - 1$ d.f. (or from a CAS) such that

$$P(-c \leqq T \leqq c) = F(c) - F(-c) = \gamma. \tag{13}$$

Since the t-distribution is symmetric, we have

$$F(-c) = 1 - F(c),$$

and (13) assumes the form (9). Substituting (11) into (13) and transforming the result as before, we obtain

$$P(\overline{X} - K \leqq \mu \leqq \overline{X} + K) = \gamma \tag{14}$$

where

$$K = cS/\sqrt{n}.$$

By inserting the observed values $\bar{x}$ of $\overline{X}$ and s^2 of S^2 into (14) we finally obtain (10). ■

Confidence Interval for the Variance σ^2 of the Normal Distribution

Table 25.3 shows the steps, which are similar to those in Tables 25.1 and 25.2.

Table 25.3 Determination of a Confidence Interval for the Variance σ^2 of a Normal Distribution, Whose Mean Need Not Be Known

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine solutions c_1 and c_2 of the equations

$$F(c_1) = \tfrac{1}{2}(1 - \gamma), \qquad F(c_2) = \tfrac{1}{2}(1 + \gamma) \tag{15}$$

from the table of the chi-square distribution with $n - 1$ degrees of freedom (Table A10 in App. 5; or use a CAS; n = sample size).

Step 3. Compute $(n - 1)s^2$, where s^2 is the variance of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k_1 = (n - 1)s^2/c_1$ and $k_2 = (n - 1)s^2/c_2$. The confidence interval is

$$\text{CONF}_\gamma \{k_2 \leqq \sigma^2 \leqq k_1\}. \tag{16}$$

EXAMPLE 4 Confidence Interval for the Variance of the Normal Distribution

Determine a 95% confidence interval (16) for the variance, using Table 25.3 and a sample (tensile strength of sheet steel in kg/mm^2, rounded to integer values)

$$89 \quad 84 \quad 87 \quad 81 \quad 89 \quad 86 \quad 91 \quad 90 \quad 78 \quad 89 \quad 87 \quad 99 \quad 83 \quad 89.$$

Solution. ***Step 1.*** $\gamma = 0.95$ is required.

Step 2. For $n - 1 = 13$ we find

$$c_1 = 5.01 \quad \text{and} \quad c_2 = 24.74.$$

Step 3. $13s^2 = 326.9$.

Step 4. $13s^2/c_1 = 65.25$, $13s^2/c_2 = 13.21$.

The confidence interval is

$$\text{CONF}_{0.95} \{13.21 \leqq \sigma^2 \leqq 65.25\}.$$

This is rather large, and for obtaining a more precise result, one would need a much larger sample. ■

Theory for Table 25.3. In Table 25.1 we used the normal distribution, in Table 25.2 the t-distribution, and now we shall use the **χ^2-distribution** (*chi-square distribution*), whose distribution function is $F(z) = 0$ if $z < 0$ and

$$F(z) = C_m \int_0^z e^{-u/2} u^{(m-2)/2}\, du \qquad \text{if } z \geqq 0 \qquad \text{(Fig. 530).}$$

The parameter m $(= 1, 2, \cdots)$ is called the **number of degrees of freedom** (d.f.), and

$$C_m = 1/[2^{m/2}\Gamma(\tfrac{1}{2}m)].$$

Note that the distribution is not symmetric (see also Fig. 531).

For deriving (16) in Table 25.3 we need the following theorem.

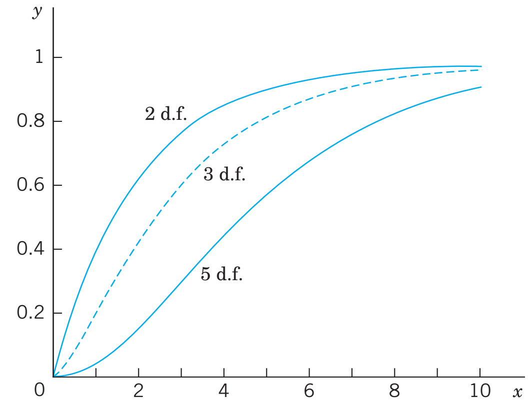


Fig. 530. Distribution function of the chi-square distribution with 2, 3, 5 d.f.

THEOREM 3

Chi-Square Distribution

Under the assumptions in Theorem 2 the random variable

(17) $$Y = (n - 1)\frac{S^2}{\sigma^2}$$

with S^2 *given by* (12) *has a chi-square distribution with* $n - 1$ *degrees of freedom.*

Proof in Ref. [G3], listed in App. 1.

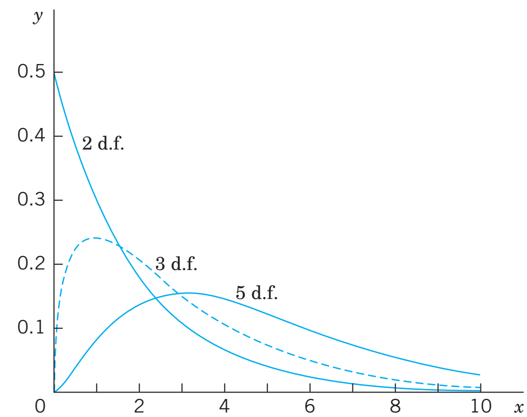


Fig. 531. Density of the chi-square distribution with 2, 3, 5 d.f.

Derivation of (16). This is similar to the derivation of (3) and (10). We choose a number γ between 0 and 1 and determine c_1 and c_2 from Table A10, App. 5, such that [see (15)]

$$P(Y \leqq c_1) = F(c_1) = \tfrac{1}{2}(1 - \gamma), \qquad P(Y \leqq c_2) = F(c_2) = \tfrac{1}{2}(1 + \gamma).$$

Subtraction yields

$$P(c_1 \leqq Y \leqq c_2) = P(Y \leqq c_2) - P(Y \leqq c_1) = F(c_2) - F(c_1) = \gamma.$$

Transforming $c_1 \leqq Y \leqq c_2$ with Y given by (17) into an inequality for σ^2, we obtain

$$\frac{n-1}{c_2} S^2 \leqq \sigma^2 \leqq \frac{n-1}{c_1} S^2.$$

By inserting the observed value s^2 of S^2 we obtain (16). ■

Confidence Intervals for Parameters of Other Distributions

The methods in Tables 25.1–25.3 for confidence intervals for μ and σ^2 are designed for the normal distribution. We now show that they can also be applied to other distributions if we use large samples.

We know that if $X_1, \cdots, X_n$ are independent random variables with the same mean μ and the same variance σ^2, then their sum $Y_n = X_1 + \cdots + X_n$ has the following properties.

(A) Y_n has the mean $n\mu$ and the variance $n\sigma^2$ (by Theorems 1 and 3 in Sec. 24.9).

(B) If those variables are normal, then Y_n is normal (by Theorem 1).

If those random variables are not normal, then **(B)** is not applicable. However, for large n the random variable Y_n is still *approximately* normal. This follows from the central limit theorem, which is one of the most fundamental results in probability theory.

THEOREM 4

Central Limit Theorem

Let $X_1, \cdots, X_n, \cdots$ be independent random variables that have the same distribution function and therefore the same mean μ and the same variance σ^2. Let $Y_n = X_1 + \cdots + X_n$. Then the random variable

$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \tag{18}$$

is **asymptotically normal** *with mean* 0 *and variance* 1; *that is, the distribution function $F_n(x)$ of Z_n satisfies*

$$\lim_{n\to\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

A proof can be found in Ref. [G3] listed in App. 1.

Hence, when applying Tables 25.1–25.3 to a nonnormal distribution, we must use ***sufficiently large samples***. As a rule of thumb, if the sample indicates that the skewness of the distribution (the asymmetry; see Team Project 20(d), Problem Set 24.6) is small, use at least $n = 20$ for the mean and at least $n = 50$ for the variance.

PROBLEM SET 25.3

1. Why are interval estimates generally more useful than point estimates?

2–6 MEAN (VARIANCE KNOWN)

2. Find a 95% confidence interval for the mean of a normal population with standard deviation 4.00 from the sample 39, 51, 49, 43, 57, 59. Does that interval get longer or shorter if we take $\gamma = 0.99$ instead of 0.95? By what factor?
3. By what factor does the length of the interval in Prob. 2 change if we double the sample size?
4. Determine a 95% confidence interval for the mean μ of a normal population with variance $\sigma^2 = 16$, using a sample of size 200 with mean 74.81.
5. What sample size would be needed for obtaining a 95% confidence interval (3) of length 2σ? Of length σ?
6. What sample size is needed to obtain a 99% confidence interval of length 2.0 for the mean of a normal population with variance 25? Use Fig. 526. Check by calculation.

MEAN (VARIANCE UNKNOWN)

7. Find a 95% confidence interval for the percentage of cars on a certain highway that have poorly adjusted brakes, using a random sample of 800 cars stopped at a roadblock on that highway, 126 of which had poorly adjusted brakes.
8. **K. Pearson result.** Find a 99% confidence interval for p in the binomial distribution from a classical result by K. Pearson, who in 24,000 trials of tossing a coin obtained 12,012 Heads. Do you think that the coin was fair?

9–11 Find a 99% confidence interval for the mean of a normal population from the sample:

9. Copper content (%) of brass 66, 66, 65, 64, 66, 67, 64, 65, 63, 64
10. Melting point (°C) of aluminum 660, 667, 654, 663, 662
11. Knoop hardness of diamond 9500, 9800, 9750, 9200, 9400, 9550
12. **CAS EXPERIMENT. Confidence Intervals.** Obtain 100 samples of size 10 of the standardized normal distribution. Calculate from them and graph the corresponding 95% confidence intervals for the mean and count how many of them do not contain 0. Does the result support the theory? Repeat the whole experiment, compare and comment.

13–17 VARIANCE

Find a 95% confidence interval for the variance of a normal population from the sample:

13. Length of 20 bolts with sample mean 20.2 cm and sample variance 0.04 cm^2
14. Carbon monoxide emission (grams per mile) of a certain type of passenger car (cruising at 55 mph): 17.3, 17.8, 18.0, 17.7, 18.2, 17.4, 17.6, 18.1
15. Mean energy (keV) of delayed neutron group (Group 3, half-life 6.2 s) for uranium U^{235} fission: a sample of 100 values with mean 442.5 and variance 9.3
16. Ultimate tensile strength (k psi) of alloy steel (Maraging H) at room temperature: 251, 255, 258, 253, 253, 252, 250, 252, 255, 256
17. The sample in Prob. 9
18. If X_1 and X_2 are independent normal random variables with mean 14 and 8 and variance 2 and 5, respectively, what distribution does $3X_1 - X_2$ have? *Hint.* Use Team Project 14(g) in Sec. 24.8.
19. A machine fills boxes weighing Y lb with X lb of salt, where X and Y are normal with mean 100 lb and 5 lb and standard deviation 1 lb and 0.5 lb, respectively. What percent of filled boxes weighing between 104 lb and 106 lb are to be expected?
20. If the weight X of bags of cement is normally distributed with a mean of 40 kg and a standard deviation of 2 kg, how many bags can a delivery truck carry so that the probability of the total load exceeding 2000 kg will be 5%?

25.4 Testing of Hypotheses. Decisions

The ideas of confidence intervals and of tests[2] are the two most important ideas in modern statistics. In a statistical **test** we make inference from sample to population through testing a **hypothesis**, resulting from experience or observations, from a theory or a quality requirement, and so on. In many cases the result of a test is used as a basis for a **decision**, for instance, to

[2]Beginning around 1930, a systematic theory of tests was developed by NEYMAN (see Sec. 25.3) and EGON SHARPE PEARSON (1895–1980), English statistician, the son of Karl Pearson (see the footnote on p. 1086).

buy (or not to buy) a certain model of car, depending on a test of the fuel efficiency (miles/gal) (and other tests, of course), to apply some medication, depending on a test of its effect; to proceed with a marketing strategy, depending on a test of consumer reactions, etc.

Let us explain such a test in terms of a typical example and introduce the corresponding standard notions of statistical testing.

EXAMPLE 1 Test of a Hypothesis. Alternative. Significance Level α

We want to buy 100 coils of a certain kind of wire, provided we can verify the manufacturer's claim that the wire has a breaking limit $\mu = \mu_0 = 200$ lb (or more). This is a test of the **hypothesis** (also called *null hypothesis*) $\mu = \mu_0 = 200$. We shall not buy the wire if the (statistical) test shows that actually $\mu = \mu_1 < \mu_0$, the wire is weaker, the claim does not hold. μ_1 is called the **alternative** (or *alternative hypothesis*) of the test. We shall **accept** the hypothesis if the test suggests that it is true, except for a small error probability α, called the **significance level** of the test. Otherwise we **reject** the hypothesis. Hence α is the probability of rejecting a hypothesis although it is true. The choice of α is up to us. 5% and 1% are popular values.

For the test we need a sample. We randomly select 25 coils of the wire, cut a piece from each coil, and determine the breaking limit experimentally. Suppose that this sample of $n = 25$ values of the breaking limit has the mean $\bar{x} = 197$ lb (somewhat less than the claim!) and the standard deviation $s = 6$ lb.

At this point we could only speculate whether this difference $197 - 200 = -3$ is due to randomness, is a chance effect, or whether it is **significant**, due to the actually inferior quality of the wire. To continue beyond speculation requires probability theory, as follows.

We assume that the breaking limit is normally distributed. (This assumption could be tested by the method in Sec. 25.7. Or we could remember the central limit theorem (Sec. 25.3) and take a still larger sample.) Then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

in (11), Sec. 25.3, with $\mu = \mu_0$ has a ***t*-distribution** with $n - 1$ degrees of freedom ($n - 1 = 24$ for our sample). Also $\bar{x} = 197$ and $s = 6$ are observed values of $\bar{X}$ and S to be used later. We can now choose a significance level, say, $\alpha = 5\%$. From Table A9 in App. 5 or from a CAS we then obtain a critical value c such that $P(T \leqq c) = \alpha = 5\%$. For $P(T \leqq \tilde{c}) = 1 - \alpha = 95\%$ the table gives $\tilde{c} = 1.71$, so that $c = -\tilde{c} = -1.71$ because of the symmetry of the distribution (Fig. 532).

We now reason as follows—this is the ***crucial idea*** of the test. If the hypothesis is true, we have a chance of only α (= 5%) that we observe a value t of T (calculated from a sample) that will fall between $-\infty$ and -1.71. Hence, if we nevertheless do observe such a t, we assert that the hypothesis cannot be true and we reject it. Then we accept the alternative. If, however, $t \geqq c$, we accept the hypothesis.

A simple calculation finally gives $t = (197 - 200)/(6/\sqrt{25}) = -2.5$ as an observed value of T. Since $-2.5 < -1.71$, we reject the hypothesis (the manufacturer's claim) and accept the alternative $\mu = \mu_1 < 200$, the wire seems to be weaker than claimed. ■

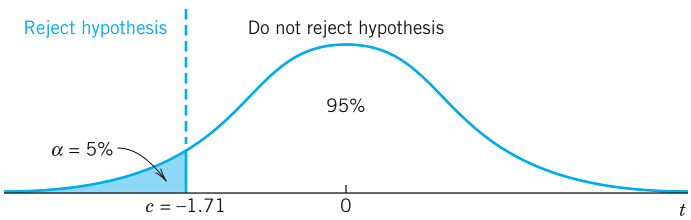


Fig. 532. *t*-distribution in Example 1

This example illustrates the ***steps of a test:***

1. Formulate the **hypothesis** $\theta = \theta_0$ to be tested. ($\theta_0 = \mu_0$ in the example.)
2. Formulate an **alternative** $\theta = \theta_1$. ($\theta_1 = \mu_1$ in the example.)
3. Choose a **significance level** α (5%, 1%, 0.1%).
4. Use a random variable $\hat{\Theta} = g(X_1, \cdots, X_n)$ whose distribution depends on the hypothesis and on the alternative, and this distribution is known in both cases. Determine

a critical value c from the distribution of $\hat{\Theta}$, assuming the hypothesis to be true. (In the example, $\hat{\Theta} = T$, and c is, obtained from $P(T \leqq c) = \alpha$.)

5. Use a sample $x_1, \cdots, x_n$ to determine an observed value $\hat{\theta} = g(x_1, \cdots, x_n)$ of $\hat{\Theta}$. (t in the example.)

6. Accept or reject the hypothesis, depending on the size of $\hat{\theta}$ relative to c. ($t < c$ in the example, rejection of the hypothesis.)

Two important facts require further discussion and careful attention. The first is the choice of an alternative. In the example, $\mu_1 < \mu_0$, but other applications may require $\mu_1 > \mu_0$ or $\mu_1 \neq \mu_0$. The second fact has to do with errors. We know that α (the significance level of the test) is the probability of *rejecting* a *true* hypothesis. And we shall discuss the probability β of *accepting* a *false* hypothesis.

One-Sided and Two-Sided Alternatives (Fig. 533)

Let θ be an unknown parameter in a distribution, and suppose that we want to test the hypothesis $\theta = \theta_0$. Then there are three main kinds of alternatives, namely,

(1) $$\theta > \theta_0$$

(2) $$\theta < \theta_0$$

(3) $$\theta \neq \theta_0.$$

(1) and (2) are **one-sided alternatives**, and (3) is a **two-sided alternative**.

We call **rejection region** (or **critical region**) the region such that we reject the hypothesis if the observed value in the test falls in this region. In ① the critical c lies to the right of θ_0 because so does the alternative. Hence the rejection region extends to the right. This is called a **right-sided test**. In ② the critical c lies to the left of θ_0 (as in Example 1), the rejection region extends to the left, and we have a **left-sided test** (Fig. 533, middle part). These are **one-sided tests**. In ③ we have two rejection regions. This is called a **two-sided test** (Fig. 533, lower part).

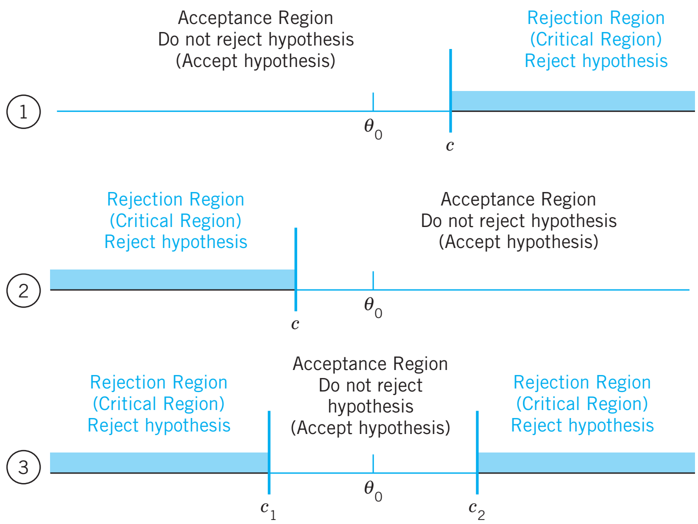


Fig. 533. Test in the case of alternative (1) (upper part of the figure), alternative (2) (middle part), and alternative (3)

All three kinds of alternatives occur in practical problems. For example, (1) may arise if θ_0 is the maximum tolerable inaccuracy of a voltmeter or some other instrument. Alternative (2) may occur in testing strength of material, as in Example 1. Finally, θ_0 in (3) may be the diameter of axle-shafts, and shafts that are too thin or too thick are equally undesirable, so that we have to watch for deviations in both directions.

Errors in Tests

Tests always involve **risks of making false decisions:**

(I) Rejecting a true hypothesis **(Type I error)**.
α = Probability of making a Type I error.

(II) Accepting a false hypothesis **(Type II error)**.
β = Probability of making a Type II error.

Clearly, we cannot avoid these errors because no absolutely certain conclusions about populations can be drawn from samples. But we show that there are ways and means of choosing suitable levels of risks, that is, of values α and β. The choice of α depends on the nature of the problem (e.g., a small risk $\alpha = 1\%$ is used if it is a matter of life or death).

Let us discuss this systematically for a test of a hypothesis $\theta = \theta_0$ against an alternative that is a single number θ_1, for simplicity. We let $\theta_1 > \theta_0$, so that we have a right-sided test. For a left-sided or a two-sided test the discussion is quite similar.

We choose a critical $c > \theta_0$ (as in the upper part of Fig. 533, by methods discussed below). From a given sample $x_1, \cdots, x_n$ we then compute a value

$$\hat{\theta} = g(x_1, \cdots, x_n)$$

with a suitable g (whose choice will be a main point of our further discussion; for instance, take $g = (x_1 + \cdots + x_n)/n$ in the case in which θ is the mean). If $\hat{\theta} > c$, we reject the hypothesis. If $\hat{\theta} \leqq c$, we accept it. Here, the value $\hat{\theta}$ can be regarded as an observed value of the random variable

$$\hat{\Theta} = g(X_1, \cdots, X_n) \tag{4}$$

because x_j may be regarded as an observed value of X_j, $j = 1, \cdots, n$. In this test there are two possibilities of making an error, as follows.

Type I Error (see Table 25.4). The hypothesis is true but is rejected (hence the alternative is accepted) because Θ assumes a value $\hat{\theta} > c$. Obviously, the probability of making such an error equals

$$P(\hat{\Theta} > c)_{\theta=\theta_0} = \alpha. \tag{5}$$

α is called the **significance level** of the test, as mentioned before.

Type II Error (see Table 25.4). The hypothesis is false but is accepted because $\hat{\Theta}$ assumes a value $\hat{\theta} \leqq c$. The probability of making such an error is denoted by β; thus

$$P(\hat{\Theta} \leqq c)_{\theta=\theta_1} = \beta. \tag{6}$$

$\eta = 1 - \beta$ is called the **power** of the test. Obviously, the power η is the probability of avoiding a Type II error.

Table 25.4 Type I and Type II Errors in Testing a Hypothesis $\theta = \theta_0$ Against an Alternative $\theta = \theta_1$

		Unknown Truth	
		$\theta = \theta_0$	$\theta = \theta_1$
Accepted	$\theta = \theta_0$	True decision $P = 1 - \alpha$	Type II error $P = \beta$
	$\theta = \theta_1$	Type 1 error $P = \alpha$	True decision $P = 1 - \beta$

Formulas (5) and (6) show that both α and β depend on c, and we would like to choose c so that these probabilities of making errors are as small as possible. But the important Figure 534 shows that these are conflicting requirements because to let α decrease we must shift c to the right, but then β increases. In practice we first choose α (5%, sometimes 1%), then determine c, and finally compute β. If β is large so that the power $\eta = 1 - \beta$ is small, we should repeat the test, choosing a larger sample, for reasons that will appear shortly.

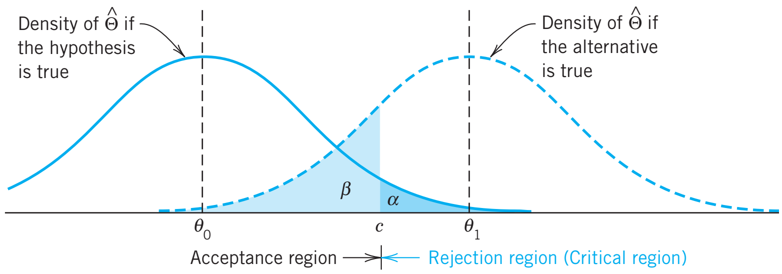


Fig. 534. Illustration of Type I and II errors in testing a hypothesis $\theta = \theta_0$ against an alternative $\theta = \theta_1$ ($> \theta_0$, right-sided test)

If the alternative is not a single number but is of the form (1)–(3), then β becomes a function of θ. This function $\beta(\theta)$ is called the **operating characteristic** (OC) of the test and its curve the **OC curve**. Clearly, in this case $\eta = 1 - \beta$ also depends on θ. This function $\eta(\theta)$ is called the **power function** of the test. (Examples will follow.)

Of course, from a test that leads to the acceptance of a certain hypothesis θ_0, it does ***not*** follow that this is the only possible hypothesis or the best possible hypothesis. Hence the terms "**not reject**" or "**fail to reject**" are perhaps better than the term "**accept**."

Test for μ of the Normal Distribution with Known σ^2

The following example explains the three kinds of hypotheses.

EXAMPLE 2 Test for the Mean of the Normal Distribution with Known Variance

Let X be a normal random variable with variance $\sigma^2 = 9$. Using a sample of size $n = 10$ with mean $\bar{x}$, test the hypothesis $\mu = \mu_0 = 24$ against the three kinds of alternatives, namely,

$$\text{(a)} \quad \mu > \mu_0 \qquad \text{(b)} \quad \mu < \mu_0 \qquad \text{(c)} \quad \mu \neq \mu_0.$$

Solution. We choose the significance level $\alpha = 0.05$. An estimate of the mean will be obtained from

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

If the hypothesis is true, $\bar{X}$ is normal with mean $\mu = 24$ and variance $\sigma^2/n = 0.9$, see Theorem 1, Sec. 25.3. Hence we may obtain the critical value c from Table A8 in App. 5.

Case (a). ***Right-Sided Test.*** We determine c from $P(\bar{X} > c)_{\mu=24} = \alpha = 0.05$, that is,

$$P(\bar{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c-24}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $(c - 24)/\sqrt{0.9} = 1.645$, and $c = 25.56$, which is greater than μ_0, as in the upper part of Fig. 533. If $\bar{x} \leqq 25.56$, the hypothesis is accepted. If $\bar{x} > 25.56$, it is rejected. The power function of the test is (Fig. 535)

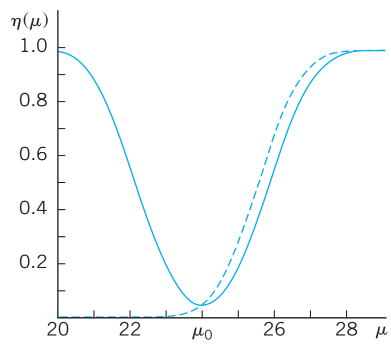


Fig. 535. Power function $\eta(\mu)$ in Example 2, case (a) (dashed) and case (c)

$$\begin{aligned} \eta(\mu) &= P(\bar{X} > 25.56)_\mu = 1 - P(\bar{X} \leqq 25.56)_\mu \\ &= 1 - \Phi\left(\frac{25.56 - \mu}{\sqrt{0.9}}\right) = 1 - \Phi(26.94 - 1.05\mu) \end{aligned} \tag{7}$$

Case (b). ***Left-Sided Test.*** The critical value c is obtained from the equation

$$P(\bar{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c-24}{\sqrt{0.9}}\right) = \alpha = 0.05.$$

Table A8 in App. 5 yields $c = 24 - 1.56 = 22.44$. If $\bar{x} \geqq 22.44$, we accept the hypothesis. If $\bar{x} < 22.44$, we reject it. The power function of the test is

$$\eta(\mu) = P(\bar{X} \leqq 22.44)_\mu = \Phi\left(\frac{22.44 - \mu}{\sqrt{0.9}}\right) = \Phi(23.65 - 1.05\mu). \tag{8}$$

Case (c). ***Two-Sided Test.*** Since the normal distribution is symmetric, we choose c_1 and c_2 equidistant from $\mu = 24$, say, $c_1 = 24 - k$ and $c_2 = 24 + k$, and determine k from

$$P(24 - k \leqq \bar{X} \leqq 24 + k)_{\mu=24} = \Phi\left(\frac{k}{\sqrt{0.9}}\right) - \Phi\left(-\frac{k}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $k/\sqrt{0.9} = 1.960$, hence $k = 1.86$. This gives the values $c_1 = 24 - 1.86 = 22.14$ and $c_2 = 24 + 1.86 = 25.86$. If $\bar{x}$ is not smaller than c_1 and not greater than c_2, we accept the hypothesis. Otherwise we reject it. The power function of the test is (Fig. 535)

$$\eta(\mu) = P(\bar{X} < 22.14)_\mu + P(\bar{X} > 25.86)_\mu = P(\bar{X} < 22.14)_\mu + 1 - P(\bar{X} \leqq 25.86)_\mu$$

$$= 1 + \Phi\left(\frac{22.14 - \mu}{\sqrt{0.9}}\right) - \Phi\left(\frac{25.86 - \mu}{\sqrt{0.9}}\right) \tag{9}$$

$$= 1 + \Phi(23.34 - 1.05\mu) - \Phi(27.26 - 1.05\mu).$$

Consequently, the operating characteristic $\beta(\mu) = 1 - \eta(\mu)$ (see before) is (Fig. 536)

$$\beta(\mu) = \Phi(27.26 - 1.05\mu) - \Phi(23.34 - 1.05\mu).$$

If we take a larger sample, say, of size $n = 100$ (instead of 10), then $\sigma^2/n = 0.09$ (instead of 0.9) and the critical values are $c_1 = 23.41$ and $c_2 = 24.59$, as can be readily verified. Then the operating characteristic of the test is

$$\beta(\mu) = \Phi\left(\frac{24.59 - \mu}{\sqrt{0.09}}\right) - \Phi\left(\frac{23.41 - \mu}{\sqrt{0.09}}\right)$$

$$= \Phi(81.97 - 3.33\mu) - \Phi(78.03 - 3.33\mu).$$

Figure 536 shows that the corresponding OC curve is steeper than that for $n = 10$. This means that the increase of n has led to an improvement of the test. In any practical case, n is chosen as small as possible but so large that the test brings out deviations between μ and μ_0 that are of practical interest. For instance, if deviations of ± 2 units are of interest, we see from Fig. 536 that $n = 10$ is much too small because when $\mu = 24 - 2 = 22$ or $\mu = 24 + 2 = 26$ β is almost 50%. On the other hand, we see that $n = 100$ is sufficient for that purpose. ■

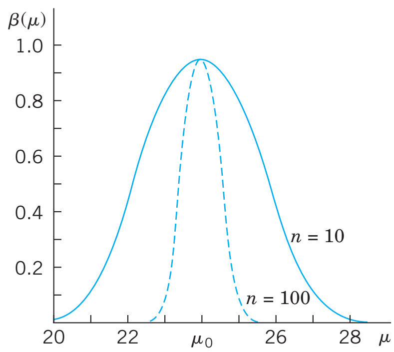


Fig. 536. Curves of the operating characteristic (OC curves) in Example 2, case (c), for two different sample sizes n

Test for μ When σ^2 Is Unknown, and for σ^2

EXAMPLE 3 Test for the Mean of the Normal Distribution with Unknown Variance

The tensile strength of a sample of $n = 16$ manila ropes (diameter 3 in.) was measured. The sample mean was $\bar{x} = 4482$ kg, and the sample standard deviation was $s = 115$ kg (N. C. Wiley, 41st Annual Meeting of the American Society for Testing Materials). Assuming that the tensile strength is a normal random variable, test the hypothesis $\mu_0 = 4500$ kg against the alternative $\mu_1 = 4400$ kg. Here μ_0 may be a value given by the manufacturer, while μ_1 may result from previous experience.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, it follows from Theorem 2 in Sec. 25.3, that the random variable

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 4500}{S/4}$$

has a t-distribution with $n - 1 = 15$ d.f. The test is left-sided. The critical value c is obtained from $P(T < c)_{\mu_0} = \alpha = 0.05$. Table A9 in App. 5 gives $c = -1.75$. As an observed value of T we obtain from the sample $t = (4482 - 4500)/(115/4) = -0.626$. We see that $t > c$ and accept the hypothesis. For obtaining numeric values of the power of the test, we would need tables called noncentral Student t-tables; we shall not discuss this question here. ■

EXAMPLE 4 **Test for the Variance of the Normal Distribution**

Using a sample of size $n = 15$ and sample variance $s^2 = 13$ from a normal population, test the hypothesis $\sigma^2 = \sigma_0^2 = 10$ against the alternative $\sigma^2 = \sigma_1^2 = 20$.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, then

$$Y = (n-1)\frac{S^2}{\sigma_0^2} = 14\frac{S^2}{10} = 1.4S^2$$

has a chi-square distribution with $n - 1 = 14$ d.f. by Theorem 3, Sec. 25.3. From

$$P(Y > c) = \alpha = 0.05, \qquad \text{that is,} \qquad P(Y \leqq c) = 0.95,$$

and Table A10 in App. 5 with 14 degrees of freedom we obtain $c = 23.68$. This is the critical value of Y. Hence to $S^2 = \sigma_0^2 Y/(n-1) = 0.714Y$ there corresponds the critical value $c^* = 0.714 \cdot 23.68 = 16.91$. Since $s^2 < c^*$, we accept the hypothesis.

If the alternative is true, the random variable $Y_1 = 14S^2/\sigma_1^2 = 0.7S^2$ has a chi-square distribution with 14 d.f. Hence our test has the power

$$\eta = P(S^2 > c^*)_{\sigma^2=20} = P(Y_1 > 0.7c^*)_{\sigma^2=20} = 1 - P(Y_1 \leqq 11.84)_{\sigma^2=20}.$$

From a more extensive table of the chi-square distribution (e.g. in Ref. [G3] or [G8]) or from your CAS, you see that $\eta \approx 62\%$. Hence the Type II risk is very large, namely, 38%. To make this risk smaller, we would have to increase the sample size. ■

Comparison of Means and Variances

EXAMPLE 5 **Comparison of the Means of Two Normal Distributions**

Using a sample $x_1, \cdots, x_{n_1}$ from a normal distribution with unknown mean μ_x and a sample $y_1, \cdots, y_{n_2}$ from another normal distribution with unknown mean μ_y, we want to test the hypothesis that the means are equal, $\mu_x = \mu_y$, against an alternative, say, $\mu_x > \mu_y$. The variances need not be known but are assumed to be equal.[3]

Two cases of comparing means are of practical importance:

Case A. *The samples have the* ***same size.*** *Furthermore, each value of the first sample corresponds to precisely one value of the other,* because corresponding values result from the same person or thing **(paired comparison)**—for example, two measurements of the same thing by two different methods or two measurements from the two eyes of the same person. More generally, they may result from pairs of *similar* individuals or things, for example, identical twins, pairs of used front tires from the same car, etc. Then we should form the differences of corresponding values and test the hypothesis that the population corresponding to the differences has mean 0, using the method in Example 3. If we have a choice, this method is better than the following.

[3]This assumption of equality of variances can be tested, as shown in the next example. If the test shows that they differ significantly, choose two samples of the same size $n_1 = n_2 = n$ (not too small, > 30, say), use the test in Example 2 together with the fact that (12) is an observed value of an approximately standardized normal random variable.

Case B. *The two samples are independent and not necessarily of the same size.* Then we may proceed as follows. Suppose that the alternative is $\mu_x > \mu_y$. We choose a significance level α. Then we compute the sample means $\bar{x}$ and $\bar{y}$ as well as $(n_1 - 1)s_x^2$ and $(n_2 - 1)s_y^2$, where s_x^2 and s_y^2 are the sample variances. Using Table A9 in App. 5 with $n_1 + n_2 - 2$ degrees of freedom, we now determine c from

$$P(T \leqq c) = 1 - \alpha. \tag{10}$$

We finally compute

$$t_0 = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}}. \tag{11}$$

It can be shown that this is an observed value of a random variable that has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, provided the hypothesis is true. If $t_0 \leqq c$, the hypothesis is accepted. If $t_0 > c$, it is rejected.

If the alternative is $\mu_x \neq \mu_y$, then (10) must be replaced by

$$P(T \leqq c_1) = 0.5\alpha, \qquad P(T \leqq c_2) = 1 - 0.5\alpha. \tag{10*}$$

Note that for samples of equal size $n_1 = n_2 = n$, formula (11) reduces to

$$t_0 = \sqrt{n} \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}}. \tag{12}$$

To illustrate the computations, let us consider the two samples $(x_1, \cdots, x_{n_1})$ and $(y_1, \cdots, y_{n_2})$ given by

105 108 86 103 103 107 124 105

and

89 92 84 97 103 107 111 97

showing the relative output of tin plate workers under two different working conditions [J. J. B. Worth, *Journal of Industrial Engineering* **9**, 249–253). Assuming that the corresponding populations are normal and have the same variance, let us test the hypothesis $\mu_x = \mu_y$ against the alternative $\mu_x \neq \mu_y$. (Equality of variances will be tested in the next example.)

Solution. We find

$$\bar{x} = 105.125, \qquad \bar{y} = 97.500, \qquad s_x^2 = 106.125. \qquad s_y^2 = 84.000.$$

We choose the significance level $\alpha = 5\%$. From (10*) with $0.5\alpha = 2.5\%$, $1 - 0.5\alpha = 97.5\%$ and Table A9 in App. 5 with 14 degrees of freedom we obtain $c_1 = -2.14$ and $c_2 = 2.14$. Formula (12) with $n = 8$ gives the value

$$t_0 = \sqrt{8} \cdot 7.625/\sqrt{190.125} = 1.56.$$

Since $c_1 \leqq t_0 \leqq c_2$, we ***accept the hypothesis*** $\mu_x = \mu_y$ that under both conditions the mean output is the same.

Case A applies to the example because the two first sample values correspond to a certain type of work, the next two were obtained in another kind of work, etc. So we may use the differences

16 16 2 6 0 0 13 8

of corresponding sample values and the method in Example 3 to test the hypothesis $\mu = 0$, where μ is the mean of the population corresponding to the differences. As a logical alternative we take $\mu \neq 0$. The sample mean is $\bar{d} = 7.625$, and the sample variance is $s^2 = 45.696$. Hence

$$t = \sqrt{8}\,(7.625 - 0)/\sqrt{45.696} = 3.19.$$

From $P(T \leqq c_1) = 2.5\%$, $P(T \leqq c_2) = 97.5\%$ and Table A9 in App. 5 with $n - 1 = 7$ degrees of freedom we obtain $c_1 = -2.36$, $c_2 = 2.36$ and ***reject the hypothesis*** because $t = 3.19$ does not lie between c_1 and c_2. Hence our present test, in which we used more information (but the same samples), shows that the difference in output is significant. ■

EXAMPLE 6 Comparison of the Variance of Two Normal Distributions

Using the two samples in the last example, test the hypothesis $\sigma_x^2 = \sigma_y^2$; assume that the corresponding populations are normal and the nature of the experiment suggests the alternative $\sigma_x^2 > \sigma_y^2$.

Solution. We find $s_x^2 = 106.125$, $s_y^2 = 84.000$. We choose the significance level $\alpha = 5\%$. Using $P(V \leqq c) = 1 - \alpha = 95\%$ and Table A11 in App. 5, with $(n_1 - 1, n_2 - 1) = (7, 7)$ degrees of freedom, we determine $c = 3.79$. We finally compute $v_0 = s_x^2/s_y^2 = 1.26$. Since $v_0 \leqq c$, we accept the hypothesis. If $v_0 < c$, we would reject it.

This test is justified by the fact that v_0 is an observed value of a random variable that has a so-called ***F*-distribution** with $(n_1 - 1, n_2 - 1)$ degrees of freedom, provided the hypothesis is true. (Proof in Ref. [G3] listed in App. 1.) The F-distribution with (m, n) degrees of freedom was introduced by R. A. Fisher[4] and has the distribution function $F(z) = 0$ if $z < 0$ and

$$F(z) = K_{mn} \int_0^z t^{(m-2)/2}(mt + n)^{-(m+n)/2}\, dt \qquad (z \geqq 0), \tag{13}$$

where $K_{mn} = m^{m/2}n^{n/2}\Gamma(\frac{1}{2}m + \frac{1}{2}n)/\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)$. (For Γ see App. A3.1.) ■

This long section contained the basic ideas and concepts of testing, along with typical applications and you may perhaps want to review it quickly before going on, because the next sections concern an adaptation of these ideas to tasks of great practical importance and resulting tests in connection with quality control, acceptance (or rejection) of goods produced, and so on.

PROBLEM SET 25.4

1. From memory: Make a list of the three types of alternatives, each with a typical example of your own.

2. Make a list of methods in this section, each with the distribution needed in testing.

3. Test $\mu = 0$ against $\mu > 0$, assuming normality and using the sample 0, 1, -1, 3, -8, 6, 1 (deviations of the azimuth [multiples of 0.01 radian] in some revolution of a satellite). Choose $\alpha = 5\%$.

4. In one of his classical experiments Buffon obtained 2048 heads in tossing a coin 4040 times. Was the coin fair?

5. Do the same test as in Prob. 4, using a result by K. Pearson, who obtained 6019 heads in 12,000 trials.

6. Assuming normality and known variance $\sigma^2 = 9$, test the hypothesis $\mu = 60.0$ against the alternative $\mu = 57.0$ using a sample of size 20 with mean $\bar{x} = 58.50$ and choosing $\alpha = 5\%$.

7. How does the result in Prob. 6 change if we use a smaller sample, say, of size 5, the other data ($\bar{x} = 58.05$, $\alpha = 5\%$, etc.) remaining as before?

8. Determine the power of the test in Prob. 6.

9. What is the rejection region in Prob. 6 in the case of a two-sided test with $\alpha = 5\%$?

10. CAS EXPERIMENT. Tests of Means and Variances.
(a) Obtain 100 samples of size 10 each from the normal distribution with mean 100 and variance 25. For each sample, test the hypothesis $\mu_0 = 100$ against the alternative $\mu_1 > 100$ at the level of $\alpha = 10\%$. Record the number of rejections of the hypothesis. Do the whole experiment once more and compare.

(b) Set up a similar experiment for the variance of a normal distribution and perform it 100 times.

11. A firm sells oil in cans containing 5000 g oil per can and is interested to know whether the mean weight differs significantly from 5000 g at the 5% level, in which case the filling machine has to be adjusted. Set up a hypothesis and an alternative and perform the test, assuming normality and using a sample of 50 fillings with mean 4990 g and standard deviation 20 g.

---

[4]After the pioneering work of the English statistician and biologist, KARL PEARSON (1857–1936), the founder of the English school of statistics, and WILLIAM SEALY GOSSET (1876–1937), who discovered the t-distribution (and published under the name "Student"), the English statistician Sir RONALD AYLMER FISHER (1890–1962), professor of eugenics in London (1933–1943) and professor of genetics in Cambridge, England (1943–1957) and Adelaide, Australia (1957–1962), had great influence on the further development of modern statistics.

12. If a sample of 25 tires of a certain kind has a mean life of 37,000 miles and a standard deviation of 5000 miles, can the manufacturer claim that the true mean life of such tires is greater than 35,000 miles? Set up and test a corresponding hypothesis at the 5% level, assuming normality.

13. If simultaneous measurements of electric voltage by two different types of voltmeter yield the differences (in volts) 0.4, -0.6, 0.2, 0.0, 1.0, 1.4, 0.4, 1.6, can we assert at the 5% level that there is no significant difference in the calibration of the two types of instruments? Assume normality.

14. If a standard medication cures about 75% of patients with a certain disease and a new medication cured 310 of the first 400 patients on whom it was tried, can we conclude that the new medication is better? Choose $\alpha = 5\%$. First guess. Then calculate.

15. Suppose that in the past the standard deviation of weights of certain 100.0-oz packages filled by a machine was 0.8 oz. Test the hypothesis H_0: $\sigma = 0.8$ against the alternative H_1: $\sigma > 0.8$ (an undesirable increase), using a sample of 20 packages with standard deviation 1.0 oz and assuming normality. Choose $\alpha = 5\%$.

16. Suppose that in operating battery-powered electrical equipment, it is less expensive to replace all batteries at fixed intervals than to replace each battery individually when it breaks down, provided the standard deviation of the lifetime is less than a certain limit, say, less than 5 hours. Set up and apply a suitable test, using a sample of 28 values of lifetimes with standard deviation $s = 3.5$ hours and assuming normality: choose $\alpha = 5\%$.

17. Brand A gasoline was used in 16 similar automobiles under identical conditions. The corresponding sample of 16 values (miles per gallon) had mean 19.6 and standard deviation 0.4. Under the same conditions, high-power brand B gasoline gave a sample of 16 values with mean 20.2 and standard deviation 0.6. Is the mileage of B significantly better than that of A? Test at the 5% level; assume normality. First guess. Then calculate.

18. The two samples 70, 80, 30, 70, 60, 80 and 140, 120, 130, 120, 120, 130, 120 are values of the differences of temperatures (°C) of iron at two stages of casting, taken from two different crucibles. Is the variance of the first population larger than that of the second? Assume normality. Choose $\alpha = 5\%$.

19. Show that for a normal distribution the two types of errors in a test of a hypothesis H_0: $\mu = \mu_0$ against an alternative H_1: $\mu = \mu_1$ can be made as small as one pleases (not zero!) by taking the sample sufficiently large.

20. Test for equality of population means against the alternative that the means are different assuming normality, choosing $\alpha = 5\%$ and using two samples of sizes 12 and 18, with mean 10 and 14, respectively, and equal standard deviation 3.

25.5 Quality Control

The ideas on testing can be adapted and extended in various ways to serve basic practical needs in engineering and other fields. We show this in the remaining sections for some of the most important tasks solvable by statistical methods. As a first such area of problems, we discuss industrial quality control, a highly successful method used in various industries.

No production process is so perfect that all the products are completely alike. There is always a small variation that is caused by a great number of small, uncontrollable factors and must therefore be regarded as a chance variation. It is important to make sure that the products have required values (for example, length, strength, or whatever property may be essential in a particular case). For this purpose one makes a test of the hypothesis that the products have the required property, say, $\mu = \mu_0$, where μ_0 is a required value. If this is done after an entire lot has been produced (for example, a lot of 100,000 screws), the test will tell us how good or how bad the products are, but it it obviously too late to alter undesirable results. It is much better to test during the production run. This is done at regular intervals of time (for example, every hour or half-hour) and is called **quality control**. Each time a sample of the same size is taken, in practice 3 to 10 times. If the hypothesis is rejected, we stop the production and look for the cause of the trouble.

If we stop the production process even though it is progressing properly, we make a Type I error. If we do not stop the process even though something is not in order, we make a Type II error (see Sec. 25.4). The result of each test is marked in graphical form on what is called a **control chart**. This was proposed by W. A. Shewhart in 1924 and makes quality control particularly effective.

Control Chart for the Mean

An illustration and example of a control chart is given in the upper part of Fig. 537. This control chart for the mean shows the **lower control limit** LCL, the **center control line** CL, and the **upper control limit** UCL. The two **control limits** correspond to the critical values c_1 and c_2 in case (c) of Example 2 in Sec. 25.4. As soon as a sample mean falls outside the range between the control limits, we reject the hypothesis and assert that the

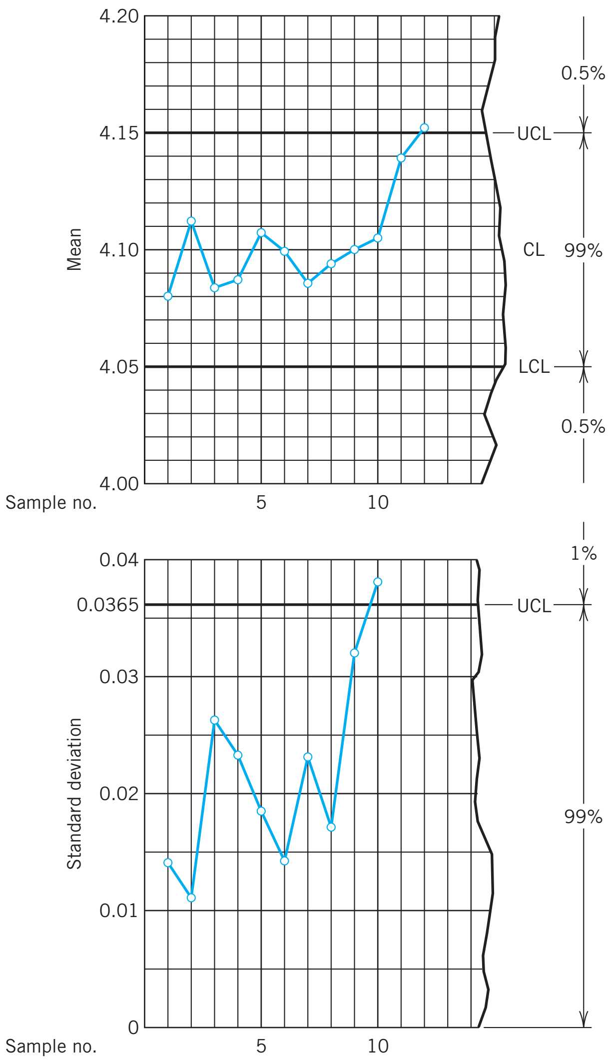


Fig. 537. Control charts for the mean (upper part of figure) and the standard deviation in the case of the samples on p. 1089

production process is "out of control"; that is, we assert that there has been a shift in process level. Action is called for whenever a point exceeds the limits.

If we choose control limits that are too loose, we shall not detect process shifts. On the other hand, if we choose control limits that are too tight, we shall be unable to run the process because of frequent searches for nonexistent trouble. The usual significance level is $\alpha = 1\%$. From Theorem 1 in Sec. 25.3 and Table A8 in App. 5 we see that in the case of the normal distribution the corresponding control limits for the mean are

$$\text{(1)} \qquad \text{LCL} = \mu_0 - 2.58\frac{\sigma}{\sqrt{n}}, \qquad \text{UCL} = \mu_0 + 2.58\frac{\sigma}{\sqrt{n}}.$$

Here σ is assumed to be known. If σ is unknown, we may compute the standard deviations of the first 20 or 30 samples and take their arithmetic mean as an approximation of σ. The broken line connecting the means in Fig. 537 is merely to display the results.

Additional, more subtle controls are often used in industry. For instance, one observes the motions of the sample means above and below the centerline, which should happen frequently. Accordingly, long runs (conventionally of length 7 or more) of means all above (or all below) the centerline could indicate trouble.

Table 25.5 Twelve Samples of Five Values Each (Diameter of Small Cylinders, Measured in Millimeters)

Sample Number	Sample Values					$\bar{x}$	s	R
1	4.06	4.08	4.08	4.08	4.10	4.080	0.014	0.04
2	4.10	4.10	4.12	4.12	4.12	4.112	0.011	0.02
3	4.06	4.06	4.08	4.10	4.12	4.084	0.026	0.06
4	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
5	4.08	4.10	4.12	4.12	4.12	4.108	0.018	0.04
6	4.08	4.10	4.10	4.10	4.12	4.100	0.014	0.04
7	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
8	4.08	4.08	4.10	4.10	4.12	4.096	0.017	0.04
9	4.06	4.08	4.10	4.12	4.14	4.100	0.032	0.08
10	4.06	4.08	4.10	4.12	4.16	4.104	0.038	0.10
11	4.12	4.14	4.14	4.14	4.16	4.140	0.014	0.04
12	4.14	4.14	4.16	4.16	4.16	4.152	0.011	0.02

Control Chart for the Variance

In addition to the mean, one often controls the variance, the standard deviation, or the range. To set up a control chart for the variance in the case of a normal distribution, we may employ the method in Example 4 of Sec. 25.4 for determining control limits. It is customary to use only one control limit, namely, an upper control limit. Now from Example 4 of Sec. 25.4 we have $S^2 = \sigma_0^2 Y/(n-1)$, where, because of our normality assumption, the random variable Y has a chi-square distribution with $n - 1$ degrees of freedom. Hence the desired control limit is

$$\text{(2)} \qquad \text{UCL} = \frac{\sigma^2 c}{n-1}$$

where c is obtained from the equation

$$P(Y > c) = \alpha, \qquad \text{that is,} \qquad P(Y \leqq c) = 1 - \alpha$$

and the table of the chi-square distribution (Table A10 in App. 5) with $n - 1$ degrees of freedom (or from your CAS); here α (5% or 1%, say) is the probability that in a properly running process an observed value s^2 of S^2 is greater than the upper control limit.

If we wanted a control chart for the variance with both an upper control limit UCL and a lower control limit LCL, these limits would be

$$\text{(3)} \qquad \text{LCL} = \frac{\sigma^2 c_1}{n-1} \qquad \text{and} \qquad \text{UCL} = \frac{\sigma^2 c_2}{n-1},$$

where c_1 and c_2 are obtained from Table A10 with $n - 1$ d.f. and the equations

$$\text{(4)} \qquad P(Y \leqq c_1) = \frac{\alpha}{2} \qquad \text{and} \qquad P(Y \leqq c_2) = 1 - \frac{\alpha}{2}.$$

Control Chart for the Standard Deviation

To set up a control chart for the standard deviation, we need an upper control limit

$$\text{(5)} \qquad \text{UCL} = \frac{\sigma\sqrt{c}}{\sqrt{n-1}}$$

obtained from (2). For example, in Table 25.5 we have $n = 5$. Assuming that the corresponding population is normal with standard deviation $\sigma = 0.02$ and choosing $\alpha = 1\%$, we obtain from the equation

$$P(Y \leqq c) = 1 - \alpha = 99\%$$

and Table A10 in App. 5 with 4 degrees of freedom the critical value $c = 13.28$ and from (5) the corresponding value

$$\text{UCL} = \frac{0.02\sqrt{13.28}}{\sqrt{4}} = 0.0365,$$

which is shown in the lower part of Fig. 537.

A control chart for the standard deviation with both an upper and a lower control limit is obtained from (3).

Control Chart for the Range

Instead of the variance or standard deviation, one often controls the **range** R (= largest sample value minus smallest sample value). It can be shown that in the case of the normal distribution, the standard deviation σ is proportional to the expectation of the random variable R^* for which R is an observed value, say, $\sigma = \lambda_n E(R^*)$ where the factor of proportionality λ_n depends on the sample size n and has the values

n	2	3	4	5	6	7	8	9	10
$\lambda_n = \sigma/E(R^*)$	0.89	0.59	0.49	0.43	0.40	0.37	0.35	0.34	0.32

n	12	14	16	18	20	30	40	50
$\lambda_n = \sigma/E(R^*)$	0.31	0.29	0.28	0.28	0.27	0.25	0.23	0.22

Since R depends on two sample values only, it gives less information about a sample than s does. Clearly, the larger the sample size n is, the more information we lose in using R instead of s. A practical rule is to use s when n is larger than 10.

PROBLEM SET 25.5

1. Suppose a machine for filling cans with lubricating oil is set so that it will generate fillings which form a normal population with mean 1 gal and standard deviation 0.02 gal. Set up a control chart of the type shown in Fig. 537 for controlling the mean, that is, find LCL and UCL, assuming that the sample size is 4.

2. Three-sigma control chart. Show that in Prob. 1, the requirement of the significance level $\alpha = 0.3\%$ leads to $\text{LCL} = \mu - 3\sigma/\sqrt{n}$ and $\text{UCL} = \mu + 3\sigma/\sqrt{n}$, and find the corresponding numeric values.

3. What sample size should we choose in Prob. 1 if we want LCL and UCL somewhat closer together, say, $\text{UCL} - \text{LCL} = 0.02$, without changing the significance level?

4. What effect on UCL − LCL does it have if we double the sample size? If we switch from $\alpha = 1\%$ to $\alpha = 5\%$?

5. How should we change the sample size in controlling the mean of a normal population if we want UCL − LCL to decrease to half its original value?

6. Graph the means of the following 10 samples (thickness of gaskets, coded values) on a control chart for means, assuming that the population is normal with mean 5 and standard deviation 1.16.

Time	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00
	5	7	7	4	5	6	5	5	3	3
Sample	2	5	3	4	6	4	5	2	4	6
values	5	4	6	3	4	6	6	5	8	6
	6	4	5	6	6	4	4	3	4	8

7. Graph the ranges of the samples in Prob. 6 on a control chart for ranges.

8. Graph $\lambda_n = \sigma/E(R^*)$ as a function of n. Why is λ_n a monotone decreasing function of n?

9. Eight samples of size 2 were taken from a lot of screws. The values (length in inches) are

Sample No.	1	2	3	4	5	6	7	8
Length	3.50	3.51	3.49	3.52	3.53	3.49	3.48	3.52
	3.51	3.48	3.50	3.50	3.49	3.50	3.47	3.49

Assuming that the population is normal with mean 3.500 and variance 0.0004 and using (1), set up a control chart for the mean and graph the sample means on the chart.

10. Attribute control charts. Fifteen samples of size 100 were taken from a production of containers. The numbers of defectives (leaking containers) in those samples (in the order observed) were

1 4 5 4 9 7 0 5 6 13 0 2 1 12 8

From previous experience it was known that the average fraction defective is $p = 4\%$ provided that the process of production is running properly. Using the binomial distribution, set up a *fraction defective chart* (also called a ***p*-chart**), that is, choose the

LCL = 0 and determine the UCL for the fraction defective (in percent) by the use of 3-sigma limits, where σ^2 is the variance of the random variable

$$\bar{X} = \textit{Fraction defective in a sample of size } 100.$$

Is the process under control?

11. Number of defectives. Find formulas for the UCL, CL, and LCL (corresponding to 3σ-limits) in the case of a control chart for the number of defectives, assuming that, in a state of statistical control, the fraction of defectives is p.

12. CAS PROJECT. Control Charts. (a) Obtain 100 samples of 4 values each from the normal distribution with mean 8.0 and variance 0.16 and their means, variances, and ranges.

(b) Use these samples for making up a control chart for the mean.

(c) Use them on a control chart for the standard deviation.

(d) Make up a control chart for the range.

(e) Describe quantitative properties of the samples that you can see from those charts (e.g., whether the corresponding process is under control, whether the quantities observed vary randomly, etc.).

13. Since the presence of a point outside control limits for the mean indicates trouble, how often would we be making the mistake of looking for nonexistent trouble if we used **(a)** 1-sigma limits, **(b)** 2-sigma limits? Assume normality.

14. What LCL and UCL should we use instead of (1) if, instead of $\bar{x}$, we use the sum $x_1 + \cdots + x_n$ of the sample values? Determine these limits in the case of Fig. 537.

15. Number of defects per unit. A so-called *c-chart* or *defects-per-unit chart* is used for the control of the number X of defects per unit (for instance, the number of defects per 100 meters of paper, the number of missing rivets in an airplane wing, etc.). **(a)** Set up formulas for CL and LCL, UCL corresponding to $\mu \pm 3\sigma$, assuming that X has a Poisson distribution. **(b)** Compute CL, LCL, and UCL in a control process of the number of imperfections in sheet glass; assume that this number is 3.6 per sheet on the average when the process is in control.

25.6 Acceptance Sampling

Acceptance sampling is usually done when products leave the factory (or in some cases even within the factory). The standard situation in acceptance sampling is that a **producer** supplies to a **consumer** (a buyer or wholesaler) a lot of N items (a carton of screws, for instance). The decision to **accept** or **reject** the lot is made by determining the number x of **defectives** (= defective items) in a sample of size n from the lot. The lot is accepted if $x \leqq c$, where c is called the **acceptance number**, giving the allowable number of defectives. If $x > c$, the consumer rejects the lot. Clearly, producer and consumer must agree on a certain **sampling plan** giving n and c.

From the hypergeometric distribution we see that the event A*: "Accept the lot"* has probability (see Sec. 24.7)

$$P(A) = P(X \leqq c) = \sum_{x=0}^{c} \binom{M}{x}\binom{N-M}{n-x} \Big/ \binom{N}{n} \tag{1}$$

where M is the number of defectives in a lot of N items. In terms of the **fraction defective** $\theta = M/N$ we can write (1) as

$$P(A;\theta) = \sum_{x=0}^{c} \binom{N\theta}{x}\binom{N-N\theta}{n-x} \Big/ \binom{N}{n}. \tag{2}$$

$P(A;\theta)$ can assume $n + 1$ values corresponding to $\theta = 0, 1/N, 2/N, \cdots, N/N$; here, n and c are fixed. A monotone smooth curve through these points is called the **operating characteristic curve (OC curve)** of the sampling plan considered.

EXAMPLE 1 Sampling Plan

Suppose that certain tool bits are packaged 20 to a box, and the following sampling plan is used. A sample of two tool bits is drawn, and the corresponding box is accepted if and only if both bits in the sample are good. In this case, $N = 20$, $n = 2$, $c = 0$, and (2) takes the form (a factor 2 drops out)

$$P(A;\theta) = \binom{20\,\theta}{0}\binom{20 - 20\,\theta}{2}\Big/\binom{20}{2}$$
$$= \frac{(20 - 20\,\theta)(19 - 20\,\theta)}{380}.$$

The values of $P(A,\theta)$ for $\theta = 0, 1/20, 2/20, \cdots, 20/20$ and the resulting OC curve are shown in Fig. 538. (Verify!) ■

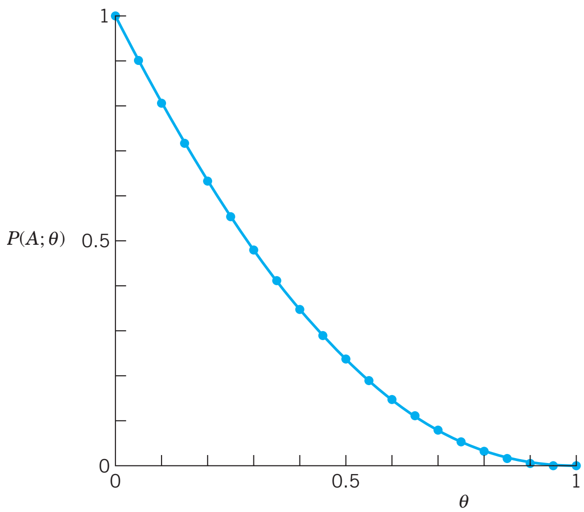


Fig. 538. OC curve of the sampling plan with $n = 2$ and $c = 0$ for lots of size $N = 20$

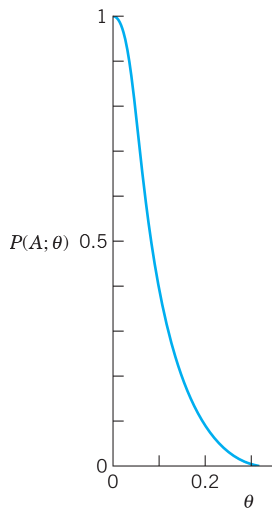


Fig. 539. OC curve in Example 2

In most practical cases θ will be small (less than 10%). Then if we take small samples compared to N, we can approximate (2) by the Poisson distribution (Sec. 24.7); thus

$$P(A;\theta) \sim e^{-\mu}\sum_{x=0}^{c}\frac{\mu^x}{x!} \qquad (\mu = n\theta). \tag{3}$$

EXAMPLE 2 Sampling Plan. Poisson Distribution

Suppose that for large lots the following sampling plan is used. A sample of size $n = 20$ is taken. If it contains not more than one defective, the lot is accepted. If the sample contains two or more defectives, the lot is rejected. In this plan, we obtain from (3)

$$P(A;\theta) \sim e^{-20\,\theta}(1 + 20\,\theta),$$

The corresponding OC curve is shown in Fig. 539. ■

Errors in Acceptance Sampling

We show how acceptance sampling fits into general test theory (Sec. 25.4) and what this means from a practical point of view. The producer wants the probability α of rejecting

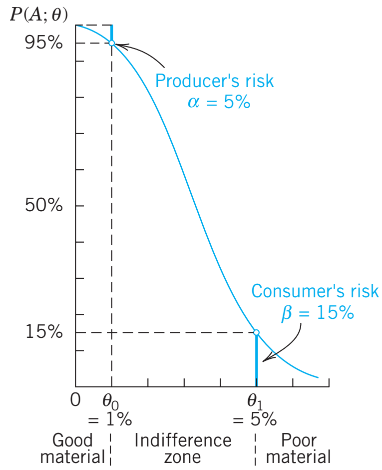


Fig. 540. OC curve, producer's and consumer's risks

an **acceptable lot** (a lot for which θ does not exceed a certain number θ_0 on which the two parties agree) to be small. θ_0 is called the **acceptable quality level** (AQL). Similarly, the consumer (the buyer) wants the probability β of accepting an **unacceptable lot** (a lot for which θ is greater than or equal to some θ_1) to be small. θ_1 is called the **lot tolerance percent defective** (LTPD) or the **rejectable quality level** (RQL). α is called **producer's risk**. It corresponds to a Type I error in Sec. 25.4. β is called **consumer's risk** and corresponds to a Type II error. Figure 540 shows an example. We see that the points $(\theta_0, 1 - \alpha)$ and (θ_1, β) lie on the OC curve. It can be shown that for large lots we can choose $\theta_0, \theta_1 (> \theta_0), \alpha, \beta$ and then determine n and c such that the OC curve runs very close to those prescribed points. Table 25.6 shows the analogy between acceptance sampling and hypothesis testing in Sec. 25.4.

Table 25.6 Acceptance Sampling and Hypothesis Testing

Acceptance Sampling	Hypothesis Testing
Acceptable quality level (AQL) $\theta = \theta_0$	Hypothesis $\theta = \theta_0$
Lot tolerance percent defectives (LTPD) $\theta = \theta_1$	Alternative $\theta = \theta_1$
Allowable number of defectives c	Critical value c
Producer's risk α of rejecting a lot with $\theta \leqq \theta_0$	Probability α of making a Type I error (significance level)
Consumer's risk β of accepting a lot with $\theta \geqq \theta_1$	Probability β of making a Type II error

Rectification

Rectification of a *rejected* lot means that the lot is inspected item by item and all defectives are removed and replaced by nondefective items. (This may be too expensive if the lot is cheap; in this case the lot may be sold at a cut-rate price or scrapped.) If a production turns out $100\theta\%$ defectives, then in K lots of size N each, $KN\theta$ of the KN items are defectives. Now $KP(A; \theta)$ of these lots are accepted. These contain $KPN\theta$ defectives, whereas the rejected and rectified lots contain no defectives, because of the rectification. Hence after the rectification the fraction defective in all K lots equals $KPN\theta/KN$. This is called the **average outgoing quality** (AOQ); thus

$$\text{AOQ}(\theta) = \theta P(A; \theta). \tag{4}$$

Figure 541 shows an example. Since $\text{AOQ}(0) = 0$ and $P(A; 1) = 0$, the AOQ curve has a maximum at some $\theta = \theta^*$, giving the **average outgoing quality limit** (AOQL). This is the worst average quality that may be expected to be accepted under rectification.

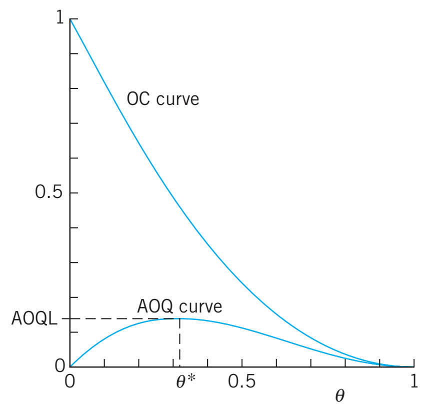


Fig. 541. OC curve and AOQ curve for the sampling plan in Fig. 538

PROBLEM SET 25.6

1. Lots of kitchen knives are inspected by a sampling plan that uses a sample of size 20 and the acceptance number $c = 1$. What is the probability of accepting a lot with $1\%, 2\%, 10\%$ defectives (knives with dull blades)? Use Table A6 of the Poisson distribution in App. 5. Graph the OC curve.

2. What happens in Prob. 1 if the sample size is increased to 50? First guess. Then calculate. Graph the OC curve and compare.

3. How will the probabilities in Prob. 1 with $n = 20$ change (up or down) if we decrease c to zero? First guess.

4. What are the producer's and consumer's risks in Prob. 1 if the AQL is 2% and the RQL is 15%?

5. Lots of copper pipes are inspected according to a sample plan that uses sample size 25 and acceptance number 1. Graph the OC curve of the plan, using the Poisson approximation. Find the producer's risk if the AQL is 1.5%.

6. Graph the AOQ curve in Prob. 5. Determine the AOQL, assuming that rectification is applied.

7. In Example 1 in the text, what are the producer's and consumer's risks if the AQL is 0.1 and the RQL is 0.6?

8. What happens in Example 1 in the text if we increase the sample size to $n = 3$, leaving the other data as before? Compute $P(A; 0.1)$ and $P(A; 0.2)$ and compare with Example 1.

9. Graph and compare sampling plans with $c = 1$ and increasing values of n, say, $n = 2, 3, 4$. (Use the binomial distribution.)

10. Find the binomial approximation of the hypergeometric distribution in Example 1 in the text and compare the approximate and the accurate values.

11. Samples of 3 fuses are drawn from lots and a lot is accepted if in the corresponding sample we find no more than 1 defective fuse. Criticize this sampling plan. In particular, find the probability of accepting a lot that is 50% defective. (Use the binomial distribution (7), Sec. 24.7.)

12. If in a sampling plan for large lots of spark plugs, the sample size is 100 and we want the AQL to be 5% and the producer's risk 2%, what acceptance number c should we choose? (Use the normal approximation of the binomial distribution in Sec. 24.8.)

13. What is the consumer's risk in Prob. 12 if we want the RQL to be 12%? Use $c = 9$ from the answer of Prob. 12.

14. A lot of batteries for wrist watches is accepted if and only if a sample of 20 contains at most 1 defective. Graph the OC and AOQ curves. Find AOQL. [Use (3).]

15. Graph the OC curve and the AOQ curve for the single sampling plan for large lots with $n = 5$ and $c = 0$, and find the AOQL.

25.7 Goodness of Fit. χ^2-Test

To test for **goodness of fit** means that we wish to test that a certain function $F(x)$ is the distribution function of a distribution from which we have a sample $x_1, \cdots, x_n$. Then we test whether the **sample distribution function** $\tilde{F}(x)$ defined by

$$\tilde{F}(x) = \textit{Sum of the relative frequencies of all sample values } x_j \textit{ not exceeding } x$$

fits $F(x)$ "sufficiently well." If this is so, we shall accept the hypothesis that $F(x)$ is the distribution function of the population; if not, we shall reject the hypothesis.

This test is of considerable practical importance, and it differs in character from the tests for parameters (μ, σ^2, etc.) considered so far.

To test in that fashion, we have to know how much $\tilde{F}(x)$ can differ from $F(x)$ if the hypothesis is true. Hence we must first introduce a quantity that measures the deviation of $\tilde{F}(x)$ from $F(x)$, and we must know the probability distribution of this quantity under the assumption that the hypothesis is true. Then we proceed as follows. We determine a number c such that, if the hypothesis is true, a deviation greater than c has a small preassigned probability. If, nevertheless, a deviation greater than c occurs, we have reason to doubt that the hypothesis is true and we reject it. On the other hand, if the deviation does not exceed c, so that $\tilde{F}(x)$ approximates $F(x)$ sufficiently well, we accept the hypothesis. Of course, if we accept the hypothesis, this means that we have insufficient evidence to reject it, and this does not exclude the possibility that there are other functions that would not be rejected in the test. In this respect the situation is quite similar to that in Sec. 25.4.

Table 25.7 shows a test of that type, which was introduced by R. A. Fisher. This test is justified by the fact that if the hypothesis is true, then χ_0^2 is an observed value of a random variable whose distribution function approaches that of the chi-square distribution with $K - 1$ degrees of freedom (or $K - r - 1$ degrees of freedom if r parameters are estimated) as n approaches infinity. The requirement that at least five sample values lie in each interval in Table 25.7 results from the fact that for finite n that random variable has only *approximately* a chi-square distribution. A proof can be found in Ref. [G3] listed in App. 1. If the sample is so small that the requirement cannot be satisfied, one may continue with the test, but then use the result with caution.

Table 25.7 Chi-square Test for the Hypothesis That $F(x)$ is the Distribution Function of a Population from Which a Sample $x_1, \cdots, x_n$ is Taken

Step 1. Subdivide the x-axis into K intervals $I_1, I_2, \cdots, I_K$ such that each interval contains at least 5 values of the given sample $x_1, \cdots, x_n$. Determine the number b_j of sample values in the interval I_j, where $j = 1, \cdots, K$. If a sample value lies at a common boundary point of two intervals, add 0.5 to each of the two corresponding b_j.

Step 2. Using $F(x)$, compute the probability p_j that the random variable X under consideration assumes any value in the interval I_j, where $j = 1, \cdots, K$. Compute

$$e_j = np_j.$$

(This is the number of sample values theoretically expected in I_j if the hypothesis is true.)

Step 3. Compute the deviation

$$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}. \tag{1}$$

Step 4. Choose a significance level (5%, 1%, or the like).

Step 5. Determine the solution c of the equation

$$P(\chi^2 \leqq c) = 1 - \alpha$$

from the table of the chi-sqare distribution with $K - 1$ degrees of freedom (Table A10 in App. 5). If r parameters of $F(x)$ are unknown and their maximum likelihood estimates (Sec. 25.2) are used, then use $K - r - 1$ degrees of freedom (instead of $K - 1$). If $\chi_0^2 \leqq c$, accept the hypothesis. If $\chi_0^2 > c$, reject the hypothesis.

Table 25.8 Sample of 100 Values of the Splitting Tensile Strength (lb/in.^{2}) of Concrete Cylinders

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

D. L. IVEY, Splitting tensile tests on structural lightweight aggregate concrete. Texas Transportation Institute, College Station, Texas.

EXAMPLE 1 Test of Normality

Test whether the population from which the sample in Table 25.8 was taken is normal.

Solution. Table 25.8 shows the values (column by column) in the order obtained in the experiment. Table 25.9 gives the frequency distribution and Fig. 542 the histogram. It is hard to guess the outcome of the test—does the histogram resemble a normal density curve sufficiently well or not?

The maximum likelihood estimates for μ and σ^2 are $\hat{\mu} = \bar{x} = 364.7$ and $\tilde{\sigma}^2 = 712.9$. The computation in Table 25.10 yields $\chi_0^2 = 2.688$. It is very interesting that the interval $375 \cdots 385$ contributes over 50% of χ_0^2. From the histogram we see that the corresponding frequency looks much too small. The second largest contribution comes from $395 \cdots 405$, and the histogram shows that the frequency seems somewhat too large, which is perhaps not obvious from inspection.

Table 25.9 Frequency Table of the Sample in Table 25.8

1 Tensile Strength x [lb/in.^{2}]	2 Absolute Frequency	3 Relative Frequency $\tilde{f}(x)$	4 Cumulative Absolute Frequency	5 Cumulative Relative Frequency $\tilde{F}(x)$
300	2	0.02	2	0.02
310	0	0.00	2	0.02
320	4	0.04	6	0.06
330	6	0.06	12	0.12
340	11	0.11	23	0.23
350	14	0.14	37	0.37
360	16	0.16	53	0.53
370	15	0.15	68	0.68
380	8	0.08	76	0.76
390	10	0.10	86	0.86
400	8	0.08	94	0.94
410	2	0.02	96	0.96
420	3	0.03	99	0.99
430	0	0.00	99	0.99
440	1	0.01	100	1.00

We choose $\alpha = 5\%$. Since $K = 10$ and we estimated $r = 2$ parameters we have to use Table A10 in App. 5 with $K - r - 1 = 7$ degrees of freedom. We find $c = 14.07$ as the solution of $P(\chi^2 \leqq c) = 95\%$. Since $\chi_0^2 < c$, we accept the hypothesis that the population is normal.

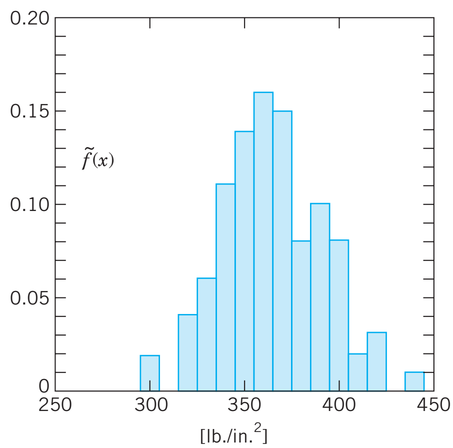


Fig. 542. Frequency histogram of the sample in Table 25.8

Table 25.10 Computations in Example 1

x_j	$\dfrac{x_j - 364.7}{26.7}$	$\Phi\left(\dfrac{x_j - 364.7}{26.7}\right)$	e_j	b_j	Term in (1)
$-\infty \cdots 325$	$-\infty \cdots -1.49$	$0.0000 \cdots 0.0681$	6.81	6	0.096
$325 \cdots 335$	$-1.49 \cdots -1.11$	$0.0681 \cdots 0.1335$	6.54	6	0.045
$335 \cdots 345$	$-1.11 \cdots -0.74$	$0.1335 \cdots 0.2296$	9.61	11	0.201
$345 \cdots 355$	$-0.74 \cdots -0.36$	$0.2296 \cdots 0.3594$	12.98	14	0.080
$355 \cdots 365$	$-0.36 \cdots 0.01$	$0.3594 \cdots 0.5040$	14.46	16	0.164
$365 \cdots 375$	$0.01 \cdots 0.39$	$0.5040 \cdots 0.6517$	14.77	15	0.0004
$375 \cdots 385$	$0.39 \cdots 0.76$	$0.6517 \cdots 0.7764$	12.47	8	1.602
$385 \cdots 395$	$0.76 \cdots 1.13$	$0.7764 \cdots 0.8708$	9.44	10	0.033
$395 \cdots 405$	$1.13 \cdots 1.51$	$0.8708 \cdots 0.9345$	6.37	8	0.417
$405 \cdots \infty$	$1.51 \cdots \infty$	$0.9345 \cdots 1.0000$	6.55	6	0.046

$\chi_0^2 = 2.688$

PROBLEM SET 25.7

1. Verify the calculations in Example 1 of the text.

2. If it is known that 25% of certain steel rods produced by a standard process will break when subjected to a load of 5000 lb, can we claim that a new, less expensive process yields the same breakage rate if we find that in a sample of 80 rods produced by the new process, 27 rods broke when subjected to that load? (Use $\alpha = 5\%$.)

3. If 100 flips of a coin result in 40 heads and 60 tails, can we assert on the 5% level that the coin is fair?

4. If in 10 flips of a coin we get the same ratio as in Prob. 3 (4 heads and 6 tails), is the conclusion the same as in Prob. 3? First conjecture, then compute.

5. Can you claim, on a 5% level, that a die is fair if 60 trials give $1, \cdots, 6$ with absolute frequencies 10, 13, 9, 11, 9, 8?

6. Solve Prob. 5 if rolling a die 180 times gives 33, 27, 29, 35, 25, 31.

7. If a service station had served 60, 49, 56, 46, 68, 39 cars from Monday through Friday between 1 P.M. and 2 P.M., can one claim on a 5% level that the differences are due to randomness? First guess. Then calculate.

8. A manufacturer claims that in a process of producing drill bits, only 2.5% of the bits are dull. Test the claim against the alternative that more than 2.5% of the bits are dull, using a sample of 400 bits containing 17 dull ones. Use $\alpha = 5\%$.

9. In a table of properly rounded function values, even and odd last decimals should appear about equally often. Test this for the 90 values of $J_1(x)$ in Table A1 in App. 5.

10. TEAM PROJECT. Difficulty with Random Selection. 77 students were asked to choose 3 of the integers $11, 12, 13, \cdots, 30$ completely arbitrarily. The amazing result was as follows.

Number	11	12	13	14	15	16	17	18	19	20
Frequ.	11	10	20	8	13	9	21	9	16	8

Number	21	22	23	24	25	26	27	28	29	30
Frequ.	12	8	15	10	10	9	12	8	13	9

If the selection were completely random, the following hypotheses should be true.

(a) The 20 numbers are equally likely.

(b) The 10 even numbers together are as likely as the 10 odd numbers together.

(c) The 6 prime numbers together have probability 0.3 and the 14 other numbers together have probability 0.7.

Test these hypotheses, using $\alpha = 5\%$. Design further experiments that illustrate the difficulties of random selection.

11. CAS EXPERIMENT. Random Number Generator. Check your generator experimentally by imitating results of n trials of rolling a fair die, with a convenient n (e.g., 60 or 300 or the like). Do this many times and see whether you can notice any "nonrandomness" features, for example, too few Sixes, too many even numbers, etc., or whether your generator seems to work properly. Design and perform other kinds of checks.

12. Test for normality at the 1% level using a sample of $n = 79$ (rounded) values x (tensile strength [kg/mm^2]

of steel sheets of 0.3 mm thickness). $a = a(x) =$ absolute frequency. (Take the first two values together, also the last three, to get $K = 5$.)

x	57	58	59	60	61	62	63	64
a	4	10	17	27	8	9	3	1

13. Mendel's pathbreaking experiments. In a famous plant-crossing experiment, the Austrian Augustinian father Gregor Mendel (1822–1884) obtained 355 yellow and 123 green peas. Test whether this agrees with Mendel's theory according to which the ratio should be 3:1.

14. Accidents in a foundry. Does the random variable $X =$ *Number of accidents per week* have a Poisson distribution if, within 50 weeks, 33 were accident-free, 1 accident occurred in 11 of the 50 weeks, 2 in 6 of the weeks, and more than 2 accidents in no week? Choose $\alpha = 5\%$.

15. Radioactivity. Rutherford-Geiger experiments. Using the given sample, test that the corresponding population has a Poisson distribution. x is the number of alpha particles per 7.5-s intervals observed by E. Rutherford and H. Geiger in one of their classical experiments in 1910, and $a(x)$ is the absolute frequency (= number of time periods during which exactly x particles were observed). Use $\alpha = 5\%$.

x	0	1	2	3	4	5	6
a	57	203	383	525	532	408	273

x	7	8	9	10	11	12	$\geqq 13$
a	139	45	27	10	4	2	0

25.8 Nonparametric Tests

Nonparametric tests, also called **distribution-free tests**, are valid for any distribution. Hence they are used in cases when the kind of distribution is unknown, or is known but such that no tests specifically designed for it are available. In this section we shall explain the basic idea of these tests, which are based on "**order statistics**" and are rather simple. If there is a choice, then tests designed for a specific distribution generally give better results than do nonparametric tests. For instance, this applies to the tests in Sec. 25.4 for the normal distribution.

We shall discuss two tests in terms of typical examples. In deriving the distributions used in the test, it is essential that the distributions, from which we sample, are continuous. (Nonparametric tests can also be derived for discrete distributions, but this is slightly more complicated.)

EXAMPLE 1 **Sign Test for the Median**

A **median** of the population is a solution $x = \tilde{\mu}$ of the equation $F(x) = 0.5$, where F is the distribution function of the population.

Suppose that eight radio operators were tested, first in rooms without air-conditioning and then in air-conditioned rooms over the same period of time, and the difference of errors (unconditioned minus conditioned) were

$$9 \quad 4 \quad 0 \quad 6 \quad 4 \quad 0 \quad 7 \quad 11.$$

Test the hypothesis $\tilde{\mu} = 0$ (that is, air-conditioning has no effect) against the alternative $\bar{\mu} > 0$ (that is, inferior performance in unconditioned rooms).

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, the probability p of a positive difference is the same as that of a negative difference. Hence in this case, $p = 0.5$, and the random variable

$$X = \textit{Number of positive values among n values}$$

has a binomial distribution with $p = 0.5$. Our sample has eight values. We omit the values 0, which do not contribute to the decision. Then six values are left, all of which are positive. Since

$$\begin{aligned} P(X = 6) &= \binom{6}{6}(0.5)^6(0.5)^0 \\ &= 0.0156 \\ &= 1.56\% \end{aligned}$$

we have observed an event whose probability is very small if the hypothesis is true; in fact $1.56\% < \alpha = 5\%$. Hence we assert that the alternative $\tilde{\mu} > 0$ is true. That is, the number of errors made in unconditioned rooms is significantly higher, so that installation of air conditioning should be considered. ■

EXAMPLE 2 Test for Arbitrary Trend

A certain machine is used for cutting lengths of wire. Five successive pieces had the lengths

$$29 \quad 31 \quad 28 \quad 30 \quad 32.$$

Using this sample, test the hypothesis that there is **no trend**, that is, the machine does not have the tendency to produce longer and longer pieces or shorter and shorter pieces. Assume that the type of machine suggests the alternative that there is *positive trend,* that is, there is the tendency of successive pieces to get longer.

Solution. We count the number of **transpositions** in the sample, that is, the number of times a larger value precedes a smaller value:

29 precedes 28 (1 transposition),

31 precedes 28 and 30 (2 transpositions).

The remaining three sample values follow in ascending order. Hence in the sample there are $1 + 2 = 3$ transpositions. We now consider the random variable

$$T = \textit{Number of transpositions.}$$

If the hypothesis is true (no trend), then each of the $5! = 120$ permutations of five elements 1 2 3 4 5 has the same probability (1/120). We arrange these permutations according to their number of transpositions:

$T = 0$	$T = 1$	$T = 2$	$T = 3$	
1 2 3 4 5	1 2 3 5 4	1 2 4 5 3	1 2 5 4 3	
	1 2 4 3 5	1 2 5 3 4	1 3 4 5 2	
	1 3 2 4 5	1 3 2 5 4	1 3 5 2 4	
	2 1 3 4 5	1 3 4 2 5	1 4 2 5 3	
		1 4 2 3 5	1 4 3 2 5	
		2 1 3 5 4	1 5 2 3 4	
		2 1 4 3 5	2 1 4 5 3	
		2 3 1 4 5	2 1 5 3 4	etc.
		3 1 2 4 5	2 3 1 5 4	
			2 3 4 1 5	
			2 4 1 3 5	
			3 1 2 5 4	
			3 1 4 2 5	
			3 2 1 4 5	
			4 1 2 3 5	

From this we obtain

$$P(T \leqq 3) = \tfrac{1}{120} + \tfrac{4}{120} + \tfrac{9}{120} + \tfrac{15}{120} = \tfrac{29}{120} = 24\%.$$

We accept the hypothesis because we have observed an event that has a relatively large probability (certainly much more than 5%) if the hypothesis is true.

Values of the distribution function of T in the case of no trend are shown in Table A12, App. 5. For instance, if $n = 3$, then $F(0) = 0.167$, $F(1) = 0.500$, $F(2) = 1 - 0.167$. If $n = 4$, then $F(0) = 0.042$, $F(1) = 0.167$, $F(2) = 0.375$, $F(3) = 1 - 0.375$, $F(4) = 1 - 0.167$, and so on.

Our method and those values refer to *continuous* distributions. Theoretically, we may then expect that all the values of a sample are different. Practically, some sample values may still be equal, because of rounding: If m values are equal, add $m(m-1)/4$ (= mean value of the transpositions in the case of the permutations of m elements), that is, $\frac{1}{2}$ for each pair of equal values, $\frac{3}{2}$ for each triple, etc. ∎

PROBLEM SET 25.8

1. What would change in Example 1 had we observed only 5 positive values? Only 4?

2. Test $\tilde{\mu} = 0$ against $\tilde{\mu} > 0$, using $1, -1, 1, 3, -8, 6, 0$ (deviations of the azimuth [multiples of 0.01 radian] in some revolution of a satellite).

3. Are oil filters of type A better than type B filters if in 11 trials, A gave cleaner oil than B in 7 cases, B gave cleaner oil than A in 1 case, whereas in 3 of the trials the results for A and B were practically the same?

4. Does a process of producing stainless steel pipes of length 20 ft for nuclear reactors need adjustment if, in a sample, 4 pipes have the exact length and 15 are shorter and 3 longer than 20 ft? Use the normal approximation of the binomial distribution.

5. Do the computations in Prob. 4 without the use of the DeMoivre–Laplace limit theorem in Sec. 24.8.

6. Thirty new employees were grouped into 15 pairs of similar intelligence and experience and were then instructed in data processing by an old method (A) applied to one (randomly selected) person of each pair, and by a new presumably better method (B) applied to the other person of each pair. Test for equality of methods against the alternative that (B) is better than (A), using the following scores obtained after the end of the training period.

A	60	70	80	85	75	40	70	45	95	80	90	60	80	75	65
B	65	85	85	80	95	65	100	60	90	85	100	75	90	60	80

7. Assuming normality, solve Prob. 6 by a suitable test from Sec. 25.4.

8. In a clinical experiment, each of 10 patients were given two different sedatives A and B. The following table shows the effect (increase of sleeping time, measured in hours). Using the sign test, find out whether the difference is significant.

A	1.9	0.8	1.1	0.1	−0.1	4.4	5.5	1.6	4.6	3.4
B	0.7	−1.6	−0.2	−1.2	−0.1	3.4	3.7	0.8	0.0	2.0
Difference	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

9. Assuming that the populations corresponding to the samples in Prob. 8 are normal, apply a suitable test for the normal distribution.

10. Test whether a thermostatic switch is properly set to 50°C against the alternative that its setting is too low. Use a sample of 9 values, 8 of which are less than 50°C and 1 is greater.

11. How would you proceed in the sign test if the hypothesis is $\tilde{\mu} = \tilde{\mu}_0$ (any number) instead of $\tilde{\mu} = 0$?

12. Test the hypothesis that, for a certain type of voltmeter, readings are independent of temperature T [°C] against the alternative that they tend to increase with T. Use a sample of values obtained by applying a constant voltage:

Temperature T [°C]	10	20	30	40	50
Reading V [volts]	99.5	101.1	100.4	100.8	101.6

13. Does the amount of fertilizer increase the yield of wheat X [kg/plot]? Use a sample of values ordered according to increasing amounts of fertilizer:

33.4 35.3 31.6 35.0 36.1 37.6 36.5 38.7.

14. Apply the test explained in Example 2 to the following data (x = diastolic blood pressure [mm Hg], y = weight of heart [in grams] of 10 patients who died of cerebral hemorrhage).

x	121	120	95	123	140	112	92	100	102	91
y	521	465	352	455	490	388	301	395	375	418

15. Does an increase in temperature cause an increase of the yield of a chemical reaction from which the following sample was taken?

Temperature [°C]	10	20	30	40	60	80
Yield [kg/min]	0.6	1.1	0.9	1.6	1.2	2.0

25.9 Regression. Fitting Straight Lines. Correlation

So far we were concerned with random experiments in which we observed a single quantity (random variable) and got samples whose values were single numbers. In this section we discuss experiments in which we observe or measure two quantities simultaneously, so that we get samples of *pairs* of values $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Most applications involve one of two kinds of experiments, as follows.

1. In **regression analysis** one of the two variables, call it x, can be regarded as an ordinary variable because we can measure it without substantial error or we can even give it values we want. x is called the **independent variable**, or sometimes the **controlled variable** because we can control it (set it at values we choose). The other variable, Y, is a random variable, and we are interested in the dependence of Y on x. Typical examples are the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x, the regression of the gain of weight Y of certain animals on the daily ration of food x, the regression of the heat conductivity Y of cork on the specific weight x of the cork, etc.
2. In **correlation analysis** both quantities are random variables and we are interested in relations between them. Examples are the relation (one says "correlation") between wear X and wear Y of the front tires of cars, between grades X and Y of students in mathematics and in physics, respectively, between the hardness X of steel plates in the center and the hardness Y near the edges of the plates, etc.

Regression Analysis

In regression analysis the dependence of Y on x is a dependence of the mean μ of Y on x, so that $\mu = \mu(x)$ is a function in the ordinary sense. The curve of $\mu(x)$ is called the **regression curve** of Y on x.

In this section we discuss the simplest case, namely, that of a straight **regression line**

$$\mu(x) = \kappa_0 + \kappa_1 x. \tag{1}$$

Then we may want to graph the sample values as n points in the xY-plane, fit a straight line through them, and use it for estimating $\mu(x)$ at values of x that interest us, so that we know what values of Y we can expect for those x. Fitting that line by eye would not be good because it would be subjective; that is, different persons' results would come out differently, particularly if the points are scattered. So we need a mathematical method that gives a unique result depending only on the n points. A widely used procedure is the method of least squares by Gauss and Legendre. For our task we may formulate it as follows.

Least Squares Principle

The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction). (Formulas below.)

To get uniqueness of the straight line, we need some extra condition. To see this, take the sample (0, 1), (0, −1). Then all the lines $y = k_1 x$ with any k_1 satisfy the principle. (Can you see it?) The following assumption will imply uniqueness, as we shall find out.

General Assumption (A1)

The x-values $x_1, \cdots, x_n$ in our sample $(x_1, y_1), \cdots, (x_n, y_n)$ are not all equal.

From a given sample $(x_1, y_1), \cdots, (x_n, y_n)$ we shall now determine a straight line by least squares. We write the line as

$$y = k_0 + k_1 x \tag{2}$$

and call it the **sample regression line** because it will be the counterpart of the population regression line (1).

Now a sample point (x_j, y_j) has the vertical distance (distance measured in the y-direction) from (2) given by

$$|y_j - (k_0 + k_1 x_j)| \qquad \text{(see Fig. 543).}$$

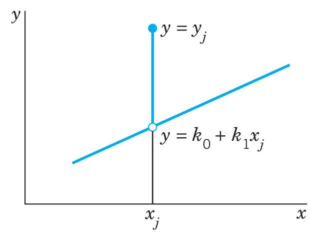


Fig. 543. Vertical distance of a point (x_j, y_j) from a straight line $y = k_0 + k_1 x$

Hence the sum of the squares of these distances is

$$q = \sum_{j=1}^{n} (y_j - k_0 - k_1 x_j)^2. \tag{3}$$

In the method of least squares we now have to determine k_0 and k_1 such that q is minimum. From calculus we know that a necessary condition for this is

$$\frac{\partial q}{\partial k_0} = 0 \qquad \text{and} \qquad \frac{\partial q}{\partial k_1} = 0. \tag{4}$$

We shall see that from this condition we obtain for the sample regression line the formula

$$y - \bar{y} = k_1 (x - \bar{x}). \tag{5}$$

Here $\bar{x}$ and $\bar{y}$ are the means of the x- and the y-values in our sample, that is,

$$\text{(6)} \qquad \begin{aligned} &\text{(a)} \quad \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \\ &\text{(b)} \quad \bar{y} = \frac{1}{n}(y_1 + \cdots + y_n). \end{aligned}$$

The slope k_1 in (5) is called the **regression coefficient** of the sample and is given by

$$\textbf{(7)} \qquad \boxed{k_1 = \frac{s_{xy}}{s_x^2}.}$$

Here the "**sample covariance**" s_{xy} is

$$\text{(8)} \quad s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)\right]$$

and s_x^2 is given by

$$\text{(9a)} \qquad s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right].$$

From (5) we see that the sample regression line passes through the point $(\bar{x}, \bar{y})$, by which it is determined, together with the regression coefficient (7). We may call s_x^2 the *variance* of the x-values, but we should keep in mind that x is an ordinary variable, not a random variable.

We shall soon also need

$$\text{(9b)} \qquad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right].$$

Derivation of (5) and (7). Differentiating (3) and using (4), we first obtain

$$\begin{aligned} \frac{\partial q}{\partial k_0} &= -2\sum(y_j - k_0 - k_1 x_j) = 0, \\ \frac{\partial q}{\partial k_1} &= -2\sum x_j(y_j - k_0 - k_1 x_j) = 0 \end{aligned}$$

where we sum over j from 1 to n. We now divide by 2, write each of the two sums as three sums, and take the sums containing y_j and $x_j y_j$ over to the right. Then we get the "**normal equations**"

$$\textbf{(10)} \qquad \boxed{\begin{aligned} k_0 n + k_1\sum x_j &= \sum y_j \\ k_0\sum x_j + k_1\sum x_j^2 &= \sum x_j y_j. \end{aligned}}$$

This is a linear system of two equations in the two unknowns k_0 and k_1. Its coefficient determinant is [see (9)]

$$\begin{vmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{vmatrix} = n\sum x_j^2 - \left(\sum x_j\right)^2 = n(n-1)s_x^2 = n\sum(x_j - \bar{x})^2$$

and is not zero because of Assumption (A1). Hence the system has a unique solution. Dividing the first equation of (10) by n and using (6), we get $k_0 = \bar{y} - k_1\bar{x}$. Together with $y = k_0 + k_1 x$ in (2) this gives (5). To get (7), we solve the system (10) by Cramer's rule (Sec. 7.6) or elimination, finding

$$k_1 = \frac{n\sum x_j y_j - \sum x_i \sum y_j}{n(n-1)s_x^2}. \tag{11}$$

This gives (7)–(9) and completes the derivation. [The equality of the two expressions in (8) and in (9) may be shown by the student]. ■

EXAMPLE 1 Regression Line

The decrease of volume y [%] of leather for certain fixed values of high pressure x [atmospheres] was measured. The results are shown in the first two columns of Table 25.11. Find the regression line of y on x.

Solution. We see that $n = 4$ and obtain the values $\bar{x} = 28000/4 = 7000$, $\bar{y} = 19.0/4 = 4.75$, and from (9) and (8)

Table 25.11 Regression of the Decrease of Volume y [%] of Leather on the Pressure x [Atmospheres]

Given Values		Auxiliary Values	
x_j	y_j	x_j^2	$x_j y_j$
4000	2.3	16,000,000	9200
6000	4.1	36,000,000	24,600
8000	5.7	64,000,000	45,600
10,000	6.9	100,000,000	69,000
28,000	19.0	216,000,000	148,400

$$s_x^2 = \frac{1}{3}\left(216{,}000{,}000 - \frac{28{,}000^2}{4}\right) = \frac{20{,}000{,}000}{3}$$

$$s_{xy} = \frac{1}{3}\left(148{,}400 - \frac{28{,}000 \cdot 19}{4}\right) = \frac{15{,}400}{3}.$$

Hence $k_1 = 15{,}400/20{,}000{,}000 = 0.00077$ from (7), and the regression line is

$$y - 4.75 = 0.00077(x - 7000) \qquad \text{or} \qquad y = 0.00077x - 0.64.$$

Note that $y(0) = -0.64$, which is physically meaningless, but typically indicates that a linear relation is merely an approximation valid on some restricted interval. ■

Confidence Intervals in Regression Analysis

If we want to get confidence intervals, we have to make assumptions about the distribution of Y (which we have not made so far; least squares is a "geometric principle," nowhere involving probabilities!). We assume normality and independence in sampling:

Assumption (A2)

For each fixed x the random variable Y is normal with mean (1), *that is,*

$$\mu(x) = \kappa_0 + \kappa_1 x \tag{12}$$

and variance σ^2 independent of x.

Assumption (A3)

The n performances of the experiment by which we obtain a sample

$$(x_1, y_1), \quad (x_2, y_2), \quad \cdots, \quad (x_n, y_n)$$

are independent.

κ_1 in (12) is called the **regression coefficient** of the population because it can be shown that, under Assumptions (A1)–(A3), the maximum likelihood estimate of κ_1 is the sample regression coefficient k_1 given by (11).

Under Assumptions (A1)–(A3), we may now obtain a confidence interval for κ_1, as shown in Table 25.12.

Table 25.12 Determination of a Confidence Interval for κ_1 in (1) under Assumptions (A1)–(A3)

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{13}$$

from the table of the t-distribution with $n - 2$ degrees of freedom (Table A9 in App. 5; n = sample size).

Step 3. Using a sample $(x_1, y_1), \cdots, (x_n, y_n)$, compute $(n - 1)s_x^2$ from (9a), $(n - 1)s_{xy}$ from (8), k_1 from (7),

$$(n - 1)s_y^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 \tag{14}$$

[as in (9b)], and

$$q_0 = (n - 1)(s_y^2 - k_1^2 s_x^2). \tag{15}$$

Step 4. Compute

$$K = c\sqrt{\frac{q_0}{(n - 2)(n - 1)s_x^2}}.$$

The confidence interval is

$$\mathrm{CONF}_\gamma\,\{k_1 - K \leqq \kappa_1 \leqq k_1 + K\}. \tag{16}$$

EXAMPLE 2 Confidence Interval for the Regression Coefficient

Using the sample in Table 25.11, determine a confidence interval for κ_1 by the method in Table 25.12.

Solution. ***Step 1.*** We choose $\gamma = 0.95$.

Step 2. Equation (13) takes the form $F(c) = 0.975$, and Table A9 in App. 5 with $n - 2 = 2$ degrees of freedom gives $c = 4.30$.

Step 3. From Example 1 we have $3s_x^2 = 20{,}000{,}000$ and $k_1 = 0.00077$. From Table 25.11 we compute

$$\begin{aligned} 3s_y^2 &= 102.0 - \frac{19^2}{4} \\ &= 11.95. \\ q_0 &= 11.95 - 20{,}000{,}000 \cdot 0.00077^2 \\ &= 0.092. \end{aligned}$$

Step 4. We thus obtain

$$\begin{aligned} K &= 4.30\sqrt{0.092/(2 \cdot 20{,}000{,}000)} \\ &= 0.000206 \end{aligned}$$

and

$$\mathrm{CONF}_{0.95}\,\{0.00056 \leqq \kappa_1 \leqq 0.00098\}.$$

Correlation Analysis

We shall now give an introduction to the basic facts in correlation analysis; for proofs see Ref. [G2] or [G8] in App. 1.

Correlation analysis is concerned with the relation between X and Y in a two-dimensional random variable (X, Y) (Sec. 24.9). A sample consists of n ordered pairs of values $(x_1, y_1), \cdots, (x_n, y_n)$, as before. The interrelation between the x and y values in the sample is measured by the sample covariance s_{xy} in (8) or by the sample **correlation coefficient**

$$r = \frac{s_{xy}}{s_x s_y} \tag{17}$$

with s_x and s_y given in (9). Here r has the advantage that it does not change under a multiplication of the x and y values by a factor (in going from feet to inches, etc.).

THEOREM 1 Sample Correlation Coefficient

The sample correlation coefficient r satisfies $-1 \leqq r \leqq 1$. In particular, $r = \pm 1$ if and only if the sample values lie on a straight line. (See Fig. 544.)

The theoretical counterpart of r is the **correlation coefficient** ρ of X and Y,

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{18}$$

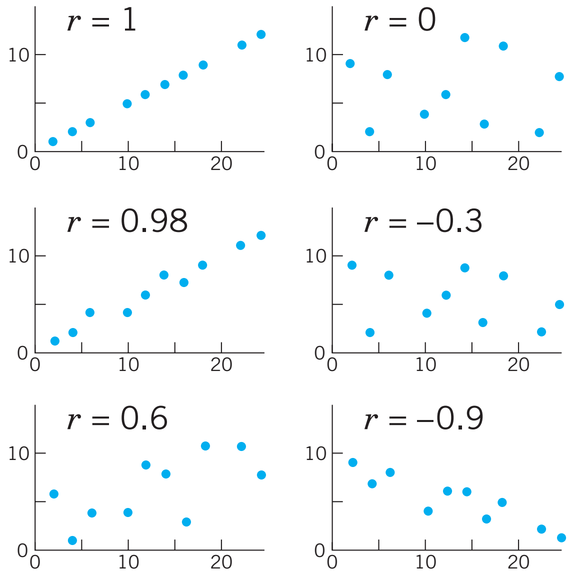


Fig. 544. Samples with various values of the correlation coefficient r

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = E([X - \mu_X]^2)$, $\sigma_Y^2 = E([Y - \mu_Y]^2)$ (the means and variances of the marginal distributions of X and Y; see Sec. 24.9), and σ_{XY} is the **covariance** of X and Y given by (see Sec. 24.9)

$$\sigma_{XY} = E([X - \mu_X][Y - \mu_Y]) = E(XY) - E(X)E(Y). \tag{19}$$

The analog of Theorem 1 is

THEOREM 2

Correlation Coefficient

The correlation coefficient ρ *satisfies* $-1 \leqq \rho \leqq 1$. *In particular,* $\rho = \pm 1$ *if and only if* X *and* Y *are* **linearly related**, *that is,* $Y = \gamma X + \delta$, $X = \gamma^* Y + \delta^*$.

X and Y are called **uncorrelated** if $\rho = 0$.

THEOREM 3

Independence. Normal Distribution

(a) *Independent* X *and* Y (see Sec. 24.9) *are uncorrelated.*

(b) *If* (X, Y) *is normal* (see below), *then uncorrelated* X *and* Y *are independent.*

Here the two-dimensional normal distribution can be introduced by taking two independent standardized normal random variables X^*, Y^*, whose joint distribution thus has the density

$$f^*(x^*, y^*) = \frac{1}{2\pi} e^{-(x^{*2} + y^{*2})/2} \tag{20}$$

(representing a surface of revolution over the x^*y^*-plane with a bell-shaped curve as cross section) and setting

$$X = \mu_X + \sigma_X X^*$$
$$Y = \mu_Y + \rho\sigma_Y X^* + \sqrt{1-\rho^2}\,\sigma_Y Y^*.$$

This gives the general **two-dimensional normal distribution** with the density

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-h(x,y)/2} \tag{21a}$$

where

$$h(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]. \tag{21b}$$

In Theorem 3(b), normality is important, as we can see from the following example.

EXAMPLE 3 Uncorrelated But Dependent Random Variables

If X assumes $-1, 0, 1$ with probability $\frac{1}{3}$ and $Y = X^2$, then $E(X) = 0$ and in (3)

$$\sigma_{XY} = E(XY) = E(X^3) = (-1)^3 \cdot \tfrac{1}{3} + 0^3 \cdot \tfrac{1}{3} + 1^3 \cdot \tfrac{1}{3} = 0,$$

so that $\rho = 0$ and X and Y are uncorrelated. But they are certainly not independent since they are even functionally related. ■

Test for the Correlation Coefficient ρ

Table 25.13 shows a test for ρ in the case of the two-dimensional normal distribution. t is an observed value of a random variable that has a t-distribution with $n - 2$ degrees of freedom. This was shown by R. A. Fisher (*Biometrika* **10** (1915), 507–521).

Table 25.13 Test of the Hypothesis $\rho = 0$ Against the Alternative $\rho > 0$ in the Case of the Two-Dimensional Normal Distribution

Step 1. Choose a significance level α (5%, 1%, or the like).

Step 2. Determine the solution c of the equation

$$P(T \leqq c) = 1 - \alpha$$

from the t-distribution (Table A9 in App. 5) with $n - 2$ degrees of freedom.

Step 3. Compute r from (17), using a sample $(x_1, y_1), \cdots, (x_n, y_n)$.

Step 4. Compute

$$t = r\left(\sqrt{\frac{n-2}{1-r^2}}\right).$$

If $t \leqq c$, accept the hypothesis. If $t > c$, reject the hypothesis.

EXAMPLE 4 Test for the Correlation Coefficient ρ

Test the hypothesis $\rho = 0$ (independence of X and Y, because of Theorem 3) against the alternative $\rho > 0$, using the data in the lower left corner of Fig. 544, where $r = 0.6$ (manual soldering errors on 10 two-sided circuit boards done by 10 workers; x = front, y = back of the boards).

Solution. We choose $\alpha = 5\%$; thus $1 - \alpha = 95\%$. Since $n = 10$, $n - 2 = 8$, the table gives $c = 1.86$. Also, $t = 0.6\sqrt{8/0.64} = 2.12 > c$. We reject the hypothesis and assert that there is a **positive correlation**. A worker making few (many) errors on the front side also tends to make few (many) errors on the reverse side of the board. ■

PROBLEM SET 25.9

1–10 SAMPLE REGRESSION LINE

Find and graph the sample regression line of y on x and the given data as points on the same axes. Show the details of your work.

1. (0, 1.0), (2, 2.1), (4, 2.9), (6, 3.6), (8, 5.2)
2. (−2, 3.5), (1, 2.6), (3, 1.3), (5, 0.4)
3. x = Revolutions per minute, y = Power of a Diesel engine [hp]

x	400	500	600	700	750
y	5800	10,300	14,200	18,800	21,000

4. x = Deformation of a certain steel [mm], y = Brinell hardness [kg/mm^2]

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

5. x = Brinell hardness, y = Tensile strength [in 1000 psi (pounds per square inch)] of steel with 0.45% C tempered for 1 hour

x	200	300	400	500
y	110	150	190	280

6. **Abrasion of quenched and tempered steel S620.** x = Sliding distance [km], y = Wear volume [mm^3]

x	1.1	3.2	3.4	4.5	5.6
y	40	65	120	150	190

7. **Ohm's law (Sec. 2.9).** x = Voltage [V], y = Current [A]. Also find the resistance R [Ω].

x	40	40	80	80	110	110
y	5.1	4.8	0.0	10.3	13.0	12.7

8. **Hooke's law** (Sec. 2.4). x = Force [lb], y = Extension [in] of a spring. Also find the spring modulus.

x	2	4	6	8
y	4.1	7.8	12.3	15.8

9. **Thermal conductivity of water.** x = Temperature [°F], y = Conductivity [Btu/(hr · ft · °F)]. Also find y at room temperature 66°F.

x	32	50	100	150	212
y	0.337	0.345	0.365	0.380	0.395

10. **Stopping distance of a car.** x = Speed [mph]. y = Stopping distance [ft]. Also find y at 35 mph.

x	30	40	50	60
y	160	240	330	435

11. **CAS EXPERIMENT. Moving Data.** Take a sample, for instance, that in Prob. 4, and investigate and graph the effect of changing y-values **(a)** for small x, **(b)** for large x, **(c)** in the middle of the sample.

12–15 CONFIDENCE INTERVALS

Find a 95% confidence interval for the regression coefficient κ_1, assuming (A2) and (A3) hold and using the sample.

12. In Prob. 2
13. In Prob. 3
14. In Prob. 4
15. x = Humidity of air [%], y = Expansion of gelatin [%],

x	10	20	30	40
y	0.8	1.6	2.3	2.8

CHAPTER 25 REVIEW QUESTIONS AND PROBLEMS

1. What is a sample? A population? Why do we sample in statistics?
2. If we have several samples from the same population, do they have the same sample distribution function? The same mean and variance?
3. Can we develop statistical methods without using probability theory? Apply the methods without using a sample?
4. What is the idea of the maximum likelihood method? Why do we say "likelihood" rather than "probability"?

5. Couldn't we make the error of interval estimation zero simply by choosing the confidence level 1?
6. What is testing? Why do we test? What are the errors involved?
7. When did we use the t-distribution? The F-distribution?
8. What is the chi-square (χ^2) test? Give a sample example from memory.
9. What are one-sided and two-sided tests? Give typical examples.
10. How do we test in quality control? In acceptance sampling?
11. What is the power of a test? What could you perhaps do when it is low?
12. What is Gauss's least squares principle (which he found at age 18)?
13. What is the difference between regression and correlation?
14. Find the mean, variance, and standard derivation of the sample 21.0 21.6 19.9 19.6 15.6 20.6 22.1 22.2.
15. Assuming normality, find the maximum likelihood estimates of mean and variance from the sample in Prob. 14.
16. Determine a 95% confidence interval for the mean μ of a normal population with variance $\sigma^2 = 25$, using a sample of size 500 with mean 22.
17. Determine a 99% confidence interval for the mean of a normal population, using the sample 32, 33, 32, 34, 35, 29, 29, 27.
18. Assuming normality, find a 95% confidence interval for the variance from the sample 145.3, 145.1, 145.4, 146.2.
19. Using a sample of 10 values with mean 14.5 from a normal population with variance $\sigma^2 = 0.25$, test the hypothesis $\mu_0 = 15.0$ against the alternative $\mu_1 = 14.5$ on the 5% level. Find the power.
20. Three specimens of high-quality concrete had compressive strength 357, 359, 413 [kg/cm^2], and for three specimens of ordinary concrete the values were 346, 358, 302. Test for equality of the population means, $\mu_1 = \mu_2$, against the alternative $\mu_1 > \mu_2$. Assume normality and equality of variance. Choose $\alpha = 5\%$.
21. Assume the thickness X of washers to be normal with mean 2.75 mm and variance 0.00024 mm^2. Set up a control chart for μ and graph the means of the five samples (2.74, 2.76), (2.74, 2.74), (2.79, 2.81), (2.78, 2.76), (2.71, 2.75) on the chart.
22. The OC curve in acceptance sampling cannot have a strictly vertical portion. Why?
23. Find the risks in the sampling plan with $n = 6$ and $c = 0$, assuming that the AQL is $\theta_0 = 1\%$ and the RQL is $\theta_1 = 15\%$. How do the risks change if we increase n?
24. Does a process of producing plastic rods of length $\tilde{\mu} = 2$ meters need adjustment if in a sample, 2 rods have the exact length and 15 are shorter and 3 longer than 2 meters? (Use the sign test.)
25. Find the regression line of y on x for the data $(x, y) = (0, 4), (2, 0), (4, -5), (6, -9), (8, -10)$.

SUMMARY OF CHAPTER 25
Mathematical Statistics

We recall from Chap. 24 that, with an experiment in which we observe some quantity (number of defectives, height of persons, etc.), there is associated a random variable X whose probability distribution is given by a distribution function

$$F(x) = P(X \leqq x) \qquad (1) \quad \text{(Sec. 24.5)}$$

which for each x gives the probability that X assumes any value not exceeding x.

In statistics we take random samples $x_1, \cdots, x_n$ of size n by performing that experiment n times (Sec. 25.1) and draw conclusions from properties of samples about properties of the distribution of the corresponding X. We do this by calculating *point estimates* or *confidence intervals* or by performing a *test* for **parameters** (μ and σ^2 in the normal distribution, p in the binomial distribution, etc.) or by a test for distribution functions.

A **point estimate** (Sec. 25.2) is an approximate value for a parameter in the distribution of X obtained from a sample. Notably, the **sample mean** (Sec. 25.1)

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + \cdots + x_n) \tag{2}$$

is an estimate of the mean μ of X, and the **sample variance** (Sec. 25.1)

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \tag{3}$$

is an estimate of the variance σ^2 of X. Point estimation can be done by the basic ***maximum likelihood method*** (Sec. 25.2).

Confidence intervals (Sec. 25.3) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ with endpoints calculated from a sample such that, with a high probability γ, we obtain an interval that contains the unknown true value of the parameter θ in the distribution of X. Here, γ is chosen at the beginning, usually 95% or 99%. We denote such an interval by $\text{CONF}_\gamma\{\theta_1 \leqq \theta \leqq \theta_2\}$.

In a **test** for a parameter we test a *hypothesis* $\theta = \theta_0$ against an *alternative* $\theta = \theta_1$ and then, on the basis of a sample, accept the hypothesis, or we reject it in favor of the alternative (Sec. 25.4). Like any conclusion about X from samples, this may involve errors leading to a false decision. There is a small probability α (which we can choose, 5% or 1%, for instance) that we reject a true hypothesis, and there is a probability β (which we can compute and decrease by taking larger samples) that we accept a false hypothesis. α is called the **significance level** and $1 - \beta$ the **power** of the test. Among many other engineering applications, testing is used in ***quality control*** (Sec. 25.5) and ***acceptance sampling*** (Sec. 25.6).

If not merely a parameter but the kind of distribution of X is unknown, we can use the **chi-square test** (Sec. 25.7) for testing the hypothesis that some function $F(x)$ is the unknown distribution function of X. This is done by determining the discrepancy between $F(x)$ and the distribution function $\tilde{F}(x)$ of a given sample.

"Distribution-free" or ***nonparametric tests*** are tests that apply to any distribution, since they are based on combinatorial ideas. These tests are usually very simple. Two of them are discussed in Sec. 25.8.

The last section deals with samples of ***pairs of values***, which arise in an experiment when we simultaneously observe two quantities. In ***regression analysis***, one of the quantities, x, is an ordinary variable and the other, Y, is a random variable whose mean μ depends on x, say, $\mu(x) = \kappa_0 + \kappa_1 x$. In ***correlation analysis*** the relation between X and Y in a two-dimensional random variable (X, Y) is investigated, notably in terms of the ***correlation coefficient*** ρ.

APPENDIX 1

References

Software *see* at the beginning of Chaps. 19 and 24.

General References

[GenRef1] Abramowitz, M. and I. A. Stegun (eds.), *Handbook of Mathematical Functions.* 10th printing, with corrections. Washington, DC: National Bureau of Standards. 1972 (also New York: Dover, 1965). See also [W1]

[GenRef2] Cajori, F., *History of Mathematics.* 5th ed. Reprinted. Providence, RI: American Mathematical Society, 2002.

[GenRef3] Courant, R. and D. Hilbert, *Methods of Mathematical Physics.* 2 vols. Hoboken, NJ: Wiley, 1989.

[GenRef4] Courant, R., *Differential and Integral Calculus.* 2 vols. Hoboken, NJ: Wiley, 1988.

[GenRef5] Graham, R. L. et al., *Concrete Mathematics.* 2nd ed. Reading, MA: Addison-Wesley, 1994.

[GenRef6] Ito, K. (ed.), *Encyclopedic Dictionary of Mathematics.* 4 vols. 2nd ed. Cambridge, MA: MIT Press, 1993.

[GenRef7] Kreyszig, E., *Introductory Functional Analysis with Applications.* New York: Wiley, 1989.

[GenRef8] Kreyszig, E., *Differential Geometry.* Mineola, NY: Dover, 1991.

[GenRef9] Kreyszig, E. *Introduction to Differential Geometry and Riemannian Geometry.* Toronto: University of Toronto Press, 1975.

[GenRef10] Szegö, G., *Orthogonal Polynomials.* 4th ed. Reprinted. New York: American Mathematical Society, 2003.

[GenRef11] Thomas, G. et al., *Thomas' Calculus, Early Transcendentals Update.* 10th ed. Reading, MA: Addison-Wesley, 2003.

Part A. Ordinary Differential Equations (ODEs) (Chaps. 1–6)

See also Part E: Numeric Analysis

[A1] Arnold, V. I., *Ordinary Differential Equations.* 3rd ed. New York: Springer, 2006.

[A2] Bhatia, N. P. and G. P. Szego, *Stability Theory of Dynamical Systems.* New York: Springer, 2002.

[A3] Birkhoff, G. and G.-C. Rota, *Ordinary Differential Equations.* 4th ed. New York: Wiley, 1989.

[A4] Brauer, F. and J. A. Nohel, *Qualitative Theory of Ordinary Differential Equations.* Mineola, NY: Dover, 1994.

[A5] Churchill, R. V., *Operational Mathematics.* 3rd ed. New York: McGraw-Hill, 1972.

[A6] Coddington, E. A. and R. Carlson, *Linear Ordinary Differential Equations.* Philadelphia: SIAM, 1997.

[A7] Coddington, E. A. and N. Levinson, *Theory of Ordinary Differential Equations.* Malabar, FL: Krieger, 1984.

[A8] Dong, T.-R. et al., *Qualitative Theory of Differential Equations.* Providence, RI: American Mathematical Society, 1992.

[A9] Erdélyi, A. et al., *Tables of Integral Transforms.* 2 vols. New York: McGraw-Hill, 1954.

[A10] Hartman, P., *Ordinary Differential Equations.* 2nd ed. Philadelphia: SIAM, 2002.

[A11] Ince, E. L., *Ordinary Differential Equations.* New York: Dover, 1956.

[A12] Schiff, J. L., *The Laplace Transform: Theory and Applications.* New York: Springer, 1999.

[A13] Watson, G. N., *A Treatise on the Theory of Bessel Functions.* 2nd ed. Reprinted. New York: Cambridge University Press, 1995.

[A14] Widder, D. V., *The Laplace Transform.* Princeton, NJ: Princeton University Press, 1941.

[A15] Zwillinger, D., *Handbook of Differential Equations.* 3rd ed. New York: Academic Press, 1998.

Part B. Linear Algebra, Vector Calculus (Chaps. 7–10)

For books on *numeric* linear algebra, *see also* Part E: Numeric Analysis.

[B1] Bellman, R., *Introduction to Matrix Analysis.* 2nd ed. Philadelphia: SIAM, 1997.

[B2] Chatelin, F., *Eigenvalues of Matrices.* New York: Wiley-Interscience, 1993.

[B3] Gantmacher, F. R., *The Theory of Matrices.* 2 vols. Providence, RI: American Mathematical Society, 2000.

[B4] Gohberg, I. P. et al., *Invariant Subspaces of Matrices with Applications.* New York: Wiley, 2006.

[B5] Greub, W. H., *Linear Algebra.* 4th ed. New York: Springer, 1975.

[B6] Herstein, I. N., *Abstract Algebra.* 3rd ed. New York: Wiley, 1996.

[B7] Joshi, A. W., *Matrices and Tensors in Physics.* 3rd ed. New York: Wiley, 1995.
[B8] Lang, S., *Linear Algebra.* 3rd ed. New York: Springer, 1996.
[B9] Nef, W., *Linear Algebra.* 2nd ed. New York: Dover, 1988.
[B10] Parlett, B., *The Symmetric Eigenvalue Problem.* Philadelphia: SIAM, 1998.

Part C. Fourier Analysis and PDEs (Chaps. 11–12)

For books on *numerics* for PDEs *see also* Part E: Numeric Analysis.

[C1] Antimirov, M. Ya., *Applied Integral Transforms.* Providence, RI: American Mathematical Society, 1993.
[C2] Bracewell, R., *The Fourier Transform and Its Applications.* 3rd ed. New York: McGraw-Hill, 2000.
[C3] Carslaw, H. S. and J. C. Jaeger, *Conduction of Heat in Solids.* 2nd ed. Reprinted. Oxford: Clarendon, 2000.
[C4] Churchill, R. V. and J. W. Brown, *Fourier Series and Boundary Value Problems.* 6th ed. New York: McGraw-Hill, 2006.
[C5] DuChateau, P. and D. Zachmann, *Applied Partial Differential Equations.* Mineola, NY: Dover, 2002.
[C6] Hanna, J. R. and J. H. Rowland, *Fourier Series, Transforms, and Boundary Value Problems.* 2nd ed. New York: Wiley, 2008.
[C7] Jerri, A. J., *The Gibbs Phenomenon in Fourier Analysis, Splines, and Wavelet Approximations.* Boston: Kluwer, 1998.
[C8] John, F., *Partial Differential Equations.* 4th edition New York: Springer, 1982.
[C9] Tolstov, G. P., *Fourier Series.* New York: Dover, 1976.
[C10] Widder, D. V., *The Heat Equation.* New York: Academic Press, 1975.
[C11] Zauderer, E., *Partial Differential Equations of Applied Mathematics.* 3rd ed. New York: Wiley, 2006.
[C12] Zygmund, A. and R. Fefferman, *Trigonometric Series.* 3rd ed. New York: Cambridge University Press, 2002.

Part D. Complex Analysis (Chaps. 13–18)

[D1] Ahlfors, L. V., *Complex Analysis.* 3rd ed. New York: McGraw-Hill, 1979.
[D2] Bieberbach, L., *Conformal Mapping.* Providence, RI: American Mathematical Society, 2000.
[D3] Henrici, P., *Applied and Computational Complex Analysis.* 3 vols. New York: Wiley, 1993.
[D4] Hille, E., *Analytic Function Theory.* 2 vols. 2nd ed. Providence, RI: American Mathematical Society, Reprint V1 1983, V2 2005.
[D5] Knopp, K., *Elements of the Theory of Functions.* New York: Dover, 1952.
[D6] Knopp, K., *Theory of Functions.* 2 parts. New York: Dover, Reprinted 1996.
[D7] Krantz, S. G., *Complex Analysis: The Geometric Viewpoint.* Washington, DC: The Mathematical Association of America, 1990.
[D8] Lang, S., *Complex Analysis.* 4th ed. New York: Springer, 1999.
[D9] Narasimhan, R., *Compact Riemann Surfaces.* New York: Springer, 1996.
[D10] Nehari, Z., *Conformal Mapping.* Mineola, NY: Dover, 1975.
[D11] Springer, G., *Introduction to Riemann Surfaces.* Providence, RI: American Mathematical Society, 2001.

Part E. Numeric Analysis (Chaps. 19–21)

[E1] Ames, W. F., *Numerical Methods for Partial Differential Equations.* 3rd ed. New York: Academic Press, 1992.
[E2] Anderson, E., et al., *LAPACK User's Guide.* 3rd ed. Philadelphia: SIAM, 1999.
[E3] Bank, R. E., *PLTMG. A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide 8.0.* Philadelphia: SIAM, 1998.
[E4] Constanda, C., *Solution Techniques for Elementary Partial Differential Equations.* Boca Raton, FL: CRC Press, 2002.
[E5] Dahlquist, G. and A. Björck, *Numerical Methods.* Mineola, NY: Dover, 2003.
[E6] DeBoor, C., *A Practical Guide to Splines.* Reprinted. New York: Springer, 2001.
[E7] Dongarra, J. J. et al., *LINPACK Users Guide.* Philadelphia: SIAM, 1979. (See also at the beginning of Chap. 19.)
[E8] Garbow, B. S. et al., *Matrix Eigensystem Routines: EISPACK Guide Extension.* Reprinted. New York: Springer, 1990.
[E9] Golub, G. H. and C. F. Van Loan, *Matrix Computations.* 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.
[E10] Higham, N. J., *Accuracy and Stability of Numerical Algorithms.* 2nd ed. Philadelphia: SIAM, 2002.
[E11] IMSL (International Mathematical and Statistical Libraries), *FORTRAN Numerical Library.* Houston, TX: Visual Numerics, 2002. (See also at the beginning of Chap. 19.)
[E12] IMSL, *IMSL for Java.* Houston, TX: Visual Numerics, 2002.
[E13] IMSL, *C Library.* Houston, TX: Visual Numerics, 2002.
[E14] Kelley, C. T., *Iterative Methods for Linear and Nonlinear Equations.* Philadelphia: SIAM, 1995.
[E15] Knabner, P. and L. Angerman, *Numerical Methods for Partial Differential Equations.* New York: Springer, 2003.

[E16] Knuth, D. E., *The Art of Computer Programming*. 3 vols. 3rd ed. Reading, MA: Addison-Wesley, 1997–2009.

[E17] Kreyszig, E., *Introductory Functional Analysis with Applications*. New York: Wiley, 1989.

[E18] Kreyszig, E., On methods of Fourier analysis in multigrid theory. *Lecture Notes in Pure and Applied Mathematics* 157. New York: Dekker, 1994, pp. 225–242.

[E19] Kreyszig, E., Basic ideas in modern numerical analysis and their origins. *Proceedings of the Annual Conference of the Canadian Society for the History and Philosophy of Mathematics*. 1997, pp. 34–45.

[E20] Kreyszig, E., and J. Todd, *QR* in two dimensions. *Elemente der Mathematik* 31 (1976), pp. 109–114.

[E21] Mortensen, M. E., *Geometric Modeling*. 2nd ed. New York: Wiley, 1997.

[E22] Morton, K. W., and D. F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction*. New York: Cambridge University Press, 1994.

[E23] Ortega, J. M., *Introduction to Parallel and Vector Solution of Linear Systems*. New York: Plenum Press, 1988.

[E24] Overton, M. L., *Numerical Computing with IEEE Floating Point Arithmetic*. Philadelphia: SIAM, 2004.

[E25] Press, W. H. et al., *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. New York: Cambridge University Press, 1992.

[E26] Shampine, L. F., *Numerical Solutions of Ordinary Differential Equations*. New York: Chapman and Hall, 1994.

[E27] Varga, R. S., *Matrix Iterative Analysis*. 2nd ed. New York: Springer, 2000.

[E28] Varga, R. S., *Geršgorin and His Circles*. New York: Springer, 2004.

[E29] Wilkinson, J. H., *The Algebraic Eigenvalue Problem*. Oxford: Oxford University Press, 1988.

Part F. Optimization, Graphs (Chaps. 22–23)

[F1] Bondy, J. A. and U.S.R. Murty, *Graph Theory with Applications*. Hoboken, NJ: Wiley-Interscience, 1991.

[F2] Cook, W. J. et al., *Combinatorial Optimization*. New York: Wiley, 1997.

[F3] Diestel, R., *Graph Theory*. 4th ed. New York: Springer, 2006.

[F4] Diwekar, U. M., *Introduction to Applied Optimization*. 2nd ed. New York: Springer, 2008.

[F5] Gass, S. L., *Linear Programming. Method and Applications*. 3rd ed. New York: McGraw-Hill, 1969.

[F6] Gross, J. T. and J.Yellen (eds.), *Handbook of Graph Theory and Applications*. 2nd ed. Boca Raton, FL: CRC Press, 2006.

[F7] Goodrich, M. T., and R. Tamassia, *Algorithm Design: Foundations, Analysis, and Internet Examples*. Hoboken, NJ: Wiley, 2002.

[F8] Harary, F., *Graph Theory*. Reprinted. Reading, MA: Addison-Wesley, 2000.

[F9] Merris, R., *Graph Theory*. Hoboken, NJ: Wiley-Interscience, 2000.

[F10] Ralston, A., and P. Rabinowitz, *A First Course in Numerical Analysis*. 2nd ed. Mineola, NY: Dover, 2001.

[F11] Thulasiraman, K., and M. N. S. Swamy, *Graph Theory and Algorithms*. New York: Wiley-Interscience, 1992.

[F12] Tucker, A., *Applied Combinatorics*. 5th ed. Hoboken, NJ: Wiley, 2007.

Part G. Probability and Statistics (Chaps. 24–25)

[G1] American Society for Testing Materials, *Manual on Presentation of Data and Control Chart Analysis*. 7th ed. Philadelphia: ASTM, 2002.

[G2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Hoboken, NJ: Wiley, 2003.

[G3] Cramér, H., *Mathematical Methods of Statistics*. Reprinted. Princeton, NJ: Princeton University Press, 1999.

[G4] Dodge, Y., *The Oxford Dictionary of Statistical Terms*. 6th ed. Oxford: Oxford University Press, 2006.

[G5] Gibbons, J. D. and S. Chakraborti, *Nonparametric Statistical Inference*. 4th ed. New York: Dekker, 2003.

[G6] Grant, E. L. and R. S. Leavenworth, *Statistical Quality Control*. 7th ed. New York: McGraw-Hill, 1996.

[G7] IMSL, *Fortran Numerical Library*. Houston, TX: Visual Numerics, 2002.

[G8] Kreyszig, E., *Introductory Mathematical Statistics. Principles and Methods*. New York: Wiley, 1970.

[G9] O'Hagan, T. et al., *Kendall's Advanced Theory of Statistics 3-Volume Set*. Kent, U.K.: Hodder Arnold, 2004.

[G10] Rohatgi, V. K. and A. K. MD. E. Saleh, *An Introduction to Probability and Statistics*. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2001.

Web References

[W1] upgraded version of [GenRef1] online at http://dlmf.nist.gov/. Hardcopy and CD-Rom: Oliver, W. J. et al. (eds.), *NIST Handbook of Mathematical Functions*. Cambridge; New York: Cambridge University Press, 2010.

[W2] O'Connor, J. and E. Robertson, MacTutor History of Mathematics Archive. St. Andrews, Scotland: University of St. Andrews, School of Mathematics and Statistics. Online at http://www-history.mcs.st-andrews.ac.uk. (Biographies of mathematicians, etc.).

APPENDIX 2

Answers to Odd-Numbered Problems

Problem Set 1.1, page 8

1. $y = \dfrac{1}{\pi}\cos 2\pi x + c$ **3.** $y = ce^x$

5. $y = 2e^{-x}(\sin x - \cos x) + c$ **7.** $y = \dfrac{1}{5.13}\sinh 5.13x + c$

9. $y = 1.65e^{-4x} + 0.35$ **11.** $y = (x + \tfrac{1}{2})e^x$

13. $y = 1/(1 + 3e^{-x})$ **15.** $y = 0$ and $y = 1$ because $y' = 0$ for these y

17. $\exp(-1.4 \cdot 10^{-11}t) = \tfrac{1}{2}$, $t = 10^{11}(\ln 2)/1.4$ [sec]

19. Integrate $y'' = g$ twice, $y'(t) = gt + v_0$, $y'(0) = v_0 = 0$ (start from rest), then $y(t) = \tfrac{1}{2}gt^2 + y_0$, where $y(0) = y_0 = 0$

Problem Set 1.2, page 11

11. Straight lines parallel to the x-axis **13.** $y = x$

15. $mv' = mg - bv^2$, $v' = 9.8 - v^2$, $v(0) = 10$, $v' = 0$ gives the limit/9.8 = 3.1 [meter/sec]

17. Errors of steps 1, 5, 10: 0.0052, 0.0382, 0.1245, approximately

19. $x_5 = 0.0286$ (error 0.0093), $x_{10} = 0.2196$ (error 0.0189)

Problem Set 1.3, page 18

1. If you add a constant later, you may not get a solution.
Example: $y' = y$, $\ln|y| = x + c$, $y = e^{x+c} = \tilde{c}e^x$ but not $e^x + c$ (with $c \neq 0$)

3. $\cos^2 y\, dy = dx$, $\tfrac{1}{2}y + \tfrac{1}{4}\sin 2y + c = x$

5. $y^2 + 36x^2 = c$, ellipses **7.** $y = x \arctan(x^2 + c)$

9. $y = x/(c - x)$ **11.** $y = 24/x$, hyperbola

13. $dy/\sin^2 y = dx/\cosh^2 x$, $-\cot y = \tanh x + c$, $c = 0$, $y = -\text{arccot}(\tanh x)$

15. $y^2 + 4x^2 = c = 25$ **17.** $y = x\arctan(x^3 - 1)$

19. $y_0e^{kt} = 2y_0$, $e^k = 2$ (1 week), $e^{2k} = 2^2$ (2 weeks), $e^{4k} = 2^4$

21. 69.6% of y_0 **23.** $PV = c =$ const

25. $T = 22 - 17e^{-0.5306t} = 21.9$ [°C] when $t = 9.68$ min

27. $e^{-k\cdot 10} = \tfrac{1}{2}$, $k = \tfrac{1}{10}$, $\ln\tfrac{1}{2}$, $e^{-kt_0} = 0.01$, $t = (\ln 100)/k = 66$ [min]

29. No. Use Newton's law of cooling.

31. $y = ax$, $y' = g(y/x) = a =$ const, independent of the point (x, y)

33. $\Delta S = 0.15S\Delta\phi$, $dS/d\phi = 0.15S$, $S = S_0e^{0.15\phi} = 1000S_0$,
$\phi = (1/0.15)\ln 1000 = 7.3 \cdot 2\pi$. Eight times.

Problem Set 1.4, page 26

1. Exact, $2x = 2x$, $x^2y = c$, $y = c/x^2$
3. Exact, $y = \arccos(c/\cos x)$
5. Not exact, $y = \sqrt{x^2 + cx}$
7. $F = e^{x^2}$, $e^{x^2} \tan y = c$
9. Exact, $u = e^{2x} \cos y + k(y)$, $u_y = -e^{2x} \sin y + k'$, $k' = 0$. *Ans.* $e^{2x} \cos y = 1$
11. $F = \sinh x$, $\sinh^2 x \cos y = c$
13. $u = e^x + k(y)$, $u_y = k' = -1 + e^y$, $k = -y + e^y$. *Ans.* $e^x - y + e^y = c$
15. $b = k$, $ax^2 + 2kxy + ly^2 = c$

Problem Set 1.5, page 34

3. $y = ce^x - 5.2$
5. $y = (x + c)e^{-kx}$
7. $y = x^2(c + e^x)$
9. $y = (x - 2.5/e)e^{\cos x}$
11. $y = 2 + c \sin x$
13. Separate. $y - 2.5 = c \cosh^4 1.5x$
15. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = 0 + 0 = 0$
17. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = r + 0 = r$
19. Solution of $cy_1' + pcy_1 = c(y_1' + py_1) = cr$
21. $y = uy^*$, $y' + py = u'y^* + uy^{*\prime} + puy^* = u'y^* + u(y^{*\prime} + py^*) = u'y^* + u \cdot 0 = r$, $u' = r/y^* = re^{\int p\,dx}$, $u = \int e^{\int p\,dx} r\,dx + c$. Thus, $y = uy_h$ gives (4). We shall see that this method extends to higher-order ODEs (Secs. 2.10 and 3.3).
23. $y^2 = 1 + 8e^{-x^2}$
25. $y = 1/u$, $u = ce^{-3.2x} + 10/3.2$
27. $dx/dy = 6e^y - 2x$, $x = ce^{-2y} + 2e^y$
31. $T = 240e^{kt} + 60$, $T(10) = 200$, $k = -0.0539$, $t = 102$ min
33. $y' = A - ky$, $y(0) = 0$, $y = A(1 - e^{-kt})/k$
35. $y' = 175(0.0001 - y/450)$, $y(0) = 450 \cdot 0.0004 = 0.18$,
$y = 0.135e^{-0.3889t} + 0.045 = 0.18/2$,
$e^{-0.3889t} = (0.09 - 0.045)/0.135 = 1/3$,
$t = (\ln 3)/0.3889 = 2.82$. *Ans.* About 3 years
37. $y' = y - y^2 - 0.2y$, $y = 1/(1.25 - 0.75e^{-0.8t})$, limit 0.8, limit 1
39. $y' = By^2 - Ay = By(y - A/B)$, $A > 0$, $B > 0$. Constant solutions $y = 0$, $y = A/B$, $y' > 0$ if $y > A/B$ (unlimited growth), $y' < 0$ if $0 < y < A/B$ (extinction). $y = A/(ce^{At} + B)$, $y(0) > A/B$ if $c < 0$, $y(0) < A/B$ if $c > 0$.

Problem Set 1.6, page 38

1. $x^2/(c^2 + 9) + y^2/c^2 - 1 = 0$
3. $y - \cosh(x - c) - c = 0$
5. $y/x = c$, $y'/x = y/x^2$, $y' = y/x$, $\tilde{y}' = -x/\tilde{y}$, $\tilde{y}^2 + x^2 = \tilde{c}$, circles
7. $2\tilde{y}^2 - x^2 = \tilde{c}$
9. $y' = -2xy$, $\tilde{y}' = 1/(2x\tilde{y})$, $x = \tilde{c}e^{\tilde{y}^2}$
11. $\tilde{y} = \tilde{c}x$
13. $y' = -4x/9y$. Trajectories $\tilde{y}' = 9\tilde{y}/4x$, $\tilde{y} = \tilde{c}x^{9/4}$ ($\tilde{c} > 0$). Sketch or graph these curves.
15. $u = c$, $u_x\,dx + u_y\,dy = 0$, $y' = -u_x/u_y$. Trajectories $\tilde{y}' = u_{\tilde{y}}/u_x$. Now $v = \tilde{c}$, $v_x\,dx + v_y\,dy = 0$, $y' = -v_x/v_y$. This agrees with the trajectory ODE in u if $u_x = v_y$ (equal denominators) and $u_y = -v_x$ (equal numerators). But these are just the Cauchy–Riemann equations.

Problem Set 1.7, page 42

1. $y' = f(x, y) = r(x) - p(x)y$; hence $\partial f/\partial y = -p(x)$ is continuous and is thus bounded in the closed interval $|x - x_0| \leqq a$.
3. In $|x - x_0| < a$; just take b in $\alpha = b/K$ large, namely, $b = aK$.
5. R has sides $2a$ and $2b$ and center $(1, 1)$ since $y(1) = 1$. In R, $f = 2y^2 \leqq 2(b+1)^2 = K$, $\quad \alpha = b/K = b/(2(b+1)^2)$, $\quad d\alpha/db = 0$ gives $b = 1$, and $\alpha_{\text{opt}} = b/K = \frac{1}{8}$. Solution by $dy/y^2 = 2\,dx$, etc., $y = 1/(3 - 2x)$.
7. $|1 + y^2| \leqq K = 1 + b^2$, $\quad \alpha = b/K$, $\quad d\alpha/db = 0$, $\quad b = 1$, $\quad \alpha = \frac{1}{2}$.
9. No. At a common point (x_1, y_1) they would both satisfy the "initial condition" $y(x_1) = y_1$, violating uniqueness.

Chapter 1 Review Questions and Problems, page 43

11. $y = ce^{-2x}$
13. $y = 1/(ce^{-4x} + 4)$
15. $y = ce^{-x} + 0.01 \cos 10x + 0.1 \sin 10x$
17. $y = ce^{-2.5x} + 0.640x - 0.256$
19. $25y^2 - 4x^2 = c$
21. $F = x$, $x^3e^y + x^2y = c$
23. $y = \sin(x + \frac{1}{4}\pi)$
25. $3 \sin x + \frac{1}{3} \sin y = 0$
27. $e^k = 1.25$, $\quad (\ln 2)/\ln 1.25 = 3.1$, $\quad (\ln 3)/\ln 1.25 = 4.9$ [days]
29. $e^k = 0.9$, 6.6 days. 43.7 days from $e^{kt} = 0.5$, $e^{kt} = 0.01$

Problem Set 2.1, page 53

1. $F(x, z, z') = 0$
3. $y = c_1e^{-x} + c_2$
5. $y = (c_1x + c_2)^{-1/2}$
7. $(dz/dy)z = -z^3 \sin y$, $-1/z = -dx/dy = \cos y + \tilde{c}_1$, $x = -\sin y + c_1y + c_2$
9. $y_2 = x^3 \ln x$
11. $y = c_1e^{2x} + c_2$
13. $y(t) = c_1e^{-t} + kt + c_2$
15. $y = 3 \cos 2.5x - \sin 2.5x$
17. $y = -0.75x^{3/2} - 2.25x^{-1/2}$
19. $y = 15e^{-x} - \sin x$

Problem Set 2.2, page 59

1. $y = c_1e^{-2.5x} + c_2e^{2.5x}$
3. $y = c_1e^{-2.8x} + c_2e^{-3.2x}$
5. $y = (c_1 + c_2x)e^{-\pi x}$
7. $y = c_1 + c_2e^{-4.5x}$
9. $y = c_1e^{-2.6x} + c_2e^{0.8x}$
11. $y = c_1e^{-x/2} + c_2e^{3x/2}$
13. $y = (c_1 + c_2x)e^{5x/3}$
15. $y = e^{-0.27x}(A \cos(\sqrt{\pi}\,x) + B \sin(\sqrt{\pi}\,x))$
17. $y'' + 2\sqrt{5}y' + 5y = 0$
19. $y'' + 4y' + 5y = 0$
21. $y = 4.6 \cos 5x - 0.24 \sin 5x$
23. $y = 6e^{2x} + 4e^{-3x}$
25. $y = 2e^{-x}$
27. $y = (4.5 - x)e^{-\pi x}$
29. $y = \dfrac{1}{\sqrt{\pi}} e^{-0.27x} \sin(\sqrt{\pi}\,x)$
31. Independent
33. $c_1x^2 + c_2x^2 \ln x = 0$ with $x = 1$ gives $c_1 = 0$; then $c_2 = 0$ for $x = 2$, say. Hence independent
35. Dependent since $\sin 2x = 2 \sin x \cos x$
37. $y_1 = e^{-x}$, $\quad y_2 = 0.001e^x + e^{-x}$

Problem Set 2.3, page 61

1. $4e^{2x}$, $-e^{-x} + 8e^{2x}$, $-\cos x - 2 \sin x$
3. 0, 0, $(D - 2I)(-4e^{-2x}) = 8e^{-2x} + 8e^{-2x}$
5. 0, $5e^{2x}$, 0
7. $(2D - I)(2D + I)$, $y = c_1 e^{0.5x} + c_2 e^{-0.5x}$
9. $(D - 2.1I)^2$, $y = (c_1 + c_2 x)e^{2.1x}$
11. $(D - 1.6I)(D - 2.4I)$, $y = c_1 e^{1.6x} + c_2 e^{2.4x}$
15. Combine the two conditions to get $L(cy + kw) = L(cy) + L(kw) = cLy + kLw$. The converse is simple.

Problem Set 2.4, page 69

1. $y' = y_0 \cos \omega_0 t + (v_0/\omega_0) \sin \omega_0 t$. At integer t (if $\omega_0 = \pi$), because of periodicity.
3. (i) Lower by a factor $\sqrt{2}$, (ii) higher by $\sqrt{2}$
5. 0.3183, 0.4775, $\sqrt{(k_1 + k_2)/m}/(2\pi) = 0.5738$
7. $mL\theta'' = -mg \sin \theta \approx -mg\theta$ (tangential component of $W = mg$),
$\theta'' + \omega_0{}^2\theta = 0$, $\omega_0/(2\pi) = \sqrt{g/L}/(2\pi)$
9. $my'' = -\tilde{a}\gamma y$, where $m = 1$ kg, $ay = \pi \cdot 0.01^2 \cdot 2y$ meter^3 is the volume of the water that causes the restoring force $a\gamma y$ with $\gamma = 9800$ nt (= weight/meter^3).
$y'' + \omega_0{}^2 y = 0$, $\omega_0{}^2 = a\gamma/m = a\gamma = 0.000628\gamma$. Frequency $\omega_0/2\pi = 0.4$ [sec^{-1}].
13. $y = [y_0 + (v_0 + \alpha y_0)t]e^{-\alpha t}$, $y = [1 + (v_0 + 1)t]e^{-t}$;
(ii) $v_0 = -2, -\frac{3}{2}, -\frac{4}{3}, -\frac{5}{4}, -\frac{6}{5}$
15. $\omega^* = [\omega_0{}^2 - c^2/(4m^2)]^{1/2} = \omega_0[1 - c^2/(4mk)]^{1/2} \approx \omega_0(1 - c^2/8mk) = 2.9583$
17. The positive solutions of $\tan t = 1$, that is, $\pi/4$ (max), $5\pi/4$ (min). etc
19. $0.0231 = (\ln 2)/30$ [kg/sec] from $\exp(-10 \cdot 3c/2m) = \frac{1}{2}$.

Problem Set 2.5, page 73

3. $y = (c_1 + c_2 \ln x)x^{-1.8}$
5. $\sqrt{x}\,(c_1 \cos(\ln x) + c_2 \sin(\ln x))$
7. $y = c_1 x^2 + c_2 x^3$
9. $y = (c_1 + c_2 \ln x)x^{0.6}$
11. $y = x^2(c_1 \cos(\sqrt{6} \ln x) + c_2 \sin(\sqrt{6} \ln x))$
13. $y = x^{-3/2}$
15. $y = (3.6 + 4.0 \ln x)/x$
17. $y = \cos(\ln x) + \sin(\ln x)$
19. $y = -0.525x^5 + 0.625x^{-3}$

Problem Set 2.6, page 79

3. $W = -2.2e^{-3x}$ **5.** $W = -x^4$ **7.** $W = a$
9. $y'' + 25y = 0$, $W = 5$, $y = 3 \cos 5x - \sin 5x$
11. $y'' + 5y + 6.34 = 0$, $W = 0.3e^{-5x}$, $3e^{-2.5} \cos 0.3x$
13. $y'' + 2y' = 0$, $W = -2e^{-2x}$, $y = 0.5(1 + e^{-2x})$
15. $y'' - 3.24y = 0$, $W = 1.8$, $y = 14.2 \cosh 1.8x + 9.1 \sinh 1.8x$

Problem Set 2.7, page 84

1. $y = c_1 e^{-x} + c_2 e^{-4x} - 5e^{-3x}$
3. $y = c_1 e^{-2x} + c_2 e^{-x} + 6x^2 - 18x + 21$
5. $y = (c_1 + c_2 x)e^{-2x} + \frac{1}{2}e^{-x} \sin x$
7. $y = c_1 e^{-x/2} + c_2 e^{-3x/2} + \frac{4}{5}e^x + 6x - 16$
9. $y = c_1 e^{4x} + c_2 e^{-4x} + 1.2xe^{4x} - 2e^x$
11. $y = \cos(\sqrt{3}x) + 6x^2 - 4$

13. $y = e^{x/4} - 2e^{x/2} + \frac{1}{5}e^{-x} + e^{x}$ **15.** $y = \ln x$
17. $y = e^{-0.1x}(1.5 \cos 0.5x - \sin 0.5x) + 2e^{0.5x}$

Problem Set 2.8, page 91

3. $y_p = 1.0625 \cos 2t + 3.1875 \sin 2t$
5. $y_p = -1.28 \cos 4.5t + 0.36 \sin 4.5t$
7. $y_p = 25 + \frac{4}{3} \cos 3t + \sin 3t$
9. $y = e^{-1.5t}(A \cos t + B \sin t) + 0.8 \cos t + 0.4 \sin t$
11. $y = A \cos \sqrt{2}t + B \sin \sqrt{2}t + t(\sin \sqrt{2}t - \cos \sqrt{2}t)/(2\sqrt{2})$
13. $y = A \cos t + B \sin t - (\cos \omega t)/(\omega^2 - 1)$
15. $y = e^{-2t}(A \cos 2t + B \sin 2t) + \frac{1}{4} \sin 2t$
17. $y = \frac{1}{3} \sin t - \frac{1}{15} \sin 3t - \frac{1}{105} \sin 5t$
19. $y = e^{-t}(0.4 \cos t + 0.8 \sin t) + e^{-t/2}(-0.4 \cos \frac{1}{2}t + 0.8 \sin \frac{1}{2}t)$
25. CAS Experiment. The choice of ω needs experimentation, inspection of the curves obtained, and then changes on a trail-and-error basis. It is interesting to see how in the case of beats the period gets increasingly longer and the maximum amplitude gets increasingly larger as $\omega/(2\pi)$ approaches the resonance frequency.

Problem Set 2.9, page 98

1. $RI' + I/C = 0, \quad I = ce^{-t/(RC)}$
3. $LI' + RI = E, \quad I = (E/R) + ce^{-Rt/L} = 4.8 + ce^{-40t}$
5. $I = 2(\cos t - \cos 20t)/399$
7. I_0 is maximum when $S = 0$; thus, $C = 1/(\omega^2 L)$.
9. $I = 0$ **11.** $I = 5.5 \cos 10t + 16.5 \sin 10t$ A
13. $I = e^{-5t}(A \cos 10t + B \sin 10t) - 400 \cos 25t + 200 \sin 25t$ A
15. $R > R_{\text{crit}} = 2\sqrt{L/C}$ is Case I, etc.
17. $E(0) = 600, I'(0) = 600, I = e^{-3t}(-100 \cos 4t + 75 \sin 4t) + 100 \cos t$
19. $R = 2\ \Omega, \quad L = 1$ H, $\quad C = \frac{1}{12}$ F, $\quad E = 4.4 \sin 10t$ V

Problem Set 2.10, page 102

1. $y = A \cos 3x + B \sin 3x + \frac{1}{9}(\cos 3x) \ln |\cos 3x| + \frac{1}{3}x \sin 3x$
3. $y = c_1 x + c_2 x^2 - x \sin x$ **5.** $y = A \cos x + B \sin x + \frac{1}{2}x(\cos x + \sin x)$
7. $y = (c_1 + c_2 x) e^{2x} + x^{-2} e^{2x}$ **9.** $y = (c_1 + c_2 x)e^{x} + 4x^{7/2} e^{x}$
11. $y = c_1 x^2 + c_2 x^3 + 1/(2x^4)$ **13.** $y = c_1 x^{-3} + c_2 x^3 + 3x^5$

Chapter 2 Review Questions and Problems, page 102

7. $y = c_1 e^{-4.5x} + c_2 e^{-3.5x}$ **9.** $y = e^{-3x}(A \cos 5x + B \sin 5x)$
11. $y = (c_1 + c_2 x)e^{0.8x}$ **13.** $y = c_1 x^{-4} + c_2 x^3$
15. $y = c_1 e^{2x} + c_2 e^{-x/2} - 3x + x^2$ **17.** $y = (c_1 + c_2 x)e^{1.5x} + 0.25x^2 e^{1.5x}$
19. $y = 5 \cos 4x - \frac{3}{4}\sin 4x + e^{x}$ **21.** $y = -4x + 2x^3 + 1/x$
23. $I = -0.01093 \cos 415t + 0.05273 \sin 415t$ A

25. $I = \frac{1}{73}(50 \sin 4t - 110 \cos 4t)$ A
27. *RLC*-circuit with $R = 20\ \Omega$, $L = 4$ H, $C = 0.1$ F, $E = -25 \cos 4t$ V
29. $\omega = 3.1$ is close to $\omega_0 = \sqrt{k/m} = 3$, $y = 25(\cos 3t - \cos 3.1t)$.

Problem Set 3.1, page 111

9. Linearly independent
11. Linearly independent
13. Linearly independent
15. Linearly dependent

Problem Set 3.2, page 116

1. $y = c_1 + c_2 \cos 5x + c_3 \sin 5x$
3. $y = c_1 + c_2 x + c_3 \cos 2x + c_4 \sin 2x$
5. $y = A_1 \cos x + B_1 \sin x + A_2 \cos 3x + B_2 \sin 3x$
7. $y = 2.398 + e^{-1.6x}(1.002 \cos 1.5x - 1.998 \sin 1.5x)$
9. $y = 4e^{-x} + 5e^{-x/2} \cos 3x$
11. $y = \cosh 5x - \cos 4x$
13. $y = e^{0.25x} + 4.3e^{-0.7x} + 12.1 \cos 0.1x - 0.6 \sin 0.1x$

Problem Set 3.3, page 122

1. $y = (c_1 + c_2 x + c_3 x^2)e^{-x} + \frac{1}{8}e^{x} - x + 2$
3. $y = c_1 \cos x + c_2 \sin x + c_3 \cos 3x + c_4 \sin 3x + 0.1 \sinh 2x$
5. $y = c_1 x^2 + c_2 x + c_3 x^{-1} - \frac{1}{12}x^{-2}$
7. $y = (c_1 + c_2 x + c_3 x^2)e^{3x} - \frac{1}{4}(\cos 3x - \sin 3x)$
9. $y = \cos x + \frac{1}{2}\sin 4x$
11. $y = e^{-3x}(-1.4 \cos x - \sin x)$
13. $y = 2 - 2 \sin x + \cos x$

Chapter 3 Review Questions and Problems, page 122

7. $y = c_1 + e^{-2x}(A \cos 3x + B \sin 3x)$
9. $y = c_1 \cosh 2x + c_2 \sinh 2x + c_3 \cos 2x + c_4 \sin 2x + \cosh x$
11. $y = (c_1 + c_2 x + c_3 x^2)e^{-1.5x}$
13. $y = (c_1 + c_2 x + c_3 x^2)e^{-2x} + x^2 - 3x + 3$
15. $y = c_1 x + c_2 x^{1/2} + c_3 x^{3/2} - \frac{10}{3}$
17. $y = 2e^{-2x} \cos 4x + 0.05\,x - 0.06$
19. $y = 4e^{-4x} + 5e^{-5x}$

Problem Set 4.1, page 136

1. Yes
5. $y_1' = 0.02(-y_1 + y_2)$, $y_2' = 0.02(y_1 - 2y_2 + y_3)$, $y_3' = 0.02(y_2 - y_3)$
7. $c_1 = 1$, $c_2 = -5$
9. $c_1 = 10$, $c_2 = 5$
11. $y_1' = y_2$, $y_2' = y_1 + \ + \frac{15}{4}y_2$, $\mathbf{y} = c_1[1 \quad 4]^{\mathsf{T}} e^{4t} + c_2 [1 \quad -\frac{1}{4}]^{\mathsf{T}} e^{-t/4}$
13. $y_1' = y_2$, $y_2' = 24y_1 - 2y_2$, $y_1 = c_1 e^{4t} + c_2 e^{-6t} = y$, $y_2 = y'$
15. **(a)** For example, $C = 1000$ gives -2.39993, -0.000167. **(b)** -2.4, 0.
(d) $a_{22} = -4 + 2\sqrt{6.4} = 1.05964$ gives the critical case. C about 0.18506.

Problem Set 4.3, page 147

1. $y_1 = c_1e^{-2t} + c_2e^{2t}, \quad y_2 = -3c_1e^{-2t} + c_2e^{2t}$
3. $y_1 = 2c_1e^{2t} + 2c_2, \quad y_2 = c_1e^{2t} - c_2$
5. $y_1 = 5c_1 + 2c_2e^{14.5t}$
$y_2 = -2c_1 + 5c_2e^{14.5t}$
7. $y_1 = -c_2 \cos \sqrt{2}t + c_3 \sin \sqrt{2}t + c_1$
$y_2 = c_2\sqrt{2} \sin \sqrt{2}t + c_3\sqrt{2} \cos \sqrt{2}t$
$y_3 = c_2 \cos \sqrt{2}t - c_3 \sin \sqrt{2}t + c_1$
9. $y_1 = \frac{1}{2}c_1e^{-18t} + 2c_2e^{9t} - c_3e^{18t}$
$y_2 = c_1e^{-18t} + c_2e^{9t} + c_3e^{18t}$
$y_3 = c_1e^{-18t} - 2c_2e^{9t} - \frac{1}{2}c_3e^{18t}$
11. $y_1 = -20e^t + 8e^{-t/2}$
$y_2 = 4e^t - 4e^{-t/2}$
13. $y_1 = 2 \sinh t, \quad y_2 = 2 \cosh t$
15. $y_1 = \frac{1}{2}e^t$
$y_2 = \frac{1}{2}e^t$
17. $y_2 = y_1' + y_1, \quad y_2' = y_1'' + y_1' = -y_1 - y_2 = -y_1 - (y_1' + y_1),$
$y_1'' + 2y_1' + 2y_1 = 0, \quad y_1 = e^{-t}(A \cos t + B \sin t),$
$y_2 = y_1' + y_1 = e^{-t}(B \cos t - A \sin t)$. Note that $r^2 = y_1^2 + y_2^2 = e^{-2t}(A^2 + B^2)$.
19. $I_1 = c_1e^{-t} + 3c_2e^{-3t}, I_2 = -3c_1e^{-t} - c_2e^{-3t}$

Problem Set 4.4, page 151

1. Unstable improper node, $y_1 = c_1e^t, \quad y_2 = c_2e^{2t}$
3. Center, always stable, $y_1 = A \cos 3t + B \sin 3t, \quad y_2 = 3B \cos 3t - 3A \sin 3t$
5. Stable spiral, $y_1 = e^{-2t}(A \cos 2t + B \sin 2t), \quad y_2 = e^{-2t}(B \cos 2t - A \sin 2t)$
7. Saddle point, always unstable, $y_1 = c_1e^{-t} + c_2e^{3t}, \quad y_2 = -c_1e^{-t} + c_2e^{3t}$
9. Unstable node, $y_1 = c_1e^{6t} + c_2e^{2t}, \quad y_2 = 2c_1e^{6t} - 2c_2e^{2t}$
11. $y = e^{-t}(A \cos t + B \sin t)$. Stable and attractive spirals
15. $p = 0.2 \neq 0$ (was 0), $\quad \Delta < 0$, spiral point, unstable.
17. For instance, **(a)** -2, **(b)** -1, **(c)** $= -\frac{1}{2}$, **(d)** $=1$, **(e)** 4.

Problem Set 4.5, page 159

5. Center at (0, 0). At (2, 0) set $y_1 = 2 + \tilde{y}_1$. Then $\tilde{y}_2' = \tilde{y}_1$. Saddle point at (2, 0).
7. (0, 0), $y_1' = -y_1 + y_2, \quad y_2' = -y_1 - y_2$, stable and attractive spiral point; $(-2, 2)$,
$y_1 = -2 + \tilde{y}_1, \quad y_2 = 2 + \tilde{y}_2, \quad \tilde{y}_1' = -\tilde{y}_1 - 3\tilde{y}_2, \quad \tilde{y}_2' = -\tilde{y}_1 - \tilde{y}_2$, saddle point
9. (0, 0) saddle point, $(-3, 0)$ and (3, 0) centers
11. $(\frac{1}{2}\pi \pm 2n\pi, 0)$ saddle points; $(-\frac{1}{2}\pi \pm 2n\pi, 0)$ centers.
Use $-\cos(\pm\frac{1}{2}\pi + \tilde{y}_1) = \sin(\pm\tilde{y}_1) \approx \pm\tilde{y}_1$.
13. $(\pm 2n\pi, 0)$ centers; $y_1 = (2n + 1)\pi + \tilde{y}_1', \quad (\pi \pm 2n\pi, 0)$ saddle points
15. By multiplication, $y_2y_2' = (4y_1 - y_1^3)y_1'$. By integration,
$y_2^2 = 4y_1^2 - \frac{1}{2}y_1^4 + c^* = \frac{1}{2}(c + 4 - y_1^2)(c - 4 + y_1^2)$, where $c^* = \frac{1}{2}c^2 - 8$.

Problem Set 4.6, page 163

3. $y_1 = c_1e^{-t} + c_2e^t, \quad y_2 = -c_1e^{-t} + c_2e^t - e^{3t}$
5. $y_1 = c_1e^{5t} + c_2e^{2t} - 0.43t - 0.24, \quad y_2 = c_1e^{5t} - 2c_2e^{2t} + 1.12t + 0.53$

7. $y_1 = c_1e^t + 4c_2e^{2t} - 3t - 4 - 2e^{-t}$, $y_2 = -c_1e^t - 5c_2e^{2t} + 5t + 7.5 + e^{-t}$

9. The formula for $\mathbf{v}$ shows that these various choices differ by multiples of the eigenvector for $\lambda = -2$, which can be absorbed into, or taken out of, c_1 in the general solution $y^{(h)}$.

11. $y_1 = -\frac{8}{3}\cosh t - \frac{4}{3}\sinh t + \frac{11}{3}e^{2t}$, $y_2 = -\frac{8}{3}\sinh t - \frac{4}{3}\cosh t + \frac{4}{3}e^{2t}$

13. $y_1 = \cos 2t + \sin 2t + 4\cos t$, $y_2 = 2\cos 2t - 2\sin 2t + \sin t$

15. $y_1 = 4e^{-t} - 4e^t + e^{2t}$, $y_2 = -4e^{-t} + t$

17. $I_1 = 2c_1e^{\lambda_1 t} + 2c_2e^{\lambda_2 t} + 100$,
$I_2 = (1.1 + \sqrt{0.41})c_1e^{\lambda_1 t} + (1.1 - \sqrt{0.41})c_2e^{\lambda_2 t}$,
$\lambda_1 = -0.9 + \sqrt{0.41}$, $\lambda_2 = -0.9 - \sqrt{0.41}$

19. $c_1 = 17.948$, $c_2 = -67.948$

Chapter 4 Review Questions and Problems, page 164

11. $y_1 = c_1e^{4t} + c_2e^{-4t}$, $y_2 = 2c_1e^{4t} - 2c_2e^{-4t}$. Saddle point

13. $y_1 = e^{-4t}(A\cos t + B\sin t)$, $y_2 = \frac{1}{5}e^{-4t}[(B - 2A)\cos t - (A + 2B)\sin t]$; asymptotically stable spiral point

15. $y_1 = c_1e^{-5t} + c_2e^{-t}$, $y_2 = c_1e^{-5t} - c_2e^{-t}$. Stable node

17. $y_1 = e^{-t}(A\cos 2t + B\sin 2t)$, $y_2 = e^{-t}(B\cos 2t - A\sin 2t)$. Stable and attractive spiral point

19. Unstable spiral point

21. $y_1 = c_1e^{-4t} + c_2e^{4t} - 1 - 8t^2$, $y_2 = -c_1e^{-4t} + c_2e^{4t} - 4t$

23. $y_1 = 2c_1e^{-t} + 2c_2e^{3t} + \cos t - \sin t$, $y_2 = -c_1e^{-t} + c_2e^{3t}$

25. $I_1' + 2.5(I_1 - I_2) = 169\sin t$, $2.5(I_2' - I_1') + 25I_2 = 0$,
$I_1 = (19 + 32.5t)e^{-5t} - 19\cos t + 62.5\sin t$,
$I_2 = (-6 - 32.5t)e^{-5t} + 6\cos t + 2.5\sin t$

27. $(0, 0)$ saddle point; $(-1, 0)$, $(1, 0)$ centers

29. $(n\pi, 0)$ center when n is even and saddle point when n is odd

Problem Set 5.1, page 174

3. $\sqrt{|k|}$

5. $\sqrt{3/2}$

7. $y = a_0(1 - x^2 + x^4/2! - x^6/3! + - \cdots) = a_0e^{-x^2}$

9. $y = a_0 + a_1x - \frac{1}{2}a_0x^2 - \frac{1}{6}a_1x^3 + \cdots = a_0\cos x + a_1\sin x$

11. $a_0(1 - \frac{1}{12}x^4 - \frac{1}{60}x^5 - \cdots) + a_1(x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{24}x^5 - \cdots)$

13. $a_0(1 - \frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{13}{720}x^6 + \cdots) + a_1(x - \frac{1}{6}x^3 - \frac{1}{24}x^5 + \frac{5}{1008}x^7 + \cdots)$

15. $\displaystyle\sum_{m=1}^{\infty} \frac{(m+1)(m+2)}{(m+1)^2 + 1}x^m$, $\displaystyle\sum_{m=5}^{\infty} \frac{(m-4)^2}{(m-3)!}x^m$

17. $s = 1 + x - x^2 - \frac{5}{6}x^3 + \frac{2}{3}x^4 + \frac{11}{24}x^5$, $s(\frac{1}{2}) = \frac{923}{768}$

19. $s = 4 - x^2 - \frac{1}{3}x^3 + \frac{1}{30}x^5$, $s(2) = -\frac{8}{5}$; but $x = 2$ is too large to give good values. Exact: $y = (x - 2)^2e^x$

Problem Set 5.2, page 179

5. $P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5)$,
$P_7(x) = \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x)$

11. Set $x = az$. $y = c_1 P_n(x/a) + c_2 Q_n(x/a)$

15. $P_1^1 = \sqrt{1 - x^2}$, $P_2^1 = 3x\sqrt{1 - x^2}$, $P_2^2 = 3(1 - x^2)$,
$P_4^2 = (1 - x^2)(105x^2 - 15)/2$

Problem Set 5.3, page 186

3. $y_1 = 1 - \dfrac{x^2}{3!} + \dfrac{x^4}{5!} - + \cdots = \dfrac{\sin x}{x}$, $y_2 = \dfrac{1}{x} - \dfrac{x}{2!} + \dfrac{x^3}{4!} - + \cdots = \dfrac{\cos x}{x}$

5. $b_0 = 1$, $c_0 = 0$, $r^2 = 0$, $y_1 = e^{-x}$, $y_2 = e^{-x} \ln x$

7. $y_1 = 1 + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{30}x^5 + \frac{1}{144}x^6 - \cdots$,
$y_2 = x + \frac{1}{6}x^3 - \frac{1}{12}x^4 + \frac{1}{120}x^5 - \frac{1}{120}x^6 + \cdots$

9. $y_1\sqrt{x}$, $y_2 = 1 + x$

11. $y_1 = e^x$, $y_2 = e^x/x$

13. $y_1 = e^x$, $y_2 = e^x \ln x$

15. $y = AF(1, 1, -\frac{1}{2}; x) + Bx^{3/2}F(\frac{5}{2}, \frac{5}{2}, \frac{5}{2}; x)$

17. $y = A(1 - 8x + \frac{32}{5}x^2) + Bx^{3/4}F(\frac{7}{4}, -\frac{5}{4}, \frac{7}{4}; x)$

19. $y = c_1F(2, -2, -\frac{1}{2}; t - 2) + c_2(t - 2)^{3/2}F(\frac{7}{2}, -\frac{1}{2}, \frac{5}{2}; t - 2)$

Problem Set 5.4, page 195

3. $c_1 J_0(\sqrt{x})$

5. $c_1 J_\nu(\lambda x) + c_2 J_{-\nu}(\lambda x)$, $\nu \neq 0, \pm 1, \pm 2, \cdots$

7. $c_1 J_{1/2}(\frac{1}{2}x) + c_2 J_{-1/2}(\frac{1}{2}x) = x^{-1/2}(\tilde{c}_1 \sin \frac{1}{2}x + \tilde{c}_2 \cos \frac{1}{2}x)$

9. $x^{-\nu}(c_1 J_\nu(x) + c_2 J_{-\nu}(x))$, $\nu \neq 0, \pm 1, \pm 2, \cdots$

13. $J_n(x_1) = J_n(x_2) = 0$ implies $x_1^{-n}J_n(x_1) = x_2^{-n}J_n(x_2) = 0$ and $[x^{-n}J_n(x)]' = 0$ somewhere between x_1 and x_2 by Rolle's theorem. Now use (21b) to get $J_{n+1}(x) = 0$ there. Conversely, $J_{n+1}(x_3) = J_{n+1}(x_4) = 0$, thus $x_3^{n+1}J_{n+1}(x_3) = x_4^{n+1}J_{n+1}(x_4) = 0$ implies $J_n(x) = 0$ in between by Rolle's theorem and (21a) with $\nu = n + 1$.

15. By Rolle, $J_0' = 0$ at least once between two zeros of J_0. Use $J_0' = -J_1$ by (21b) with $\nu = 0$. Together $J_1 = 0$ at least once between two zeros of J_0. Also use $(xJ_1)' = xJ_0$ by (21a) with $\nu = 1$ and Rolle.

19. Use (21b) with $\nu = 0$, (21a) with $\nu = 1$, (21d) with $\nu = 2$, respectively.

21. Integrate (21a).

23. Use (21a) with $\nu = 1$, partial integration, (21b) with $\nu = 0$, partial integration.

25. Use (21d) to get

$$\int J_5(x)\,dx = -2J_4(x) + \int J_3(x)\,dx = -2J_4(x) - 2J_2(x) + \int J_1(x)\,dx$$
$$= -2J_4(x) - 2J_2(x) - J_0(x) + c.$$

Problem Set 5.5, page 200

1. $c_1 J_4(x) + c_2 Y_4(x)$

3. $c_1 J_{2/3}(x^2) + c_2 Y_{2/3}(x^2)$

5. $c_1 J_0(\sqrt{x}) + c_2 Y_0(\sqrt{x})$

7. $\sqrt{x}\,(c_1 J_{1/4}(\frac{1}{2}kx^2) + c_2 Y_{1/4}(\frac{1}{2}kx^2))$
9. $x^3(c_1 J_3(x) + c_2 Y_3(x))$
11. Set $H^{(1)} = kH^{(2)}$ and use (10).
13. Use (20) in Sec. 5.4.

Chapter 5 Review Questions and Problems, page 200

11. $\cos 2x$, $\sin 2x$
13. $(x-1)^{-5}$, $(x-1)^7$; Euler–Cauchy with $x - 1$ instead of x
15. $J_{\sqrt{5}}(x)$, $J_{-\sqrt{5}}(x)$
17. e^x, $1 + x$
19. $\sqrt{x}\,J_1(\sqrt{x})$, $\sqrt{x}\,Y_1(\sqrt{x})$

Problem Set 6.1, page 210

1. $3/s^2 + 12/s$
3. $s/(s^2 + \pi^2)$
5. $1/((s-2)^2 - 1)$
7. $(\omega \cos\theta + s\sin\theta)/(s^2 + \omega^2)$
9. $\dfrac{1}{s} + \dfrac{e^{-s} - 1}{s^2}$
11. $\dfrac{1 - e^{-bs}}{s^2} - \dfrac{be^{-bs}}{s}$
13. $\dfrac{(1 - e^{-s})^2}{s}$
15. $\dfrac{e^{-s} - 1}{2s^2} - \dfrac{e^{-s}}{2s} + \dfrac{1}{s}$
19. Use $e^{at} = \cosh at + \sinh at$.
23. Set $ct = p$. Then $\mathscr{L}(f(ct)) = \displaystyle\int_0^\infty e^{-st} f(ct)\,dt = \int_0^\infty e^{-(s/c)p} f(p)\,dp/c = F(s/c)/c.$
25. $0.2\cos 1.8t + \sin 1.8t$
27. $\dfrac{1}{L^2}\cos\dfrac{n\pi t}{L}$
29. $2t^3 - 1.9t^5$
31. $\mathscr{L}^{-1}\left(\dfrac{4}{s-2} - \dfrac{3}{s+1}\right) = 4e^{2t} - 3e^{-t}$
33. $\dfrac{2}{(s+3)^3}$
35. $\dfrac{0.5 \cdot 2\pi}{(s + 4.5)^2 + 4\pi^2}$
37. $\pi t e^{-\pi t}$
39. $\frac{7}{2}t^3 e^{-t\sqrt{2}}$
41. $e^{-5\pi t}\sinh \pi t$
43. $e^{3t}(2\cos 3t + \frac{5}{3}\sin 3t)$
45. $(k_0 + k_1 t)e^{-at}$

Problem Set 6.2, page 216

1. $y = 1.25e^{-5.2t} - 1.25\cos 2t + 3.25\sin 2t$
3. $(s-3)(s+2) = 11s + 28 - 11 = 11s + 17$, $\quad Y = 10/(s-3) + 1/(s+2)$,
$y = 10e^{3t} + e^{-2t}$
5. $(s^2 - \frac{1}{4})Y = 12s$, $\quad y = 12\cosh\frac{1}{2}t$
7. $y = \frac{1}{2}e^{3t} + \frac{5}{2}e^{-4t} + \frac{1}{2}e^{-3t}$
9. $y = e^t - e^{3t} + 2t$
11. $(s + 1.5)^2 Y = s + 31.5 + 3 + 54/s^4 + 64/s$,
$Y = 1/(s + 1.5) + 1/(s + 1.5)^2 + 24/s^4 - 32/s^3 + 32/s^2$,
$y = (1 + t)e^{-1.5t} + 4t^3 - 16t^2 + 32t$
13. $t = \tilde{t} - 1$, $\quad \tilde{Y} = 4/(s-6)$, $\quad \tilde{y} = 4e^{6t}$, $\quad y = 4e^{6(t+1)}$

15. $t = \tilde{t} + 1.5$, $(s-1)(s+4)\tilde{Y} = 4s + 17 + 6/(s-2)$, $y = 3e^{t-1.5} + e^{2(t-1.5)}$

17. $\dfrac{1}{(s+a)^2}$

19. $\dfrac{2\omega^2}{s(s^2+4\omega^2)}$

21. $\mathcal{L}(f') = \mathcal{L}(\sinh 2t) = s\mathcal{L}(f) - 1$. *Answer:* $(s^2-2)/(s^3-4s)$

23. $12(1 - e^{-t/4})$

25. $(1 - \cos \omega t)/\omega^2$

27. $\frac{1}{9}(1 + t - \cos 3t - \frac{1}{3}\sin 3t)$

29. $\dfrac{1}{a^2}(e^{-at} - 1) + \dfrac{t}{a}$

Problem Set 6.3, page 223

3. $\mathcal{L}((t-2)u(t-2)) = e^{-2s}/s^2$

5. $\left(e^t\left(1 - u\left(t - \frac{1}{2}\pi\right)\right)\right) = \dfrac{1}{s-1}(1 - e^{-\pi s/2 + \pi/2})$

7. $\dfrac{1}{s+\pi}(e^{-2(s+\pi)} - e^{-4(s+\pi)})$

9. $e^{-3s/2}\left(\dfrac{2}{s^3} + \dfrac{3}{s^2} + \dfrac{\frac{9}{4}}{s}\right)$

11. $(se^{-\pi s/2} + e^{-\pi s})/(s^2+1)$

13. $2[1 + u(t-\pi)]\sin 3t$

15. $(t-3)^3 u(t-3)/6$

17. $e^{-t}\cos t\ (0 < t < 2\pi)$

19. $\frac{1}{3}(e^t - 1)^3 e^{-5t}$

21. $\sin 3t + \sin t\ (0 < t < \pi)$; $\frac{4}{3}\sin 3t\ (t > \pi)$

23. $e^t - \sin t\ (0 < t < 2\pi)$, $e^t - \frac{1}{2}\sin 2t\ (t > 2\pi)$

25. $t - \sin t\ (0 < t < 1)$, $\cos(t-1) + \sin(t-1) - \sin t\ (t > 1)$

27. $t = 1 + \tilde{t}$, $\tilde{y}'' + 4\tilde{y} = 8(1+\tilde{t})^2(1 - u(\tilde{t} - 4))$, $\cos 2t + 2t^2 - 1$ if $t < 5$, $\cos 2t + 49\cos(2t - 10) + 10\sin(2t - 10)$ if $t > 5$

29. $0.1i' + 25i = 490e^{-5t}[1 - u(t-1)]$,
$i = 20(e^{-5t} - e^{-250t}) + 20u(t-1)[-e^{-5t} + e^{-250t+245}]$

31. $Rq' + q/C = 0$, $Q = \mathcal{L}(q)$, $q(0) = CV_0$, $i = q'(t)$,
$R(sQ - CV_0) + Q/C = 0$, $q = CV_0 e^{-t/(RC)}$

33. $10I + \dfrac{100}{s}I = \dfrac{100}{s^2}e^{-2s}$, $I = e^{-2s}\left(\dfrac{1}{s} - \dfrac{1}{s+10}\right)$, $i = 0$ if $t < 2$ and
$1 - e^{-10(t-2)}$ if $t > 2$

35. $i = (10\sin 10t + 100\sin t)(u(t-\pi) - u(t-3\pi))$

37. $(0.5s^2 + 20)I = 78s(1 + e^{-\pi s})/(s^2+1)$,
$i = 4\cos t - 4\cos\sqrt{40}t - 4u(t-\pi)[\cos t + \cos(\sqrt{40}(t-\pi))]$

39. $i' + 2i + 2\displaystyle\int_0^t i(\tau)\,d\tau = 1000(1 - u(t-2))$, $I = 1000(1 - e^{-2s})/(s^2+2s+2)$,

$i = 1000e^{-t}\sin t - 1000u(t-2)e^{-t+2}\sin(t-2)$

Problem Set 6.4, page 230

3. $y = 8\cos 2t + \frac{1}{2}u(t-\pi)\sin 2t$

5. $\sin t\ (0 < t < \pi)$; $0\ (\pi < t < 2\pi)$; $-\sin t\ (t > 2\pi)$

7. $y = e^{-t} + 4e^{-3t}\sin\frac{1}{2}t + \frac{1}{2}u(t - \frac{1}{2})e^{-3(t-1/2)}\sin(\frac{1}{2}t - \frac{1}{4})$

9. $y = 0.1[e^t + e^{-2t}(-\cos t + 7\sin t)] + 0.1u(t-10)[-e^{-t} + e^{-2t+30}(\cos(t-10) - 7\sin(t-10))]$

11. $y = -e^{-3t} + e^{-2t} + \frac{1}{6}u(t-1)(1 - 3e^{-2(t-1)} + 2e^{-3(t-1)}) + u(t-2)(e^{-2(t-2)} - e^{-3(t-2)})$

15. $ke^{-ps}/(s - se^{-ps})$ $(s > 0)$

Problem Set 6.5, page 237

1. t

3. $(e^t - e^{-t})/2 = \sinh t$

5. $\frac{1}{2}t \sin \omega t$

7. $e^t - t - 1$

9. $y - 1 * y = 1, \quad y = e^t$

11. $y = \cos t$

13. $y(t) + 2\int_0^t e^{t-\tau} y(\tau)\, d\tau = te^t, \quad y = \sinh t$

17. $e^{4t} - e^{-1.5t}$

19. $t \sin \pi t$

21. $(\omega t - \sin \omega t)/\omega^2$

23. $4.5(\cosh 3t - 1)$

25. $1.5t \sin 6t$

Problem Set 6.6, page 241

3. $\dfrac{\frac{1}{2}}{(s+3)^2}$

5. $\dfrac{s^2 - \omega^2}{(s^2 + \omega^2)^2}$

7. $\dfrac{2s^3 + 24s}{(s^2 - 4)^3}$

9. $\dfrac{\pi(3s^2 - \pi^2)}{(s^2 + \pi^2)^3}$

11. $\dfrac{4s^2 - \pi^2}{(s^2 + \frac{1}{4}\pi^2)^2}$

15. $F(s) = -\dfrac{1}{2}\left(\dfrac{1}{s^2 - 9}\right)', \quad f(t) = \frac{1}{6}t \sinh 3t$

17. $\ln s - \ln(s-1); \quad (-1 + e^t)/t$

19. $[\ln(s^2 + 1) - 2\ln(s-1)]' = 2s/(s^2+1) - 2/(s-1); \quad 2(-\cos t + e^t)/t$

Problem Set 6.7, page 246

3. $y_1 = -e^{-5t} + 4e^{2t}, \quad y_2 = e^{-5t} + 3e^{2t}$

5. $y_1 = -\cos t + \sin t + 1 + u(t-1)[-1 + \cos(t-1) - \sin(t-1)]$
$y_2 = \cos t + \sin t - 1 + u(t-1)[1 - \cos(t-1) - \sin(t-1)]$

7. $y_1 = -e^{-2t} + 4e^t + \frac{1}{3}u(t-1)(-e^{3-2t} + e^t),$
$y_2 = -e^{-2t} + e^t + \frac{1}{3}u(t-1)(-e^{3-2t} + e^t)$

9. $y_1 = (3 + 4t)e^{3t}, \quad y_2 = (1 - 4t)e^{3t}$

11. $y_1 = e^t + e^{2t}, \quad y_2 = e^{2t}$

13. $y_1 = -4e^t + \sin 10t + 4\cos t, \quad y_2 = 4e^t - \sin 10t + 4\cos t$

15. $y_1 = e^t, \quad y_2 = e^{-t}, \quad y_3 = e^t - e^{-t}$

19. $4i_1 + 8(i_1 - i_2) + 2i_1' = 390\cos t, \quad 8i_2 + 8(i_2 - i_1) + 4i_2' = 0,$
$i_1 = -26e^{-2t} - 16e^{-8t} + 42\cos t + 15\sin t,$
$i_2 = -26e^{-2t} + 8e^{-8t} + 18\cos t + 12\sin t$

Chapter 6 Review Questions and Problems, page 251

11. $\dfrac{5s}{s^2 - 4} - \dfrac{3}{s^2 - 1}$

13. $\frac{1}{2}(1 - \cos \pi t), \quad \pi^2/(2s^3 + 2\pi^2 s)$

15. $e^{-3s+3/2}/(s - \frac{1}{2})$

17. Sec. 6.6; $2s^2/(s^2+1)^2$

19. $12/(s^2(s+3))$

21. $tu(t-1)$

23. $\sin(\omega t+\theta)$

25. $3t^2+t^3$

27. $e^{-2t}(3\cos t - 2\sin t)$

29. $y = e^{-2t}(13\cos t + 11\sin t) + 10t - 8$

31. $e^{-t} + u(t-\pi)[1.2\cos t - 3.6\sin t + 2e^{-t+\pi} - 0.8e^{2t-2\pi}]$

33. $0\ (0 \leqq t \leqq 2),\quad 1 - 2e^{-(t-2)} + e^{-2(t-2)}\quad (t > 2)$

35. $y_1 = 4e^t - e^{-2t},\quad y_2 = e^t - e^{-2t}$

37. $y_1 = \cos t - u(t-\pi)\sin t + 2u(t-2\pi)\sin^2 \frac{1}{2}t,$
$y_2 = -\sin t - 2u(t-\pi)\cos^2\frac{1}{2}t + u(t-2\pi)\sin t$

39. $y_1 = (1/\sqrt{10})\sin\sqrt{10}t,\quad y_2 = -(1/\sqrt{10})\sin\sqrt{10}t$

41. $1 - e^{-t}\ (0 < t < 4),\quad (e^4 - 1)e^{-t}\quad (t > 4)$

43. $i(t) = e^{-4t}(\frac{3}{26}\cos 3t - \frac{10}{39}\sin 3t) - \frac{3}{26}\cos 10t + \frac{8}{65}\sin 10t$

45. $5i_1' + 20(i_1 - i_2) = 60,\quad 30i_2' + 20(i_2' - i_1') + 20i_2 = 0,$
$i_1 = -8e^{-2t} + 5e^{-0.8t} + 3,\quad i_2 = -4e^{-2t} + 4e^{-0.8t}$

Problem Set 7.1, page 261

3. $3\times 3,\quad 3\times 4,\quad 3\times 6,\quad 2\times 2,\quad 2\times 3,\quad 3\times 2$

5. $\mathbf{B} = \frac{1}{5}\mathbf{A},\quad \frac{1}{10}\mathbf{A}$

7. No, no, yes, no, no

9. $\begin{bmatrix} 0 & 6 & 12 \\ 18 & 15 & 15 \\ 3 & 0 & -9 \end{bmatrix},\quad \begin{bmatrix} 0 & 2.5 & 1 \\ 2.5 & 1.5 & 2 \\ -1 & 2 & -1 \end{bmatrix},\quad \begin{bmatrix} 0 & 8.5 & 13 \\ 20.5 & 16.5 & 17 \\ 2 & 2 & -10 \end{bmatrix},$ undefined

11. $\begin{bmatrix} 0 & 26 \\ 34 & 32 \\ 28 & -10 \end{bmatrix},$ same, $\begin{bmatrix} 5.4 & 0.6 \\ -4.2 & 2.4 \\ -0.6 & 0.6 \end{bmatrix},$ same

13. $\begin{bmatrix} 70 & 28 \\ -28 & 56 \\ 14 & 0 \end{bmatrix},$ same, $-\mathbf{D}$, undefined

15. $\begin{bmatrix} 5.5 \\ 33.0 \\ -11.0 \end{bmatrix},$ same, undefined, undefined

17. $\begin{bmatrix} -4.5 \\ -27.0 \\ 9.0 \end{bmatrix}$

Problem Set 7.2, page 270

5. $10,\ n(n+1)/2$

7. $\mathbf{0},\quad \mathbf{I},\quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},\quad \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} 10 & -14 & -6 \\ -5 & 7 & -12 \\ -5 & -1 & -4 \end{bmatrix}$, same, $\begin{bmatrix} 10 & -5 & -15 \\ -14 & 7 & -33 \\ -2 & -4 & -4 \end{bmatrix}$, same

13. $\begin{bmatrix} 1 & 2 & 0 \\ 2 & 13 & -6 \\ 0 & -6 & 4 \end{bmatrix}$, $\begin{bmatrix} -9 & -5 \\ 3 & -1 \\ 4 & 0 \end{bmatrix}$, undefined, $\begin{bmatrix} -9 & 3 & 4 \\ -5 & -1 & 0 \end{bmatrix}$

15. Undefined, $\begin{bmatrix} 8 \\ -4 \\ -3 \end{bmatrix}$, $[7 \quad -1 \quad 3]$, same

17. $\begin{bmatrix} -30 & -18 \\ 45 & 9 \\ 5 & -7 \end{bmatrix}$, undefined, $\begin{bmatrix} 22 \\ 4 \\ -12 \end{bmatrix}$, undefined

19. Undefined, $\begin{bmatrix} 10.5 \\ 0 \\ -3 \end{bmatrix}$, $\begin{bmatrix} 7 \\ -3 \\ 1 \end{bmatrix}$, same

25. **(d)** $\mathbf{AB} = (\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} = \mathbf{BA}$; etc.
(e) *Answer.* If $\mathbf{AB} = -\mathbf{BA}$.
29. $\mathbf{p} = [85 \quad 62 \quad 30]^{\mathsf{T}}$, $\mathbf{v} = [44{,}920 \quad 30{,}940]^{\mathsf{T}}$

Problem Set 7.3, page 280

1. $x = -2$, $y = 0.5$
3. $x = 1$, $y = 3$, $z = -5$
5. $x = 6$, $y = -7$
7. $x = -3t$, $y = t$ arb., $z = 2t$
9. $x = 3t - 1$, $y = -t + 4$, $z = t$ arb.
11. $w = 1$, $x = t_1$ arb., $y = 2t_2 - t_1$, $z = t_2$ arb.
13. $w = 4$, $x = 0$, $y = 2$, $z = 6$
17. $I_1 = 2$, $I_2 = 6$, $I_3 = 8$
19. $I_1 = (R_1 + R_2)E_0/(R_1R_2)$ A, $I_2 = E_0/R_1$ A, $I_3 = E_0/R_2$ A
21. $x_2 = 1600 - x_1$, $x_3 = 600 + x_1$, $x_4 = 1000 - x_1$. No
23. C: $3x_1 - x_3 = 0$, H: $8x_1 - 2x_4 = 0$, O: $2x_2 - 2x_3 - x_4 = 0$, thus $C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$

Problem Set 7.4, page 287

1. 1; $[2 \quad -1 \quad 3]$; $[2 \quad -1]^{\mathsf{T}}$
3. 3; $\{[3 \quad 5 \quad 0],\ [0 \quad 3 \quad 5],\ [0 \quad 0 \quad 1]\}$
5. 3; $\{[2 \quad -1 \quad 4],\ [0 \quad 1 \quad -46],\ [0 \quad 0 \quad 1]\}$; $\{[2 \quad 0 \quad 1],\ [0 \quad 3 \quad 23],\ [0 \quad 0 \quad 1]\}$

7. 2; [8 0 4 0], [0 2 0 4]; [8 0 4], [0 2 0]

9. 3; [9 0 1 0], [0 9 8 9], [0 0 1 0]

11. **(c)** 1

17. No

19. Yes

21. No

23. Yes

25. Yes

27. 2, [−2 0 1], [0 2 1]

29. No

31. No

33. 1, solution of the given system $c[1 \quad \frac{10}{3} \quad 3]$, basis $[1 \quad \frac{10}{3} \quad 3]$

35. 1, $[4 \quad 2 \quad \frac{4}{3} \quad 1]$

Problem Set 7.7, page 300

7. $\cos(\alpha + \beta)$

9. 1

11. 40

13. 289

15. −64

17. 2

19. 2

21. $x = 3.5, \quad y = -1.0$

23. $x = 0, \quad y = 4, \quad z = -1$

25. $w = 3, \quad x = 0, \quad y = 2, \quad z = -2$

Problem Set 7.8, page 308

1. $\begin{bmatrix} 1.20 & 4.64 \\ 0.50 & 3.60 \end{bmatrix}$

3. $\begin{bmatrix} 54 & 0.9 & -3.4 \\ 2 & 0.2 & -0.2 \\ -30 & -0.5 & 2 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -4 & 1 \end{bmatrix}$

7. $\mathbf{A}^{-1} = \mathbf{A}$

9. $\begin{bmatrix} 0 & 0 & \frac{1}{2} \\ \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$

11. $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2 = \begin{bmatrix} 3.760 & 22.272 \\ 2.400 & 15.280 \end{bmatrix}$

15. $\mathbf{AA}^{-1} = \mathbf{I}, \quad (\mathbf{AA}^{-1})^{-1} = (\mathbf{A}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{I}$. Multiply by $\mathbf{A}$ from the right.

Problem Set 7.9, page 318

1. $[1 \quad 0]^T, \quad [0 \quad 1]^T; \quad [1 \quad 0]^T, \quad [0 \quad -1]^T; \quad [1 \quad 1]^T, \quad [-1 \quad 1]^T$

3. 1, $[1 \quad 11 \quad -7]^T$

5. No

7. Dimension 2, basis xe^{-x}, e^{-x}

9. 3; basis $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$

11. $x_1 = 5y_1 - y_2, \quad x_2 = 3y_1 - y_2$

13. $x_1 = 2y_1 - 3y_2, \quad x_2 = -10y_1 + 16y_2 + y_3, \quad x_3 = -7y_1 + 11y_2 + y_3$

15. $\sqrt{26}$

17. $\sqrt{5}$

19. 1

21. $k = -20$

23. $\mathbf{a} = [3 \quad 1 \quad -4]^\mathsf{T}, \quad \mathbf{b} = [-4 \quad 8 \quad -1]^\mathsf{T}, \quad \|\mathbf{a} + \mathbf{b}\| = \sqrt{107} \leqq 5.099 + 9$

25. $\mathbf{a} = [5 \quad 3 \quad 2]^\mathsf{T}, \quad \mathbf{b} = [3 \quad 2 \quad -1]^\mathsf{T}, \quad 90 + 14 = 2(38 + 14)$

Chapter 7 Review Questions and Problems, page 318

11. $\begin{bmatrix} -1 & 6 & 1 \\ -18 & 8 & -7 \\ -13 & -2 & -7 \end{bmatrix}, \quad \begin{bmatrix} 1 & 18 & 13 \\ -6 & -8 & 2 \\ -1 & 7 & 7 \end{bmatrix}$

13. $[21 \quad -8 \quad -31]^\mathsf{T}, \quad [21 \quad -8 \quad 31]$

15. 197, 0

17. $-5, \quad \det \mathbf{A}^2 = (\det \mathbf{A})^2 = 25, \quad 0$

19. $\begin{bmatrix} -2 & -12 & -12 \\ -12 & 16 & -9 \\ -12 & -9 & -14 \end{bmatrix}$

21. $x = 4, \quad y = -2, \quad z = 8$

23. $x = 6, \quad y = 2t + 2, \quad z = t$ arb.

25. $x = 0.4, \quad y = -1.3, \quad z = 1.7$

27. $x = 10, \quad y = -2$

29. Ranks 2, 2, ∞

31. Ranks 2, 2, 1

33. $I_1 = 16.5$ A, $\quad I_2 = 11$ A, $\quad I_3 = 5.5$ A

35. $I_1 = 4$ A, $\quad I_2 = 5$ A, $\quad I_3 = 1$ A

Problem Set 8.1, page 329

1. $3, [1 \quad 0]^\mathsf{T}; \quad -0.6, [0 \quad 1]^\mathsf{T}$

3. $-4, [2 \quad 9]^\mathsf{T}; \quad 3, [1 \quad 1]^\mathsf{T}$

5. $-3i, [1 \quad -i]; \quad 3i, [1 \quad i], i = \sqrt{-1}$

7. $\lambda^2 = 0, \quad [1 \quad 0]^\mathsf{T}$

9. $0.8 + 0.6i, [1 \quad -i]^\mathsf{T}; \quad 0.8 - 0.6i, [1 \quad i]^\mathsf{T}$

11. $-(\lambda^3 - 18\lambda^2 + 99\lambda - 162)/(\lambda - 3) = -(\lambda^2 - 15\lambda + 54); \quad 3, [2 \quad -2 \quad 1]^\mathsf{T};$
$6, [1 \quad 2 \quad 2]^\mathsf{T}; \quad 9, [2 \quad 1 \quad -2]^\mathsf{T}$

13. $-(\lambda - 9)^3; \quad 9, [2 \quad -2 \quad 1]^\mathsf{T}$, defect 2

15. $(\lambda + 1)^2(\lambda^2 + 2\lambda - 15); \quad -1, [1 \quad 0 \quad 0 \quad 0]^\mathsf{T}, [0 \quad 1 \quad 0 \quad 0]^\mathsf{T};$
$-5, [-3 \quad -3 \quad 1 \quad 1]^\mathsf{T}, 3, [3 \quad -3 \quad 1 \quad -1]^\mathsf{T}$

17. $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Eigenvalues $i, -i$. Corresponding eigenvectors are complex, indicating that no direction is preserved under a rotation.

19. $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}; \quad 1, \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \quad 0, \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. A point onto the x_2-axis goes onto itself, a point on the x_1-axis onto the origin.

23. Use that real entries imply real coefficients of the characteristic polynomial.

Problem Set 8.2, page 333

1. $1.5,\ [1 \quad -1]^\mathsf{T},\ -45°;\quad 4.5,\ [1 \quad 1]^\mathsf{T},\ 45°$

3. $1,\ [-1/\sqrt{6} \quad 1]^\mathsf{T},\ 112.2°;\quad 8,\ [1 \quad 1/\sqrt{6}]^\mathsf{T},\ 22.2°$

5. $0.5,\ [1 \quad -1]^\mathsf{T};\quad 1.5,\ [1 \quad 1]^\mathsf{T};$ directions $-45°$ and $45°$

7. $[5 \quad 8]^\mathsf{T}$

9. $[11 \quad 12 \quad 16]^\mathsf{T}$

11. 1.8

13. $c[10 \quad 18 \quad 25]^\mathsf{T}$

15. $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y} = [0.6747 \quad 0.7128 \quad 0.7543]^\mathsf{T}$

17. $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j\ (\mathbf{x}_j \neq 0),\quad (\mathbf{A} - k\mathbf{I})\mathbf{x}_j = \lambda_j\mathbf{x}_j - k\mathbf{x}_j = (\lambda_j - k)\mathbf{x}_j.$

19. From $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j\ (\mathbf{x}_j \neq \mathbf{0})$ and Prob. 18 follows $k_p\mathbf{A}^p\mathbf{x}_j = k_p\lambda_j^p\mathbf{x}_j$ and $k_q\mathbf{A}^q\mathbf{x}_j = k_q\lambda_j^q\mathbf{x}_j$ ($p \geqq 0,\ q \geqq 0$, integer). Adding on both sides, we see that $k_p\mathbf{A}^p + k_q\mathbf{A}^q$ has the eigenvalue $k_p\lambda_j^p + k_q\lambda_j^q$. From this the statement follows.

Problem Set 8.3, page 338

1. $0.8 \pm 0.6i,\ [1 \quad \pm i]^\mathsf{T};$ orthogonal

3. $2 \pm 0.8i,\ [1 \quad \pm i]$. Not skew–symmetric!

5. $1,\ [0 \quad 2 \quad 1]^\mathsf{T};\quad 6,\ [1 \quad 0 \quad 0]^\mathsf{T},\quad [0 \quad 1 \quad -2]^\mathsf{T};$ symmetric

7. $0,\quad \pm 25i,$ skew–symmetric

9. $1,\ [0 \quad 1 \quad 0]^\mathsf{T};\quad i,\ [1 \quad 0 \quad i]^\mathsf{T};\quad -i,\ [1 \quad 0 \quad -i]^\mathsf{T},$ orthogonal

15. No

17. $\mathbf{A}^{-1} = (-\mathbf{A}^\mathsf{T})^{-1} = -(\mathbf{A}^{-1})^\mathsf{T}$

19. No since $\det \mathbf{A} = \det(\mathbf{A}^\mathsf{T}) = \det(-\mathbf{A}) = (-1)^3 \det(\mathbf{A}) = -\det(\mathbf{A}) = 0$.

Problem Set 8.4, page 345

1. $\begin{bmatrix} -25 & 12 \\ -50 & 25 \end{bmatrix},\quad -5,\ \begin{bmatrix} 3 \\ 5 \end{bmatrix};\quad 5,\ \begin{bmatrix} 2 \\ 5 \end{bmatrix};\quad \mathbf{x} = \begin{bmatrix} -2 \\ 4 \end{bmatrix},\quad \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

3. $\begin{bmatrix} 3.008 & -0.544 \\ 5.456 & 6.992 \end{bmatrix},\quad 4,\ \begin{bmatrix} -17 \\ 31 \end{bmatrix};\quad 6,\ \begin{bmatrix} -2 \\ 11 \end{bmatrix};\quad \mathbf{x} = \begin{bmatrix} 25 \\ 25 \end{bmatrix},\quad \begin{bmatrix} 10 \\ 5 \end{bmatrix}$

5. $\begin{bmatrix} 4 & 3 & -9 \\ 0 & -5 & 15 \\ 0 & -5 & 15 \end{bmatrix},\quad 0,\ \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix};\quad 4,\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix};\quad 10,\ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix};\quad \mathbf{x} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix},\quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},\quad \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

9. $\begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ -\frac{2}{5} & \frac{1}{5} \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} -2 & 1 \\ 3 & -1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -5 \end{bmatrix}$

13. $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

15. $\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & -2 & 0 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$

17. $\mathbf{C} = \begin{bmatrix} 7 & 3 \\ 3 & 7 \end{bmatrix}$, $4y_1^2 + 10y_2^2 = 200$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \mathbf{y}$, ellipse

19. $\mathbf{C} = \begin{bmatrix} 3 & 11 \\ 11 & 3 \end{bmatrix}$, $14y_1^2 - 8y_2^2 = 0$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{y}$; pair of straight lines

21. $\mathbf{C} = \begin{bmatrix} 1 & -6 \\ -6 & 1 \end{bmatrix}$, $7y_1^2 - 5y_2^2 = 70$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{y}$, hyperbola

23. $\mathbf{C} = \begin{bmatrix} -11 & 42 \\ 42 & 24 \end{bmatrix}$, $52y_1^2 - 39y_2^2 = 156$, $\mathbf{x} = \frac{1}{\sqrt{13}} \begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} \mathbf{y}$, hyperbola

Problem Set 8.5, page 351

1. Hermitian, 5, $[-i \quad 1]^\mathsf{T}$, 7, $[i \quad 1]^\mathsf{T}$
3. Unitary, $(1 - i\sqrt{3})/2$, $[-1 \quad 1]^\mathsf{T}$; $(1 + i\sqrt{3})/2$, $[1 \quad 1]^\mathsf{T}$
5. Skew-Hermitian, unitary, $-i$, $[0 \quad -1 \quad 1]^\mathsf{T}$, i, $[1 \quad 0 \quad 0]^\mathsf{T}$, $[0 \quad 1 \quad 1]^\mathsf{T}$
7. Eigenvalues -1, 1; eigenvectors $[1 \quad -1]^\mathsf{T}$, $[1 \quad 1]^\mathsf{T}$; $[1 \quad -i]^\mathsf{T}$, $[1 \quad i]^\mathsf{T}$; $[0 \quad 1]^\mathsf{T}$, $[1 \quad 0]^\mathsf{T}$, resp.
9. Hermitian, 16
11. Skew-Hermitian, $-6i$
13. $\overline{(\mathbf{ABC})}^\mathsf{T} = \overline{\mathbf{C}}^\mathsf{T}\overline{\mathbf{B}}^\mathsf{T}\overline{\mathbf{A}}^\mathsf{T} = \mathbf{C}^{-1}(-\mathbf{B})\mathbf{A}$
15. $\mathbf{A} = \mathbf{H} + \mathbf{S}$, $\mathbf{H} = \frac{1}{2}(\mathbf{A} + \overline{\mathbf{A}}^\mathsf{T})$, $\mathbf{S} = \frac{1}{2}(\mathbf{A} - \overline{\mathbf{A}}^\mathsf{T})$ (**H** Hermitian, **S** skew-Hermitian)
19. $\mathbf{A}\overline{\mathbf{A}}^\mathsf{T} - \overline{\mathbf{A}}^\mathsf{T}\mathbf{A} = (\mathbf{H} + \mathbf{S})(\mathbf{H} - \mathbf{S}) - (\mathbf{H} - \mathbf{S})(\mathbf{H} + \mathbf{S}) = 2(-\mathbf{HS} + \mathbf{SH}) = \mathbf{0}$ if and only if $\mathbf{HS} = \mathbf{SH}$.

Chapter 8 Review Questions and Problems, page 352

11. 3, $[1 \quad 1]^\mathsf{T}$; 2, $[1 \quad -1]^\mathsf{T}$
13. 3, $[1 \quad 5]^\mathsf{T}$; 7, $[1 \quad 1]^\mathsf{T}$
15. 0, $[2 \quad -2 \quad 1]^\mathsf{T}$; $9i$, $[-1 + 3i \quad 1 + 3i \quad 4]^\mathsf{T}$; $-9i$, $[-1 - 3i \quad 1 - 3i \quad 4]^\mathsf{T}$

17. $-1, 1$; $\mathbf{A} = \frac{1}{16} \begin{bmatrix} 5 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 23 & 2 \\ 39 & 1 \end{bmatrix} = \frac{1}{8} \begin{bmatrix} -1 & 1 \\ 63 & 1 \end{bmatrix}$

19. $\frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \mathbf{A} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -0.9 & 0 \\ 0 & 0.6 \end{bmatrix}$

21. $\frac{1}{3}\begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -20 & 0 \\ 0 & 0 & 22 \end{bmatrix}$

23. $\mathbf{C} = \begin{bmatrix} 4 & 12 \\ 12 & -14 \end{bmatrix}$, $10y_1^2 - 20y_2^2 = 20$, $\mathbf{x} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}\mathbf{y}$, hyperbola

25. $\mathbf{C} = \begin{bmatrix} 3.7 & 1.6 \\ 1.6 & 1.3 \end{bmatrix}$, $4.5y_1^2 + 0.5y_2^2 = 4.5$, $\mathbf{x} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}\mathbf{y}$, ellipse

Problem Set 9.1, page 360

1. 5, 1, 0; $\sqrt{26}$; $[5/\sqrt{26}, 1/\sqrt{26}, 0]$
3. 8.5, −4.0, 1.7; $\sqrt{91.14}$, [0.890, −0.419, 0.178]
5. 2, 1, −2; $\mathbf{u} = [\frac{2}{3}, \frac{1}{3}, -\frac{2}{3}]$, position vector of Q
7. Q: $(4, 0, \frac{1}{2})$, $|\mathbf{v}| = \sqrt{16.25}$
9. Q: (0, 0, −8), $|\mathbf{v}| = 8$
11. [6, 4, 0], $[\frac{3}{2}, 1, 0]$, [−3, −2, 0]
13. [1, 5, 8]
15. 7[9, −7, 8] = [63, −49, 56]
17. [12, 8, 0]
21. [4, 9, −3], $\sqrt{106}$
23. [0, 0, 5], 5
25. [6, 2, −14] = 2**u**, $\sqrt{236}$
27. $\mathbf{p} = [0, 0, -5]$
29. $\mathbf{v} = [v_1, v_2, 3]$, v_1, v_2 arbitrary
31. $k = 10$
33. $|\mathbf{p} + \mathbf{q} + \mathbf{u}| \leqq 18$. Nothing
35. $v_B - v_A = [-19, 0] - [22/\sqrt{2}, 22/\sqrt{2}] = [-19 - 22/\sqrt{2}, -22/\sqrt{2}]$
37. $\mathbf{u} + \mathbf{v} + \mathbf{p} = [-k, 0] + [l, l] + [0, -1000] = \mathbf{0}$, $-k + l + 0 = 0$,
$0 + l - 1000 = 0$, $l = 1000, k = 1000$

Problem Set 9.2, page 367

1. 44, 44, 0
3. $\sqrt{35}$, $\sqrt{320}$, $\sqrt{86}$
5. $|[2, 9, 9]| = \sqrt{166} = 12.88 < \sqrt{80} + \sqrt{86} = 18.22$
7. $|-24| = 24$, $|a||c| = \sqrt{35}\sqrt{86} = \sqrt{3010} = 54.86$; cf. (6)
9. 300; cf. (5a) and (5b)
13. Use (1) and $|\cos\gamma| \leqq 1$.
15. $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = \mathbf{a}\cdot\mathbf{a} + 2\mathbf{a}\cdot\mathbf{b} + \mathbf{b}\cdot\mathbf{b} + (\mathbf{a}\cdot\mathbf{a} - 2\mathbf{a}\cdot\mathbf{b} + \mathbf{b}\cdot\mathbf{b})$
$= 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$
17. [2, 5, 0] • [2, 2, 2] = 14
19. [0, 4, 3] • [−3, −2, 1] = −5 is negative! Why?
21. Yes, because $W = (\mathbf{p} + \mathbf{q})\cdot\mathbf{d} = \mathbf{p}\cdot\mathbf{d} + \mathbf{q}\cdot\mathbf{d}$.
23. arccos 0.5976 = 53.3°
27. $\beta - \alpha$ is the angle between the unit vectors **a** and **b**. Use (2).
29. $\gamma = \arccos(12/(6\sqrt{13})) = 0.9828 = 56.3°$ and 123.7°
31. $a_1 = -\frac{28}{3}$
33. $\pm[\frac{3}{5}, -\frac{4}{5}]$
35. $(\mathbf{a} + \mathbf{b})\cdot(\mathbf{a} - \mathbf{b}) = |\mathbf{a}|^2 - |\mathbf{b}|^2 = 0$, $|\mathbf{a}| = |\mathbf{b}|$. A square.
37. 0. Why?
39. If $|\mathbf{a}| = |\mathbf{b}|$ or if **a** and **b** are orthogonal.

Problem Set 9.3, page 374

5. $-\mathbf{m}$ instead of $\mathbf{m}$, tendency to rotate in the opposite sense.

7. $|\mathbf{v}| = |[0, 20, 0] \times [8, 6, 0]| = |[0, 0, -160]| = 160$

9. Zero volume in Fig. 191, which can happen in several ways.

11. $[0, 0, 7]$, $[0, 0, -7]$, -4

13. $[6, 2, 7]$, $[-6, -2, -7]$

15. $\mathbf{0}$

17. $[-32, -58, 34]$, $[-42, -63, 19]$

19. 1, -1

21. $[-48, -72, -168]$, $12\sqrt{248} = 189.0, 189.0$

23. 0, 0, 13

25. $\mathbf{m} = [-2, -2, 0] \times [2, 3, 0] = [0, 0, -10]$, $m = 10$ clockwise

27. $[6, 2, 0] \times [1, 2, 0] = [0, 0, 10]$

29. $\frac{1}{2}|[-12, 2, 6]| = \sqrt{46}$

31. $3x + 2y - z = 5$

33. $474/6 = 79$

Problem Set 9.4, page 380

1. Hyperbolas

3. Parallel straight lines (planes in space) $y = \frac{3}{4}x + c$

5. Circles, centers on the y-axis

7. Ellipses

9. Parallel planes

11. Elliptic cylinders

13. Paraboloids

Problem Set 9.5, page 390

1. Circle, center $(3, 0)$, radius 2

3. Cubic parabola $x = 0$, $z = y^3$

5. Ellipse

7. Helix

9. A "Lissajous curve"

11. $\mathbf{r} = [3 + \sqrt{13}\cos t, 2 + \sqrt{13}\sin t, 1]$

13. $\mathbf{r} = [2 + t, 1 + 2t, 3]$

15. $\mathbf{r} = [t, 4t - 1, 5t]$

17. $\mathbf{r} = [\sqrt{2}\cos t, \sin t, \sin t]$

19. $\mathbf{r} = [\cosh t, (\sqrt{3}/2)\sinh t, -2]$

21. Use $\sin(-\alpha) = -\sin\alpha$.

25. $\mathbf{u} = [-\sin t, 0, \cos t]$. At P, $\mathbf{r}' = [-8, 0, 6]$. $\mathbf{q}(w) = [6 - 8w, i, 8 + 6w]$.

27. $\mathbf{q}(w) = [2 + w, \frac{1}{2} - \frac{1}{4}w, 0]$

29. $\sqrt{\mathbf{r}' \bullet \mathbf{r}'} = \cosh t$, $l = \sinh l = 1.175$

31. $\sqrt{\mathbf{r}' \bullet \mathbf{r}'} = a$, $l = a\pi/2$

33. Start from $\mathbf{r}(t) = [t, f(t)]$.

35. $\mathbf{v} = \mathbf{r}' = [1, 2t, 0]$, $|\mathbf{v}| = \sqrt{1 + 4t^2}$, $\mathbf{a} = [0, 2, 0]$

37. $\mathbf{v}(0) = (\omega + 1)R\mathbf{i}$, $\mathbf{a}(0) = -\omega^2 R\mathbf{j}$

39. $\mathbf{v} = [-\sin t - 2\sin 2t, \cos t - 2\cos 2t]$, $|\mathbf{v}|^2 = 5 - 4\cos 3t$,

$$\mathbf{a} = [-\cos t - 4\cos 2t, -\sin t + 4\sin 2t], \text{ and } \mathbf{a}_{\tan} = \frac{6\sin 3t}{5 - 4\cos 3t}\mathbf{v}.$$

41. $\mathbf{v} = [-\sin t, 2\cos 2t, -2\sin 2t]$, $|\mathbf{v}|^2 = 4 + \sin^2 t$,

$$\mathbf{a} = [-\cos t, -4\sin 2t, -4\cos 2t], \text{ and } \mathbf{a}_{\tan} = \frac{\frac{1}{2}\sin 2t}{4 + \sin^2 t}\mathbf{v}.$$

43. 1 year $= 365 \cdot 86{,}400$ sec, $R = 30 \cdot 365 \cdot 86{,}400/2\pi = 151 \cdot 10^6$ [km], $|\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R = 5.98 \cdot 10^{-6}$ [km/sec^2]

45. $R = 3960 + 80$ mi $= 2.133 \cdot 10^7$ ft, $g = |\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R$, $|\mathbf{v}| = \sqrt{gR} = \sqrt{6.61 \cdot 10^8} = 25{,}700$ [ft/sec] $= 17{,}500$ [mph]

49. $\mathbf{r}(t) = [t, y(t), 0]$, $\mathbf{r}' = [1, y', 0]$ $\mathbf{r} \bullet \mathbf{r}' = 1 + y'^2$, etc.

51. $\dfrac{d\mathbf{r}}{ds} = \dfrac{d\mathbf{r}}{dt} \Big/ \dfrac{ds}{dt}, \qquad \dfrac{d^2\mathbf{r}}{ds^2} = \dfrac{d^2\mathbf{r}}{dt^2} \Big/ \left(\dfrac{ds}{dt}\right)^2 + \cdots, \qquad \dfrac{d^3\mathbf{r}}{ds^3} = \dfrac{d^3\mathbf{r}}{dt^3} \Big/ \left(\dfrac{ds}{dt}\right)^3 + \cdots$

53. $3/(1 + 9t^2 + 9t^4)$

Problem Set 9.7, page 402

1. $[2y - 1, 2x + 2]$
3. $[-y/x^2, 1/x]$
5. $[4x^3, 4y^3]$
7. Use the chain rule.
9. Apply the quotient rule to each component and collect terms.
11. $[y, x], \quad [5, -4]$
13. $[2x/(x^2 + y^2), 2y/(x^2 + y^2)], \quad [0.16, 0.12]$
15. $[8x, 18y, 2z], \quad [40, -18, -22]$
17. For P on the x- and y-axes.
19. $[-1.25, 0]$
21. $[0, -e]$
23. Points with $y = 0, \pm\pi, \pm 2\pi, \cdots$.
25. $-\nabla T(P) = [0, 4, -1]$
31. $\nabla f = [32x, -2y], \quad \nabla f(P) = [160, -2]$
33. $[12x, 4y, 2z], \quad [60, 20, 10]$
35. $[-2x, -2y, 1], \quad [-6, -8, 1]$
37. $[2, 1] \bullet [1, -1]/\sqrt{5} = 1/\sqrt{5}$
39. $[1, 1, 1] \bullet [-3/125, 0, -4/125]/\sqrt{3} = -7/(125\sqrt{3})$
41. $\sqrt{8/3}$
43. $f = xyz$
45. $f = \int v_1\,dx + \int v_2\,dy + \int v_3\,dz$

Problem Set 9.8, page 405

1. $2x + 8y + 18z; \quad 7$
3. 0, after simplification; solenoidal
5. $9x^2y^2z^2; \quad 1296$
7. $-2e^x (\cos y)z$
9. **(b)** $(fv_1)_x + (fv_2)_y + (fv_3)_z = f[(v_1)_x + (v_2)_y + (v_3)_z] + f_xv_1 + f_yv_2 + f_zv_3$, etc.
11. $[v_1, v_2, v_3] = \mathbf{r}' = [x', y', z'] = [y, 0, 0], \quad z' = 0, z = c_3, \quad y' = 0, y = c_2$, and $x' = y = c_2, x = c_2t + c_1$. Hence as t increases from 0 to 1, this "shear flow" transforms the cube into a parallelepiped of volume 1.
13. div $(\mathbf{w} \times \mathbf{r}) = 0$ because v_1, v_2, v_3 do not depend on x, y, z, respectively.
15. $-2 \cos 2x + 2 \cos 2y$
17. 0
19. $2/(x^2 + y^2 + z^2)^2$

Problem Set 9.9, page 408

3. Use the definitions and direct calculation.
5. $[x(z^2 - y^2), y(x^2 - z^2), z(y^2 - x^2)]$
7. $e^{-x}[\cos y, \sin y, 0]$
9. curl $\mathbf{v} = [-6z, 0, 0]$ incompressible, $\mathbf{v} = \mathbf{r}' = [x', y', z'] = [0, 3z^2, 0], \quad x = c_1$, $z = c_3, \quad y' = 3z^2 = 3c_3^2, \quad y = 3c_3^2t + c_2$
11. curl $\mathbf{v} = [0, 0, -3]$, incompressible, $x' = y, \quad y' = -2x, \quad 2xx' + yy' = 0$, $x^2 + \frac{1}{2}y^2 = c, z = c_3$
13. curl $\mathbf{v} = 0$, irrotational, div $\mathbf{v} = 1$, compressible, $\mathbf{r} = [c_1e^t, c_2e^t, c_3e^{-t}]$. Sketch it.
15. $[-1, -1, -1]$, same (why?)
17. $-yz - zx - xy$, 0 (why?), $-y - z - x$
19. $[-2z - y, -2x - z, -2y - x]$, same (why?)

Chapter 9 Review Questions and Problems, page 409

11. $-10, \quad 1080, \quad 1080, \quad 65$
13. $[-10, -30, 0], \quad [10, 30, 0], \quad \mathbf{0}, \quad 40$
15. $[-1260, -1830, -300], \quad [-210, 120, -540]$, undefined
17. $-125, \quad 125, \quad -125$
19. $[70, -40, -50], \quad 0, \quad \sqrt{35^2 + 20^2 + 25^2} = \sqrt{2250}$
21. $[-2, -6, -13]$
23. $\gamma_1 = \arccos(-10/\sqrt{65 \cdot 40}) = 1.7682 = -101.3°$, $\gamma_2 = 23.7°$
25. $[5, 2, 0] \bullet [4 - 1, 3 - 1, 0] = 19$
27. $\mathbf{v} \bullet \mathbf{w}/|\mathbf{w}| = 22/\sqrt{8} = 7.78$
29. $[0, 0, -14]$, tendency of clockwise rotation
31. 4
33. $1, \quad -2y$
35. 0, same (why?), $2(y^2 + x^2 - xz)$
37. $[0, -2, 0]$
39. $9/\sqrt{225} = \frac{3}{5}$

Problem Set 10.1, page 418

3. 4
5. $\mathbf{r} = [2\cos t, \quad 2\sin t], \quad 0 \leqq t \leqq \pi/2; \quad \frac{8}{5}$
7. "Exponential helix," $(e^{6\pi} - 1)/3$
9. $23.5, \quad 0$
11. $2e^{-t} + 2te^{-t^2}, \quad -2e^{-2} - e^{-4} + 3$
15. $18\pi, \quad \frac{4}{3}(4\pi)^3, \quad 18\pi$
17. $[4\cos t, \quad + \sin t, \quad \sin t, \quad 4\cos t], \quad [2, 2, 0]$
19. $144t^4, \quad 1843.2$

Problem Set 10.2, page 425

3. $\sin \frac{1}{2}x \cos 2y, \quad 1 - 1/\sqrt{2} = 0.293$
5. $e^{xy}\sin z, \quad e - 0$
7. $\cosh 1 - 2 = -0.457$
9. $e^x \cosh y + e^z \sinh y, \quad e - (\cosh 1 + \sinh 1) = 0$
13. $e^{a^2}\cos 2b$
15. Dependent, $x^2 \neq -4y^2$, etc.
17. Dependent, $4 \neq 0$, etc.
19. $\sin(a^2 + 2b^2 + c^2)$

Problem Set 10.3, page 432

3. $8y^3/3, \quad 54$
5. $\displaystyle\int_0^1 [x - x^3 - (x^2 - x^5)]\, dx = \frac{1}{12}$
7. $\cosh 2x - \cosh x, \quad \frac{1}{2}\sinh 4 - \sinh 2$
9. $36 + 27y^2, \quad 144$
11. $z = 1 - r^2, \quad dx\,dy = r\,dr\,d\theta$, *Answer:* $\pi/2$
13. $\bar{x} = 2b/3, \quad \bar{y} = h/3$
15. $\bar{x} = 0, \quad \bar{y} = 4r/3\pi$
17. $I_x = bh^3/12, \quad I_y = b^3h/4$
19. $I_x = (a + b)h^3/24, \quad I_y = h(a^4 - b^4)/(48(a - b))$

Problem Set 10.4, page 438

1. $(-1 - 1) \cdot \pi/4 = -\pi/2$
3. $9(e^2 - 1) - \frac{8}{3}(e^3 - 1)$
5. $2x - 2y, \quad 2x(1 - x^2) - (2 - x^2)^2 + 1, \quad x = -1 \cdots 1, \quad -\frac{56}{15}$
7. 0. Why?
9. $\frac{16}{5}$
13. $\nabla^2 w = \cosh x, \quad y = x/2 \cdots 2, \quad \frac{1}{2}\cosh 4 - \frac{1}{2}$

15. $\nabla^2 w = 6xy, \quad 3x(10 - x^2)^2 - 3x, \quad 486$ **17.** $\nabla^2 w = 6x - 6y, \quad -38.4$
19. $|\text{grad } w|^2 = e^{2x}, \quad \frac{5}{2}(e^4 - 1)$

Problem Set 10.5, page 442

1. Straight lines, $\mathbf{k}$
3. $z = c\sqrt{x^2 + y^2}$, circles, straight lines, $[-cu \cos v, \quad -cu \sin v, \quad u]$
5. $z = x^2 + y^2$, circles, parabolas, $[-2u^2 \cos v, \quad -2u^2 \sin v, \quad u]$
7. $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1, \quad [bc \cos^2 v \cos u, \quad ac \cos^2 v \sin u, \quad ab \sin v \cos v]$, ellipses
11. $[\tilde{u}, \quad \tilde{v}, \quad \tilde{u}^2, \quad + \tilde{v}^2], \quad \tilde{\mathbf{N}} = [-2\tilde{u}, \quad -2\tilde{v}, \quad 1]$
13. Set $x = u$ and $y = v$.
15. $[2 + 5 \cos u, \quad -1 + 5 \sin u, \quad v], \quad [5 \cos u, \quad 5 \sin u, \quad 0]$
17. $[a \cos v \cos u, \quad -2.8 + a \cos v \sin u, \quad 3.2 + a \sin v], \quad a = 1.5;$
$[a^2 \cos^2 v \cos u, \quad a^2 \cos^2 v \sin u, \quad a^2 \cos v \sin v]$
19. $[\cosh u, \quad \sinh u, \quad v], \quad [\cosh u, \quad -\sinh u, \quad 0]$

Problem Set 10.6, page 450

1. $\mathbf{F}(\mathbf{r}) \cdot \mathbf{N} = [-u^2, \quad v^2, \quad 0] \cdot [-3, \quad 2, \quad 1] = 3u^2 + 2v^2, \quad 29.5$
3. $\mathbf{F}(\mathbf{r}) \cdot \mathbf{N} = \cos^3 v \cos u \sin u$ from (3), Sec. 10.5. *Answer*: $\frac{1}{3}$
5. $\mathbf{F}(\mathbf{r}) \cdot \mathbf{N} = -u^3, -128\pi$
7. $\mathbf{F} \cdot \mathbf{N} = [0, \quad \sin u, \quad \cos v] \cdot [1, -2u, 0], \quad 4 + (-2 + \pi^2/16 - \pi/2)\sqrt{2} = -0.1775$
9. $\mathbf{r} = [2 \cos u, \quad 2 \sin u, \quad v], \quad 0 \leqq u \leqq \pi/4, \quad 0 \leqq v \leqq 5$. Integrate $2 \sinh v \sin u$ to get $2(1 - 1/\sqrt{2})(\cosh 5 - 1) = 42.885$.
13. $7\pi^3/\sqrt{6} = 88.6$
15. $G(\mathbf{r}) = (1 + 9u^4)^{3/2}, \quad |\mathbf{N}| = (1 + 9u^4)^{1/2}$. *Answer*: 54.4
21. $I_{x=y} = \iint_S [\frac{1}{2}(x - y)^2 + z^2]\, \sigma\, dA$
23. $[u \cos v, \quad u \sin v, \quad u], \quad \int_0^{2\pi} \int_0^h u^2 \cdot u\sqrt{2}\, du\, dv = \dfrac{\pi}{\sqrt{2}} h^4$
25. $[\cos u \cos v, \quad \cos u \sin v, \quad \sin u], \quad dA = (\cos u)\, du\, dv$, B the z-axis, $\quad I_B = 8\pi/3$, $I_K = I_B + 1^2 \cdot 4\pi = 20.9$.

Problem Set 10.7, page 457

1. 224
3. $-e^{-1-z} + e^{-y-z}, \quad -2e^{-1-z} + e^{-z}, \quad 2e^{-3} - e^{-2} - 2e^{-1} + 1$
5. $\frac{1}{2}(\sin 2x)(1 - \cos 2x), \quad \frac{1}{8}, \quad \frac{3}{4}$
7. $[r \cos u \cos v, \quad \cos u \sin v, \quad r \sin u], \quad dV = r^2 \cos u\, dr\, du\, dv, \quad \sigma = v, \quad 2\pi^2 a^3/3$
9. $\text{div } \mathbf{F} = 2x + 2z$, 48
11. $12(e - 1/e) = 24 \sinh 1$
13. $\text{div } \mathbf{F} = -\sin z$, 0
15. $1/\pi + \frac{5}{24} = 0.5266$
17. $h^4\pi/2$
19. $8abc(b^2 + c^2)/3$
21. $(a^4/4) \cdot 2\pi \cdot h = ha^4\pi/2$
23. $h^5\pi/10$
25. Do Prob. 20 as the last one.

Problem Set 10.8, page 462

1. $x = 0, y = 0, z = 0$, no contributions. $x = a$: $\partial f/\partial n = \partial f/\partial x = -2x = -2a$, etc. Integrals $x = a$: $(-2a)bc$, $y = b$: $(-2b)ac$, $z = c$: $(4c)\,ab$. Sum 0

3. The volume integral of $8y^2 + [0, 8y] \cdot [2x, 0] = 8y^2$ is $8y^3/3 = \frac{8}{3}$. The surface integral of $f\partial g/\partial n = f \cdot 2x = 2f = 8y^2$ over $x = 1$ is $8y^3/3 = \frac{8}{3}$. Others 0.

5. The volume integral of $6y^2 \cdot 4 - 2x^2 \cdot 12$ is 0; $8(x = 1)$, $-8(y = 1)$, others 0.

7. $\mathbf{F} = [x, \quad 0, \quad 0]$, $\operatorname{div} \mathbf{F} = 1$, use (2*), Sec. 10.7, etc.

9. $z = 0$ and $z = \sqrt{a^2 - x^2 - y^2} = \sqrt{a^2 - r^2}$, $dx\,dy = r\,dr\,d\theta$, $-2\pi \cdot \frac{1}{2}(a^2 - r^2)^{3/2} \cdot \frac{2}{3}\big|_0^a = \frac{2}{3}\pi a^3$

11. $r = a$, $\phi = 0$, $\cos\phi = 1$, $v = \frac{1}{3}a \cdot (4\pi a^2)$

Problem Set 10.9, page 468

1. S: $z = y\ (0 \leqq x \leqq 1, 0 \leqq y \leqq 4)$, $[0, 2z, -2z] \bullet [0, -1, 1]$, ± 20

3. $[2e^{-z}\cos y, \quad -e^{-z}, \quad 0] \bullet [0, \quad -y, \quad 1] = ye^{-z}$, $\pm(2 - 2/\sqrt{e})$

5. $[0, 2z, \frac{3}{2}] \bullet [0, 0, 1] = \frac{3}{2}$, $\pm\frac{3}{2}a^2$

7. $[-e^z, \quad -e^x, \quad -e^y] \bullet [-2x, \quad 0, \quad 1]$, $\pm(e^4 - 2e + 1)$

9. The sides contribute a, $3a^2/2$, $-a$, 0.

11. -2π; curl $\mathbf{F} = \mathbf{0}$

13. $5\mathbf{k}$, 80π

15. $[0, -1, 2x - 2y] \bullet [0, 0, 1]$, $\frac{1}{3}$

17. $\mathbf{r} = [\cos u, \quad \sin u, \quad v]$, $[-3v^2, \quad 0, \quad 0] \bullet [\cos u, \quad \sin u, \quad 0]$, -1

19. $\mathbf{r} = [u\cos v, \quad u \sin v, \quad u]$, $0 \leqq u \leqq 1$, $0 \leqq v \leqq \pi/2$, $[-e^z, \quad 1, \quad 0] \bullet [-u\cos v, \quad -u\sin v, \quad u]$. *Answer*: 1/2

Chapter 10 Review Questions and Problems, page 469

11. $\mathbf{r} = [4 - 10t, \quad 2 + 8t]$, $\mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = [2(4 - 10t)^2, \quad -4(2t + 8t)^2] \bullet [-10, \quad 8]\,dt$; $-4528/3$. Or using exactness.

13. Not exact, curl $\mathbf{F} = (5\cos x)\mathbf{k}$, ± 10

15. 0 since curl $\mathbf{F} = \mathbf{0}$

17. By Stokes, $\pm 18\pi$

19. $\mathbf{F} = \operatorname{grad}(y^2 + xz)$, 2π

21. $M = 8$, $\bar{x} = \frac{8}{5}$, $\bar{y} = \frac{16}{5}$

23. $M = \frac{63}{20}$, $\bar{x} = \frac{8}{7} = 1.14$, $\bar{y} = \frac{118}{49} = 2.41$

25. $M = 4k/15$, $\bar{x} = \frac{5}{16}$, $\bar{y} = \frac{4}{7}$

27. $288(a + b + c)\pi$

29. $\operatorname{div} \mathbf{F} = 20 + 6z^2$. *Answer*: 21

31. $24 \sinh 1 = 28.205$

33. Direct integration, $\frac{224}{3}$

35. 72π

Problem Set 11.1, page 482

1. $2\pi, 2\pi, \pi, \pi, 1, 1, \frac{1}{2}, \frac{1}{2}$

5. There is no *smallest* $p > 0$.

13. $\dfrac{4}{\pi}\left(\cos x + \dfrac{1}{9}\cos 3x + \dfrac{1}{25}\cos 5x + \cdots\right) + 2\left(\sin x + \dfrac{1}{3}\sin 3x + \dfrac{1}{5}\sin 5x + \cdots\right)$

15. $\frac{4}{3}\pi^2 + 4\left(\cos x + \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x + \cdots\right) - 4\pi\left(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \cdots\right)$

17. $\dfrac{\pi}{2} + \dfrac{4}{\pi}\left(\cos x + \dfrac{1}{9}\cos 3x + \dfrac{1}{25}\cos 5x + \cdots\right)$

19. $\dfrac{\pi}{4} - \dfrac{2}{\pi}\left(\cos x + \dfrac{1}{9}\cos 3x + \dfrac{1}{25}\cos 5x + \cdots\right) + \sin x - \dfrac{1}{2}\sin 2x + \dfrac{1}{3}\sin 3x - + \cdots$

21. $2\left(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \frac{1}{4}\sin 4x + \frac{1}{5}\sin 5x + \cdots\right)$

Problem Set 11.2, page 490

1. Neither, even, odd, odd, neither **3.** Even **5.** Even

9. Odd, $L = 2$, $\dfrac{4}{\pi}\left(\sin\dfrac{\pi x}{2} + \dfrac{1}{3}\sin\dfrac{3\pi x}{2} + \dfrac{1}{5}\sin\dfrac{5\pi x}{2} + \cdots\right)$

11. Even, $L = 1$, $\dfrac{1}{3} - \dfrac{4}{\pi^2}\left(\cos \pi x - \dfrac{1}{4}\cos 2\pi x + \dfrac{1}{9}\cos 3\pi x - + \cdots\right)$

13. Rectifier, $L = \dfrac{1}{2}$, $\dfrac{1}{8} - \dfrac{1}{\pi^2}\left(\cos 2\pi x + \dfrac{1}{9}\cos 6\pi x + \dfrac{1}{25}\cos 10\pi x + \cdots\right) + \dfrac{1}{\pi}\left(\dfrac{1}{2}\sin 2\pi x - \dfrac{1}{4}\sin 4\pi x + \dfrac{1}{6}\sin 6\pi x - \dfrac{1}{8}\sin 8\pi x + - \cdots\right)$

15. Odd, $L = \pi$, $\dfrac{4}{\pi}\left(\sin x - \dfrac{1}{9}\sin 3x + \dfrac{1}{25}\sin 5x - + \cdots\right)$

17. Even, $L = 1$, $\dfrac{1}{2} + \dfrac{4}{\pi^2}\left(\cos \pi x + \dfrac{1}{9}\cos 3\pi x + \dfrac{1}{25}\cos 5\pi x + \cdots\right)$

19. $\frac{3}{8} + \frac{1}{2}\cos 2x + \frac{1}{8}\cos 4x$

23. $L = 4$, **(a)** 1, **(b)** $\dfrac{4}{\pi}\left(\sin\dfrac{\pi x}{4} + \dfrac{1}{3}\sin\dfrac{3\pi x}{4} + \dfrac{1}{5}\sin\dfrac{5\pi x}{4} + \cdots\right)$

25. $L = \pi$, **(a)** $\dfrac{\pi}{2} + \dfrac{4}{\pi}\left(\cos x + \dfrac{1}{9}\cos 3x + \dfrac{1}{25}\cos 5x + \cdots\right)$,
(b) $2\left(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \frac{1}{4}\sin 4x + \cdots\right)$

27. $L = \pi$, **(a)** $\dfrac{3\pi}{8} + \dfrac{2}{\pi}\left(\cos x - \dfrac{1}{2}\cos 2x + \dfrac{1}{9}\cos 3x + \dfrac{1}{25}\cos 5x - \dfrac{1}{18}\cos 6x + \dfrac{1}{49}\cos 7x + \dfrac{1}{81}\cos 9x - \dfrac{1}{50}\cos 10x + \dfrac{1}{121}\cos 11x + \cdots\right)$
(b) $\left(1 + \dfrac{2}{\pi}\right)\sin x + \dfrac{1}{2}\sin 2x + \left(\dfrac{1}{3} - \dfrac{2}{9\pi}\right)\sin 3x + \dfrac{1}{4}\sin 4x + \left(\dfrac{1}{5} + \dfrac{2}{25\pi}\right)\sin 5x + \dfrac{1}{6}\sin 6x + \cdots$

29. Rectifier, $L = \pi$,
(a) $\dfrac{2}{\pi} - \dfrac{4}{\pi}\left(\dfrac{1}{1\cdot 3}\cos x + \dfrac{1}{3\cdot 5}\cos 3x + \dfrac{1}{5\cdot 7}\cos 5x + \cdots\right)$, **(b)** $\sin x$

Problem Set 11.3, page 494

3. The output becomes a pure cosine series.

5. For A_n this is similar to Fig. 54 in Sec. 2.8, whereas for the phase shift B_n the sense is the same for all n.

7. $y = C_1 \cos \omega t + C_2 \sin \omega t + a(\omega) \sin t, \quad a(\omega) = 1/(\omega^2 - 1) = -1.33,$ $-5.26, 4.76, 0.8, 0.01$. Note the change of sign.

11. $y = C_1 \cos \omega t + C_2 \sin \omega t + \frac{4}{\pi}\left(\frac{1}{\omega^2 - 9} \sin t + \frac{1}{\omega^2 - 49} \sin 3t + \frac{1}{\omega^2 - 121} \sin 5t + \cdots\right)$

13. $y = \sum_{n=1}^{N} (A_n \cos nt + B_n \sin nt), \quad A_n = [(1 - n^2)a_n - nb_n c]/D_n,$
$B_n = [(1 - n^2)b_n + nca_n]/D_n, \quad D_n = (1 - n^2)^2 + n^2c^2$

15. $b_n = (-1)^{n+1} \cdot 12/n^3$ (n odd), $\quad y = \sum_{n=1}^{\infty} (A_n \cos nt + B_n \sin nt),$
$A_n = (-1)^n \cdot 12nc/n^3D_n, \quad B_n = (-1)^{n+1} \cdot 12(1 - n^2)/(n^3D_n)$ with D_n as in Prob. 13.

17. $I = 50 + A_1 \cos t + B_1 \sin t + A_3 \cos 3t + B_3 \sin 3t + \cdots, A_n = (10 - n^2)a_n/D_n,$
$B_n = 10na_n/D_n, \quad a_n = -400/(n^2\pi), D_n = (n^2 - 10)^2 + 100n^2$

19. $I(t) = \sum_{n=1}^{\infty} (A_n \cos nt + B_n \sin nt), \quad A_n = (-1)^{n+1} \frac{2400(10 - n^2)}{n^2D_n},$
$B_n = (-1)^{n+1} \frac{24{,}000}{nD_n}, \quad D_n = (10 - n^2)^2 + 100n^2$

Section 11.4, page 498

3. $F = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{1}{9} \cos 3x + \frac{1}{25} \cos 5x + \cdots\right), E^* = 0.0748,$ $0.0748, 0.0119, 0.0119, 0.0037$

5. $F = \frac{4}{\pi}\left(\sin x + \frac{1}{3} \sin 3x + \frac{1}{5} \sin 5x + \cdots\right), E^* = 1.1902, 1.1902, 0.6243, 0.6243,$ $0.4206 \quad (0.1272$ when $N = 20)$

7. $F = 2[(\pi^2 - 6) \sin x - \frac{1}{8}(4\pi^2 - 6) \sin 2x + \frac{1}{27}(9\pi^2 - 6) \sin 3x - + \cdots];$
$E^* = 674.8, 454.7, 336.4, 265.6, 219.0$. Why is E^* so large?

Section 11.5, page 503

3. Set $x = ct + k$.

5. $x = \cos \theta, dx = -\sin \theta \, d\theta$, etc.

7. $\lambda_m = (m\pi/10)^2, m = 1, 2, \cdots; y_m = \sin(m\pi x/10)$

9. $\lambda = [(2m + 1)\pi/(2L)]^2, m = 0, 1, \cdots, y_m = \sin((2m + 1)\pi x/(2L))$

11. $\lambda_m = m^2, m = 1, 2, \cdots, y_m = x \sin(m \ln |x|)$

13. $p = e^{8x}, q = 0, r = e^{8x}, \lambda_m = m^2, y_m = e^{-4x} \sin mx, m = 1, 2, \cdots$

Section 11.6, page 509

1. $8(P_1(x) - P_3(x) + P_5(x))$

3. $\frac{4}{5}P_0(x) - \frac{4}{7}P_2(x) - \frac{8}{35}P_4(x)$

9. $-0.4775P_1(x) - 0.6908P_3(x) + 1.844P_5(x) - 0.8236P_7(x) + 0.1658P_9(x) + \cdots,$
$m_0 = 9$. ***Rounding*** seems to have considerable influence in Probs. 8–13.

11. $0.7854P_0(x) - 0.3540P_2(x) + 0.0830P_4(x) - \cdots, m_0 = 4$
13. $0.1212P_0(x) - 0.7955P_2(x) + 0.9600P_4(x) - 0.3360P_6(x) + \cdots, m_0 = 8$
15. (c) $a_m = (2/J_1^2(\alpha_{0,m}))(J_1(\alpha_{0,m})/\alpha_{0,m}) = 2/(\alpha_{0,m}J_1(\alpha_{0,m}))$

Section 11.7, page 517

1. $f(x) = \pi e^{-x} (x > 0)$ gives $A = \displaystyle\int_0^\infty e^{-v} \cos wv \, dv = \frac{1}{1 + w^2}$, $B = \dfrac{w}{1 + w^2}$ (see Example 3), etc.

3. Use (11); $B = \dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{\pi}{2} \sin wv \, dv = \frac{1 - \cos \pi w}{w}$

5. $B(w) = \dfrac{2}{\pi} \displaystyle\int_0^1 \frac{1}{2} \pi v \sin wv \, dv = \frac{\sin w - w \cos w}{w^2}$

7. $\dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{\sin w \cos xw}{w} \, dw$

9. $A(w) = \dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{\cos wv}{1 + v^2} \, dv = e^{-w} \, (w > 0)$

11. $\dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{\cos \pi w + 1}{1 - w^2} \cos xw \, dw$

15. For $n = 1, 2, 11, 12, 31, 32, 49, 50$ the value of $\mathrm{Si}(n\pi) - \pi/2$ equals $0.28, -0.15, 0.029, -0.026, 0.0103, -0.0099, 0.0065, -0.0064$ (rounded).

17. $\dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{1 - \cos w}{w} \sin xw \, dw$

19. $\dfrac{2}{\pi} \displaystyle\int_0^\infty \frac{w - e(w \cos w - \sin w)}{1 + w^2} \sin xw \, dw$

Section 11.8, page 522

1. $\hat{f}_c(w) = \sqrt{(2/\pi)}\,(2 \sin w - \sin 2w)/w$
3. $\hat{f}_c(w) = \sqrt{(2/\pi)}\,(\cos 2w + 2w \sin 2w - 1)/w^2$
5. $\hat{f}_c(w) = \sqrt{\dfrac{2}{\pi}}\, \dfrac{(w^2 - 2) \sin w + 2w \cos w}{w^3}$
7. Yes. No
9. $\sqrt{2/\pi}\, w/(a^2 + w^2)$
11. $\sqrt{2/\pi}\,((2 - w^2) \cos w + 2w \sin w - 2)/w^3$
13. $\mathcal{F}_s(e^{-x}) = \dfrac{1}{w}\left(-\mathcal{F}_c(e^{-x}) + \sqrt{\dfrac{2}{\pi}} \cdot 1\right) = \dfrac{1}{w}\left(\sqrt{\dfrac{2}{\pi}} \cdot \dfrac{1}{w^2 + 1} + \sqrt{\dfrac{2}{\pi}}\right) = \sqrt{\dfrac{2}{\pi}}\, \dfrac{w}{w^2 + 1}$

Problem Set 11.9, page 533

3. $i(e^{-ibw} - e^{-iaw})/(w\sqrt{2\pi})$ if $a < b$; 0 otherwise
5. $[e^{(1-iw)a} - e^{-(1-iw)a}]/(\sqrt{2\pi}(1 - iw))$
7. $(e^{-iaw}(1 + iaw) - 1)/(\sqrt{2\pi} w^2)$
9. $\sqrt{2/\pi}(\cos w + w \sin w - 1)/w^2$
11. $i\sqrt{2/\pi}\,(\cos w - 1)/w$
13. $e^{-w^2/2}$ by formula 9

17. No, the assumptions in Theorem 3 are not satisfied.

19. $[f_1 + f_2 + f_3 + f_4,\quad f_1 - if_2 - f_3 + if_4,\quad f_1 - f_2 + f_3 - f_4,\quad f_1 + if_2 - f_3 - if_4]$

21. $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_1 + f_2 \\ f_1 - f_2 \end{bmatrix}$

Chapter 11 Review Questions and Problems, page 537

11. $1 + \dfrac{4}{\pi}\left(\sin\dfrac{\pi x}{2} + \dfrac{1}{3}\sin\dfrac{3\pi x}{2} + \dfrac{1}{5}\sin\dfrac{5\pi x}{2} + \cdots\right)$

13. $\dfrac{1}{4} - \dfrac{2}{\pi^2}\left(\cos \pi x + \dfrac{1}{9}\cos 3\pi x + \dfrac{1}{25}\cos 5\pi x + \cdots\right) +$
$\dfrac{1}{\pi}\left(\sin \pi x - \dfrac{1}{2}\sin 2\pi x + \dfrac{1}{3}\sin 3\pi x - + \cdots\right)$

15. $\cosh x$, $\sinh x$ $(-5 < x < 5)$, respectively

17. Cf. Sec. 11.1.

19. $\dfrac{1}{2} - \dfrac{4}{\pi^2}\left(\cos \pi x + \dfrac{1}{9}\cos 3\pi x + \cdots\right),\quad \dfrac{2}{\pi}\left(\sin \pi x - \dfrac{1}{2}\sin 2\pi x + - \cdots\right)$

21. $y = C_1 \cos \omega t + C_2 \sin \omega t + \dfrac{\pi^2}{\omega^2} - 12\left(\dfrac{\cos t}{\omega^2 - 1} - \dfrac{1}{4}\cdot\dfrac{\cos 2t}{\omega^2 - 4} + \dfrac{1}{9}\cdot\dfrac{\cos 3t}{\omega^2 - 9}\right.$
$\left. - \dfrac{1}{16}\cdot\dfrac{\cos 4t}{\omega^2 - 16} + - \cdots\right)$

23. 0.82, 0.50, 0.36, 0.28, 0.23

25. 0.0076, 0.0076, 0.0012, 0.0012, 0.0004

27. $\dfrac{1}{\pi}\displaystyle\int_0^\infty \frac{(\cos w + w \sin w - 1)\cos wx + (\sin w - w\cos w)\sin wx}{w^2}\,dw$

29. $\sqrt{2/\pi}\,(\cos aw - \cos w + aw \sin aw - w \sin w)/w^2$

Problem Set 12.1, page 542

1. $L(c_1u_1 + c_2u_2) = c_1L(u_1) + c_2L(u_2) = c_1 \cdot 0 + c_2 \cdot 0 = 0$

3. $c = 2$

5. $c = a/b$

7. Any c and ω

9. $c = \pi/25$

15. $u = 110 - (110/\ln 100)\ln(x^2 + y^2)$

17. $u = a(y)\cos 4\pi x + b(y)\sin 4\pi x$

19. $u = c(x)\,e^{-y^3/3}$

21. $u = e^{-3y}(a(x)\cos 2y + b(x)\sin 2y) + 0.1e^{3y}$

23. $u = c_1(y)x + c_2(y)/x^2$ (Euler–Cauchy)

25. $u(x, y) = axy + bx + cy + k$; a, b, c, k arbitrary constants

Problem Set 12.3, page 551

5. $k \cos 3\pi t \sin 3\pi x$

7. $\dfrac{8k}{\pi^3}\left(\cos \pi t \sin \pi x + \dfrac{1}{27}\cos 3\pi t \sin 3\pi x + \dfrac{1}{125}\cos 5\pi t \sin 5\pi x + \cdots\right)$

9. $\dfrac{0.8}{\pi^2}\left(\cos \pi t \sin \pi x - \dfrac{1}{9}\cos 3\pi t \sin 3\pi x + \dfrac{1}{25}\cos 5\pi t \sin 5\pi x - + \cdots\right)$

11. $\dfrac{2}{\pi^2}\left((2-\sqrt{2})\cos \pi t \sin \pi x - \dfrac{1}{9}(2+\sqrt{2})\cos 3\pi t \sin 3\pi x\right.$

$\left. + \dfrac{1}{25}(2+\sqrt{2})\cos 5\pi t \sin 5\pi x - + \cdots\right)$

13. $\dfrac{4}{\pi^3}\left((4-\pi)\cos \pi t \sin \pi x + \cos 2\pi t \sin 2\pi x + \dfrac{4+3\pi}{27}\cos 3\pi t \sin 3\pi x\right.$

$\left. + \dfrac{4-5\pi}{125}\cos 5\pi t \sin 5\pi x + \cdots\right)$. No terms with $n = 4, 8, 12, \cdots$.

17. $u = \dfrac{8L^2}{\pi^3}\left(\cos\left[c\left(\dfrac{\pi}{L}\right)^2 t\right]\sin\dfrac{\pi x}{L} + \dfrac{1}{3^3}\cos\left[c\left(\dfrac{3\pi}{L}\right)^2 t\right]\sin\dfrac{3\pi x}{L} + \cdots\right)$

19. **(a)** $u(0, t) = 0$, **(b)** $u(L, t) = 0$, **(c)** $u_x(0, t) = 0$, **(d)** $u_x(L, t) = 0$. $C = -A, D = -B$ from (a), (c). Insert this. The coefficient determinant resulting from (b), (d) must be zero to have a nontrivial solution. This gives (22).

Problem Set 12.4, page 556

3. $c^2 = 300/[0.9/(2 \cdot 9.80)] = 80.83^2\ [\mathrm{m^2/sec^2}]$

9. Elliptic, $u = f_1(y + 2ix) + f_2(y - 2ix)$

11. Parabolic, $u = xf_1(x - y) + f_2(x - y)$

13. Hyperbolic, $u = f_1(y - 4x) + f_2(y - x)$

15. Hyperbolic, $xy'^2 + yy' = 0, y = v, xy = w, u_w = z, u = \dfrac{1}{y}f_1(xy) + f_2(y)$

17. Elliptic, $u = f_1(y - (2 - i)x) + f_2(y - (2 + i)x)$. Real or imaginary parts of any function u of this form are solutions. Why?

Problem Set 12.6, page 566

3. $u_1 = \sin x\, e^{-t}$, $u_2 = \sin 2x\, e^{-4t}$, $u_3 = \sin 3x\, e^{-9t}$ differ in rapidity of decay.

5. $u = \sin 0.1\pi x\, e^{-1.752\pi^2 t/100}$

7. $u = \dfrac{800}{\pi^3}\left(\sin 0.1\pi x\, e^{-0.01752\pi^2 t} + \dfrac{1}{3^3}\sin 0.3\pi x\, e^{-0.01752(3\pi)^2 t} + \cdots\right)$

9. $u = u_{\mathrm{I}} + u_{\mathrm{II}}$, where $u_{\mathrm{II}} = u - u_{\mathrm{I}}$ satisfies the boundary conditions of the text,

so that $u_{\mathrm{II}} = \displaystyle\sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L}\, e^{-(cn\pi/L)^2 t}$, $B_n = \dfrac{2}{L}\displaystyle\int_0^L [f(x) - u_{\mathrm{I}}(x)]\sin\frac{n\pi x}{L}\,dx$.

11. $F = A\cos px + B\sin px$, $F'(0) = Bp = 0$, $B = 0$, $F'(L) = -Ap\sin pL = 0$, $p = n\pi/L$, etc.

13. $u = 1$

15. $\dfrac{1}{2} + \dfrac{4}{\pi^2}\left(\cos x\, e^{-t} + \dfrac{1}{9}\cos 3x\, e^{-9t} + \dfrac{1}{25}\cos 5x\, e^{-25t} + \cdots\right)$

17. $-\dfrac{K\pi}{L}\displaystyle\sum_{n=1}^{\infty} nB_n\, e^{-\lambda_n^2 t}$

19. $u = 1000\,(\sin\frac{1}{2}\pi x \sinh\frac{1}{2}\pi y)/\sinh \pi$

21. $u = \dfrac{100}{\pi}\displaystyle\sum_{n=1}^{\infty}\frac{1}{(2n-1)\sinh(2n-1)\pi}\sin\frac{(2n-1)\pi x}{24}\sinh\frac{(2n-1)\pi y}{24}$

23. $u = A_0 x + \sum_{n=1}^{\infty} A_n \frac{\sinh(n\pi x/24)}{\sinh n\pi} \cos\frac{n\pi y}{24}$,

$$A_0 = \frac{1}{24^2}\int_0^{24} f(y)\,dy, \quad A_n = \frac{1}{12}\int_0^{24} f(y)\cos\frac{n\pi y}{24}\,dy$$

25. $\sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{a}\sinh\frac{n\pi(b-y)}{a},\ A_n = \frac{2}{a\sinh(n\pi b/a)}\int_0^a f(x)\sin\frac{n\pi x}{a}\,dx$

Problem Set 12.7, page 574

3. $A = \frac{2}{\pi}\int_0^{\infty}\frac{\cos pv}{1+v^2}\,dv = \frac{2}{\pi}\cdot\frac{\pi}{2}e^{-p},\ u = \int_0^{\infty} e^{-p-c^2p^2t}\cos px\,dp$

5. $A = \frac{2}{\pi}\int_0^1 v\cos pv\,dv = \frac{2}{\pi}\cdot\frac{\cos p + p\sin p - 1}{p^2}$, etc.

7. $A = \frac{2}{\pi}\int_0^{\infty}\frac{\sin v}{v}\cos pv\,dv = \frac{2}{\pi}\cdot\frac{\pi}{2} = 1$ if $0 < p < 1$ and 0 if $p > 1$,

$$u = \int_0^1 \cos px\, e^{-c^2p^2t}\,dp$$

9. Set $w = -v$ in (21) to get $\operatorname{erf}(-x) = -\operatorname{erf} x$.

13. In (12) the argument $x + 2cz\sqrt{t}$ is 0 (the point where f jumps) when $z = -x/(2c\sqrt{t})$. This gives the lower limit of integration.

15. Set $w = s/\sqrt{2}$ in (21).

Problem Set 12.9, page 584

1. **(a)**, **(b)** It is multiplied by $\sqrt{2}$. **(c)** Half

5. $B_{mn} = (-1)^{n+1}8/(mn\pi^2)$ if m odd, 0 if m even

7. $B_{mn} = (-1)^{m+n}4ab/(mn\pi^2)$

11. $u = 0.1\cos\sqrt{20}t\sin 2x\sin 4y$

13. $\frac{6.4}{\pi^2}\sum_{\substack{m=1\\ m,n}}^{\infty}\sum_{\substack{n=1\\ \text{odd}}}^{\infty}\frac{1}{m^3n^3}\cos\left(t\sqrt{m^2+n^2}\right)\sin mx\sin ny$

17. $c\pi\sqrt{260}$ (corresponding eigenfunctions $F_{4,16}$ and $F_{16,14}$), etc.

19. $\cos\left(\pi t\sqrt{\frac{36}{a^2}+\frac{4}{b^2}}\right)\sin\frac{6\pi x}{a}\sin\frac{4\pi y}{b}$

Problem Set 12.10, page 591

5. $110 + \frac{440}{\pi}\left(r\cos\theta - \frac{1}{3}r^3\cos 3\theta + \frac{1}{5}r^5\cos 5\theta - + \cdots\right)$

7. $55\pi - \frac{440}{\pi}\left(r\cos\theta + \frac{1}{9}r^3\cos 3\theta + \frac{1}{25}r^5\cos 5\theta + \cdots\right)$

11. Solve the problem in the disk $r < a$ subject to u_0 (given) on the upper semicircle and $-u_0$ on the lower semicircle.

$$u = \frac{4u_0}{\pi}\left(\frac{r}{a}\sin\theta + \frac{1}{3a^3}r^3\sin 3\theta + \frac{1}{5a^5}r^5\sin 5\theta + \cdots\right)$$

13. Increase by a factor $\sqrt{2}$

15. $T = 6.826\rho R^2 f_1^2$

17. No

25. $\alpha_{11}/(2\pi) = 0.6098$; See Table A1 in App. 5.

Problem Set 12.11, page 598

5. $A_4 = A_6 = A_8 = A_{10} = 0, \quad A_5 = 605/16, \quad A_7 = -4125/128, \quad A_9 = 7315/256$

9. $\nabla^2 u = u'' + 2u'/r = 0, \quad u''/u' = -2/r, \quad \ln|u'| = -2\ln|r| + c_1,$
$u' = \tilde{c}/r^2, u = c/r + k$

13. $u = 320/r + 60$ is smaller than the potential in Prob. 12 for $2 < r < 4$.

17. $u = 1$

19. $\cos 2\phi = 2\cos^2\phi - 1, 2w^2 - 1 = \frac{4}{3}P_2(w) - \frac{1}{3}, u = \frac{4}{3}r^2 P_2(\cos\phi) - \frac{1}{3}$

25. Set $1/r = \rho$. Then $u(\rho, \theta, \phi) = rv(r, \theta, \phi), u_\rho = (v + rv_r)(-1/\rho^2)$,
$u_{\rho\rho} = (2v_r + rv_{rr})(1/\rho^4) + (v + rv_r)(2/\rho^3), u_{\rho\rho} + (2/\rho)u_\rho = r^5(v_{rr} + (2/r)v_r)$.
Substitute this and $u_{\phi\phi} = rv_{\phi\phi}$ etc. into (7) [written in terms of ρ] and divide by r^5.

Problem Set 12.12, page 602

5. $W = \dfrac{c(s)}{x^s} + \dfrac{x}{s^2(s+1)}$, $W(0, s) = 0, c(s) = 0, w(x, t) = x(t - 1 + e^{-t})$

7. $w = f(x)g(t), xf'g + f\dot{g} = xt$, take $f(x) = x$ to get $g = ce^{-t} + t - 1$ and $c = 1$ from $w(x, 0) = x(c - 1) = 0$.

11. Set $x^2/(4c^2\tau) = z^2$. Use z as a new variable of integration. Use $\operatorname{erf}(\infty) = 1$.

Chapter 12 Review Questions and Problems, page 603

17. $u = c_1(x)e^{-3y} + c_2(x)e^{2y} - 3$

19. Hyperbolic, $f_1(x) + f_2(y + x)$

21. Hyperbolic, $f_1(y + 2x) + f_2(y - 2x)$

23. $\frac{3}{4}\cos 2t\sin x - \frac{1}{4}\cos 6t\sin 3x$

25. $\sin 0.01\pi x\, e^{-0.001143t}$

27. $\frac{3}{4}\sin 0.01\pi x\, e^{-0.001143t} - \frac{1}{4}\sin 0.03\pi x\, e^{-0.01029t}$

29. $100\cos 2x\, e^{-4t}$

39. $u = (u_1 - u_0)(\ln r)/\ln(r_1/r_0) + (u_0\ln r_1 - u_1\ln r_0)/\ln(r_1/r_0)$

Problem Set 13.1, page 612

1. $1/i = i/i^2 = -i, \quad 1/i^3 = i/i^4 = i$

3. $4.8 - 1.4i$

5. $x - iy = -(x + iy), \quad x = 0$

9. $-117, \quad 4$

11. $-8 - 6i$

13. $-120 - 40i$

15. $3 - i$

17. $-4x^2y^2$

19. $(x^2 - y^2)/(x^2 + y^2), \quad 2xy/(x^2 + y^2)$

Problem Set 13.2, page 618

1. $\sqrt{2}\,(\cos\frac{1}{4}\pi + i\sin\frac{1}{4}\pi)$

3. $2(\cos\frac{1}{2}\pi + i\sin\frac{1}{2}\pi), \quad 2(\cos\frac{1}{2}\pi - i\sin\frac{1}{2}\pi)$

5. $\frac{1}{2}(\cos \pi + i \sin \pi)$

7. $\sqrt{1 + \frac{1}{4}\pi^2}\,(\cos \arctan \frac{1}{2}\pi + i \sin \arctan \frac{1}{2}\pi)$

9. $3\pi/4$

11. $\pm\arctan(\frac{4}{3}) = \pm 0.9273$

13. -1024. *Answer:* π

15. $-3i$

17. $2 + 2i$

21. $\sqrt[6]{2}\,(\cos \frac{1}{12}k\pi + i \sin \frac{1}{12}\pi), \quad k = 1, 9, 17$

23. $6, \quad -3 \pm 3\sqrt{3}i$

25. $\cos(\frac{1}{8}\pi + \frac{1}{2}k\pi) + i \sin(\frac{1}{8}\pi + \frac{1}{2}k\pi), \quad k = 0, 1, 2, 3$

27. $\cos \frac{1}{5}\pi \pm i \sin \frac{1}{5}\pi, \quad \cos \frac{3}{5}\pi \pm i \sin \frac{3}{5}\pi, \quad -1$

29. $i, \quad -1 - i$

31. $\pm(1 - i), \quad \pm(2 + 2i)$

33. $|z_1 + z_2|^2 = (z_1 + z_2)\overline{(z_1 + z_2)} = (z_1 + z_2)(\bar{z}_1 + \bar{z}_2)$. Multiply out and use $\text{Re}\, z_1\bar{z}_2 \leqq |z_1\bar{z}_2|$ (Prob. 34).
$z_1\bar{z}_1 + z_1\bar{z}_2 + z_2\bar{z}_1 + z_2\bar{z}_2 = |z_1|^2 + 2\,\text{Re}\, z_1\bar{z}_2 + |z_2|^2 \leqq |z_1|^2 + 2|z_1||z_2| + |z_2|^2 = (|z_1| + |z_2|)^2$. Hence $|z_1 + z_1|^2 \leqq (|z_1| + |z_2|)^2$. Taking square roots gives (6).

35. $[(x_1 + x_2)^2 + (y_1 + y_2)^2] + [(x_1 - x_2)^2 + (y_1 - y_2)^2] = 2(x_1^2 + y_1^2 + x_2^2 + y_2^2)$

Problem Set 13.3, page 624

1. Closed disk, center $-1 + 5i$, radius $\frac{3}{2}$

3. Annulus (circular ring), center $4 - 2i$, radii π and 3π

5. Domain between the bisecting straight lines of the first quadrant and the fourth quadrant.

7. Half-plane extending from the vertical straight line $x = -1$ to the right.

11. $u(x, y) = (1 - x)/((1 - x)^2 + y^2), \quad u(1, -1) = 0,$
$v(x, y) = y((1 - x)^2 + y^2), \quad v(1, -1) = -1$

15. Yes, since $\text{Im}(|z|^2/z) = \text{Im}(|z|^2\bar{z}/(z\bar{z})) = \text{Im}\,\bar{z} = -r \sin\theta \to 0$.

17. Yes, because $\text{Re}\, z = r\cos\theta \to 0$ and $1 - |z| \to 1$ as $r \to 0$.

19. $f'(z) = 8(z - 4i)^7$. Now $z - 4i = 3$, hence $f'(3 + 4i) = 8 \cdot 3^7 = 17{,}496$.

21. $n(1 - z)^{-n-1}i, \quad ni$

23. $3iz^2/(z + i)^4, \quad -3i/16$

Problem Set 13.4, page 629

1. $r_x = x/r = \cos\theta, \quad r_y = \sin\theta, \quad \theta_x = -(\sin\theta)/r, \quad \theta_y = (\cos\theta)/r$
(a) $0 = u_x - v_y = u_r \cos\theta + u_\theta(-\sin\theta)/r - v_r \sin\theta - v_\theta(\cos\theta)/r$
(b) $0 = u_y + v_x = u_r \sin\theta + u_\theta(\cos\theta)/r + v_r \cos\theta + v_\theta(-\sin\theta)/r$
Multiply (a) by $\cos\theta$, (b) by $\sin\theta$, and add. Etc.

3. Yes

5. No, $f(z) = \overline{(z^2)}$

7. Yes, when $z \neq 0$. Use (7).

9. Yes, when $z \neq 0, \quad -2\pi i, \quad 2\pi i$

11. Yes

13. $f(z) = -\frac{1}{2}i(z^2 + c)$, c real

15. $f(z) = 1/z + c$ (c real)

17. $f(z) = z^2 + z + c$ (c real)

19. No

21. $a = \pi, \quad v = e^{\pi x} \sin \pi y$

23. $a = 0, \quad v = \frac{1}{2}b(y^2 - x^2) + c$

27. $f = u + iv$ implies $if = -v + iu$.

29. Use (4), (5), and (1).

Problem Set 13.5, page 632

3. $e^{2\pi i}e^{-2\pi} = e^{-2\pi} = 0.001867$

5. $e^2(-1) = -7.389$

7. $e^{\sqrt{2}}i = 4.113i$

9. $5e^{i \arctan(3/4)} = 5e^{0.644i}$

11. $6.3e^{\pi i}$

13. $\sqrt{2}e^{\pi i/4}$

15. $\exp(x^2 - y^2)\cos 2xy, \quad \exp(x^2 - y^2)\sin 2xy$
17. $\text{Re}(\exp(z^3)) = \exp(x^3 - 3xy^2)\cos(3x^2y - y^3)$
19. $z = 2n\pi i, \quad n = 0, 1, \cdots$

Problem Set 13.6, page 636

1. Use (11), then (5) for e^{iy}, and simplify.
7. $\cosh 1 = 1.543,\ i \sinh 1 = 1.175i$
9. Both $-0.642 - 1.069i$. Why?
11. $i \sinh \pi = 11.55i$, both
15. Insert the definitions on the left, multiply out, and simplify.
17. $z = \pm(2n + 1)i/2$
19. $z = \pm n\pi i$

Problem Set 13.7, page 640

5. $\ln 11 + \pi i$
7. $\frac{1}{2}\ln 32 - \pi i/4 = 1.733 - 0.785i$
9. $i \arctan(0.8/0.6) = 0.927i$
11. $\ln e + \pi i/2 = 1 + \pi i/2$
13. $\pm 2n\pi i, \quad n = 0, 1, \cdots$
15. $\ln|e^i| + i \arctan \dfrac{\sin 1}{\cos 1} \pm 2n\pi i = 0 + i + 2n\pi i, \quad n = 0, 1, \cdots$
17. $\ln(i^2) = \ln(-1) = (1 \pm 2n)\pi i, \quad 2\ln i = (1 \pm 4n)\pi i, n = 0, 1, \cdots$
19. $e^{4-3i} = e^4(\cos 3 - i \sin 3) = -54.05 - 7.70i$
21. $e^{0.6}e^{0.4i} = e^{0.6}(\cos 0.4 + i \sin 0.4) = 1.678 + 0.710i$
23. $e^{(1-i)\,\text{Ln}\,(1+i)} = e^{\ln\sqrt{2} + \pi i/4 - i\ln\sqrt{2} + \pi/4} = 2.8079 + 1.3179i$
25. $e^{(3-i)(\ln 3 + \pi i)} = 27e^{\pi}(\cos(3\pi - \ln 3) + i \sin(3\pi - \ln 3)) = -284.2 + 556.4i$
27. $e^{(2-i)\,\text{Ln}(-1)} = e^{(2-i)\pi i} = e^{\pi} = 23.14$

Chapter 13 Review Questions and Problems, page 641

1. $2 - 3i$
3. $27.46e^{0.9929i}, \quad 7.616e^{1.976i}$
11. $-5 + 12i$
13. $0.16 - 0.12i$
15. i
17. $4\sqrt{2}e^{-3\pi i/4}$
19. $15e^{-\pi i/2}$
21. $\pm 3, \quad \pm 3i$
23. $(\pm 1 \pm i)/\sqrt{2}$
25. $f(z) = -iz^2/2$
27. $f(z) = e^{-2z}$
29. $f(z) = e^{-z^2/2}$
31. $\cos 3 \cosh 1 + i \sin 3 \sinh 1 = -1.528 + 0.166i$
33. $i \tanh 1 = 0.7616i$
35. $\cosh \pi \cos \pi + i \sinh \pi \sin \pi = -11.592$

Problem Set 14.1, page 651

1. Straight segment from (2, 1) to (5, 2.5).
3. Parabola $y = x^2$ from (1, 2) to (2, 8).
5. Circle through (0, 0), center $(3, -1)$, radius $\sqrt{10}$, oriented clockwise.
7. Semicircle, center 2, radius 4.
9. Cubic parabola $y = x^3 \quad (-2 \leqq x \leqq 2)$
11. $z(t) = t + (2 + t)i \quad (-1 \leqq t \leqq 1)$
13. $z(t) = 2 - i + 2e^{it} \quad (0 \leqq t \leqq \pi)$

15. $z(t) = 2\cosh t + i \sinh t \ (-\infty < t < \infty)$
17. Circle $z(t) = -a - ib + re^{-it} \quad (0 \leqq t \leqq 2\pi)$
19. $z(t) = t + (1 - \frac{1}{4}t^2)i \quad (-2 \leqq t \leqq 2)$
21. $z(t) = (1 + i)t \quad (1 \leqq t \leqq 3)$, $\quad \text{Re}\, z = t$, $\quad z'(t) = 1 + i$. *Answer:* $4 + 4i$
23. $e^{2\pi i} - e^{\pi i} = 1 - (-1) = 2$
25. $\frac{1}{2}\exp z^2|_1^i = \frac{1}{2}(e^{-1} - e^1) = -\sinh 1$
27. $\tan \frac{1}{4}\pi i - \tan \frac{1}{4} = i \tanh \frac{1}{4} - 1$
29. $\text{Im}\, z^2 = 2xy = 0$ on the axes. $z = 1 + (-1 + i)t \quad (0 \leqq t \leqq 1)$, $(\text{Im}\, z^2)\, \dot{z} = 2(1 - t)y(-1 + i)$ integrated: $(-1 + i)/3$.
35. $|\text{Re}\, z| = |x| \leqq 3 = M$ on C, $L = \sqrt{8}$

Problem Set 14.2, page 659

1. Use (12), Sec. 14.1, with $m = 2$. **3.** Yes **5.** 5
7. **(a)** Yes. **(b)** No, we would have to move the contour across $\pm 2i$.
9. 0, yes
11. πi, no
13. 0, yes
15. $-\pi$, no
17. 0, no
19. 0, yes
21. $2\pi i$
23. $1/z + 1/(z - 1)$, hence $2\pi i + 2\pi i = 4\pi i$.
25. 0 (Why?)
27. 0 (Why?)
29. 0

Problem Set 14.3, page 663

1. $2\pi i z^2/(z - 1)|_{z=-1} = -\pi i$
3. 0
5. $2\pi i(\cos 3z)/6|_{z=0} = \pi i/3$
7. $2\pi i(i/2)^3/2 = \pi/8$
11. $2\pi i \cdot \left.\dfrac{1}{z + 2i}\right|_{z=2i} = \dfrac{\pi}{2}$
13. $2\pi i(z + 2)|_{z=2} = 8\pi i$
15. $2\pi i \cosh(-\pi^2 - \pi i) = -2\pi i \cosh \pi^2 = -60{,}739i$ since $\cosh \pi i = \cos \pi = -1$ and $\sinh \pi i = i \sin \pi = 0$.
17. $2\pi i \left.\dfrac{\text{Ln}\,(z + 1)}{z + i}\right|_{z=i} = 2\pi i \dfrac{\text{Ln}\,(1 + i)}{2i} = \pi(\ln\sqrt{2} + i\pi/4) = 1.089 + 2.467i$
19. $2\pi i e^{2i}/(2i) = \pi e^{2i}$

Problem Set 14.4, page 667

1. $(2\pi i/3!)(-\cos 0) = -\pi i/3$
3. $(2\pi i/(n - 1)!)e^0$
5. $\dfrac{2\pi i}{3!}(\cosh 2z)''' = \dfrac{\pi i}{3} \cdot 8 \sinh 1 = 9.845i$
7. $(2\pi i/(2n)!)\,(\cos z)^{(2n)}|_{z=0} = (2\pi i/(2n)!)(-1)^n \cos 0 = (-1)^n 2\pi i/(2n)!$
9. $-2\pi i(\tan \pi z)'\Big|_{z=0} = \left.\dfrac{-2\pi i \cdot \pi}{\cos^2 \pi z}\right|_{z=0} = -2\pi^2 i$
11.
$$\frac{2\pi i}{4}((1 + z)\sin z)'\Big|_{z=1/2} = \tfrac{1}{2}\pi i(\sin z + (1 + z)\cos z)|_{z=1/2}$$
$$= \tfrac{1}{2}\pi i(\sin \tfrac{1}{2} + \tfrac{3}{2}\cos \tfrac{1}{2})$$
$$= 2.821i$$

13. $2\pi i \cdot \left.\frac{1}{z}\right|_{z=2} = \pi i$ **15.** 0. Why?

17. 0 by Cauchy's integral theorem for a doubly connected domain; see (6) in Sec. 14.2.

19. $(2\pi i/2!)4^{-3}(e^{3z})''|_{z=\pi i/4} = -9\pi(1+i)/(64\sqrt{2})$

Chapter 14 Review Questions and Problems, page 668

21. $\frac{1}{2}\cosh(-\frac{1}{4}\pi^2) - \frac{1}{2} = 2.469$

23. $2\pi i(e^z)^{(4)}|_{z=0} = ie^z/12|_{z=0} = \pi i/12$ by Cauchy's integral formula.

25. $-2\pi i(\tan \pi z)'|_{z=1} = -2\pi^2 i/\cos^2 \pi z|_{z=1} = -2\pi^2 i$

27. 0 since $z^2 + \bar{z} - 2 = 2(x^2 - y^2)$ and $y = x$

29. $-4\pi i$

Problem Set 15.1, page 679

1. $z_n = (2i/2)^n$; bounded, divergent, ± 1, $\pm i$

3. $z_n = -\frac{1}{2}\pi i/(1 + 2/(ni))$ by algebra; convergent to $-\pi i/2$

5. Bounded, divergent, $\pm 1 + 10i$

7. Unbounded, hence divergent

9. Convergent to 0, hence bounded

17. Divergent; use $1/\ln n > 1/n$. **19.** Convergent; use $\Sigma 1/n^2$.

21. Convergent **23.** Convergent

25. Divergent

29. By absolute convergence and Cauchy's convergence principle, for given $\epsilon > 0$ we have for every $n > N(\epsilon)$ and $p = 1, 2, \cdots$

$$|z_{n+1}| + \cdots + |z_{n+p}| < \epsilon,$$

hence $|z_{n+1} + \cdots + z_{n+p}| < \epsilon$ by (6*), Sec. 13.2, hence convergence by Cauchy's principle.

Problem Set 15.2, page 684

1. No! Nonnegative integer powers of z (or $z - z_0$) only!

3. At the center, in a disk, in the whole plane

5. $\Sigma a_n z^{2n} = \Sigma a_n (z^2)^n$, $|z^2| < R = \lim |a_n/a_{n+1}|$; hence $|z| < \sqrt{R}$.

7. $\pi/2, \infty$ **9.** $i, \sqrt{3}$ **11.** $0, \sqrt{\frac{26}{5}}$

13. $-i, \frac{1}{2}$ **15.** $2i$, 1 **17.** $1/\sqrt{2}$

Problem Set 15.3, page 689

3. $f = \sqrt[n]{n}$. Apply l'Hôpital's rule to $\ln f = (\ln n)/n$.

5. 2 **7.** $\sqrt{3}$ **9.** $1/\sqrt{2}$

11. $\sqrt{\frac{7}{3}}$ **13.** 1 **15.** $\frac{3}{4}$

Problem Set 15.4, page 697

3. $2z^2 - \frac{(2z^2)^3}{3!} + \cdots = 2z^2 - \frac{4}{3}z^6 + \frac{4}{15}z^{10} - + \cdots, \quad R = \infty$

5. $\frac{1}{2} - \frac{1}{4}z^4 + \frac{1}{8}z^8 - \frac{1}{16}z^{12} + \frac{1}{32}z^{16} - + \cdots, \quad R = \sqrt[4]{2}$

7. $\dfrac{1}{2} + \dfrac{1}{2}\cos z = 1 - \dfrac{1}{2 \cdot 2!}z^2 + \dfrac{1}{2 \cdot 4!}z^4 - \dfrac{1}{2 \cdot 6!}z^6 + - \cdots, \quad R = \infty$

9. $\displaystyle\int_0^z \left(1 - \frac{1}{2}t^2 + \frac{1}{8}t^4 - + \cdots\right) dt = z - \frac{1}{6}z^3 + \frac{1}{40}z^5 - + \cdots, \quad R = \infty$

11. $z^3/(1!3) - z^7/(3!7) + z^{11}/(5!11) - + \cdots, \quad R = \infty$

13. $(2/\sqrt{\pi})(z - z^3/3 + z^5/(2!5) - z^7/(3!7) + \cdots), \quad R = \infty$

17. Team Project. (a) $(\text{Ln}(1 + z))' = 1 - z + z^2 - + \cdots = 1/(1 + z)$.

(c) Use that the terms of $(\sin iy)/(iy)$ are all positive, so that the sum cannot be zero.

19. $\frac{1}{2} + \frac{1}{2}i + \frac{1}{2}i(z - i) + (-\frac{1}{4} + \frac{1}{4}i)(z - i)^2 - \frac{1}{4}(z - i)^3 + \cdots, \quad R = \sqrt{2}$

21. $1 - \dfrac{1}{2!}\left(z - \dfrac{1}{2}\pi\right)^2 + \dfrac{1}{4!}\left(z - \dfrac{1}{2}\pi\right)^4 - \dfrac{1}{6!}\left(z - \dfrac{1}{2}\pi\right)^6 + - \cdots, \quad R = \infty$

23. $-\frac{1}{4} - \frac{2}{8}i(z - i) + \frac{3}{16}(z - i)^2 + \frac{4}{32}i(z - i)^3 - \frac{5}{64}(z - i)^4 + \cdots, \quad R = 2$

25. $2\left(z - \dfrac{1}{2}i\right) + \dfrac{2^3}{3!}\left(z - \dfrac{1}{2}i\right)^3 + \dfrac{2^5}{5!}\left(z - \dfrac{1}{2}i\right)^5 + \cdots, \quad R = \infty$

Problem Set 15.5, page 704

3. $|z + i| \leqq \sqrt{3} - \delta, \quad \delta > 0$

5. $|z + \frac{1}{2}i| \leqq \frac{1}{4} - \delta, \quad \delta > 0$

7. Nowhere

9. $|z - 2i| \leqq 2 - \delta, \quad \delta > 0$

11. $|z^n| \leqq 1$ and $\Sigma 1/n^2$ converges. Use Theorem 5.

13. $|\sin^n |z|| \leqq 1$ for all z, and $\Sigma 1/n^2$ converges. Use Theorem 5.

15. $R = 4$ by Theorem 2 in Sec. 15.2; use Theorem 1.

17. $R = 1/\sqrt{\pi} > 0.56$; use Theorem 1.

Chapter 15 Review Questions and Problems, page 706

11. 1

13. 3

15. $\frac{1}{2}$

17. $\infty, \quad e^{2z}$

19. $\infty, \quad \cosh \sqrt{z}$

21. $\displaystyle\sum_{n=0}^{\infty} \frac{z^{4n}}{(2n + 1)!}, \quad R = \infty$

23. $\displaystyle\frac{1}{2} + \frac{1}{2}\cos 2z = 1 + \frac{1}{2}\sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)!}(2z)^{2n}, \quad R = \infty$

25. $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n!} z^{2n-2}, \quad R = \infty$

27. $\cos[(z - \frac{1}{2}\pi) + \frac{1}{2}\pi] = -(z - \frac{1}{2}\pi) + \frac{1}{6}(z - \frac{1}{2}\pi)^3 - + \cdots = -\sin(z - \frac{1}{2}\pi)$

29. $\ln 3 + \dfrac{1}{3}(z - 3) - \dfrac{1}{2 \cdot 9}(z - 3)^2 + \dfrac{1}{3 \cdot 27}(z - 3)^3 - + \cdots, \quad R = 3$

Problem Set 16.1, page 714

1. $z^{-4} - \frac{1}{2}z^{-2} + \frac{1}{24} - \frac{1}{720}z^2 + - \cdots, \quad 0 < |z| < \infty$

3. $z^{-3} + z^{-1} + \frac{1}{2}z + \frac{1}{6}z^3 + \frac{1}{24}z^5 + \cdots, \quad 0 < |z| < \infty$

5. $z^{-2} + z^{-1} + 1 + z + z^2 + \cdots, \quad 0 < |z| < 1$

7. $z^3 + \frac{1}{2}z + \frac{1}{24}z^{-1} + \frac{1}{720}z^3 + \cdots, \quad 0 < |z| < \infty$

9. $\exp[1 + (z-1)](z-1)^{-2} = e \cdot [(z-1)^{-2} + (z-1)^{-1} + \frac{1}{2} + \frac{1}{6}(z-1) + \cdots]$, $0 < |z-1| < \infty$

11. $$\frac{[\pi i + (z - \pi i)]^2}{(z - \pi i)^4} = \frac{(\pi i)^2}{(z - \pi i)^4} + \frac{2\pi i}{(z - \pi i)^3} + \frac{1}{(z - \pi i)^2}$$

13. $$i^{-3}\left(1 + \frac{z-i}{i}\right)^{-3}(z-i)^{-2} = \sum_{n=0}^{\infty}\binom{-3}{n} i^{-3-n}(z-i)^{n-2} = i(z-i)^{-2}$$
$-3(z-i)^{-1} - 6i + 10(z-i) + \cdots, \quad 0 < |z-i| < 1$

15. $(-\cos(z-\pi))(z-\pi)^{-2} = -(z-\pi)^{-2} + \frac{1}{2} - \frac{1}{24}(z-\pi)^2 + - \cdots$, $0 < |z-\pi| < \infty$

19. $\displaystyle\sum_{n=0}^{\infty} z^{2n}, \quad |z| < 1, \qquad -\sum_{n=0}^{\infty}\frac{1}{z^{2n+2}}, \quad |z| > 1$

21. $-(z + \frac{1}{2}\pi)^{-1}\cos(z + \frac{1}{2}\pi) = -(z + \frac{1}{2}\pi)^{-1} + \frac{1}{2}(z + \frac{1}{2}\pi) - \frac{1}{24}(z + \frac{1}{2}\pi)^3 + - \cdots$, $|z + \frac{1}{2}\pi| > 0$

23. $z^8 + z^{12} + z^{16} + \cdots, \quad |z| < 1, \quad -z^4 - 1 - z^{-4} - z^{-8} - \cdots, \quad |z| > 1$

25. $\dfrac{i}{(z-i)^2} + \dfrac{1}{z-i} + i + (z-i)$

Section 16.2, page 719

1. $0 \pm 2\pi, \pm 4\pi, \cdots$, fourth order

3. $-81i$, fourth order

5. $\pm 1, \pm 2, \cdots$, second order

7. $\pm(2 + 2i), \pm i$, simple

9. $\frac{1}{2}\sin 4z$, $z = 0, \pm\pi/4, \pm\pi/2, \cdots$, simple

11. $f(z) = (z - z_0)^n g(z)$, $g(z_0) \neq 0$, hence $f^2(z) = (z - z_0)^{2n} g^2(z)$.

13. Second-order poles at i and $-2i$

15. Simple pole at ∞, essential singularity at $1 + i$

17. Fourth-order poles at $\pm n\pi i$, $n = 0, 1, \cdots$, essential singularity at ∞

19. $e^z(1 - e^z) = 0$, $e^z = 1$, $z = \pm 2n\pi i$ simple zeros. *Answer:* simple poles at $\pm 2n\pi i$, essential singularity at ∞

21. $1, \infty$ essential singularities, $\pm 2n\pi i$, $n = 0, 1, \cdots$, simple poles

Section 16.3, page 725

3. $\frac{4}{15}$ at 0

5. $\pm 4i$ at $\mp i$

7. $1/\pi$ at $0, \pm 1, \cdots$

9. -1 at $\pm 2n\pi i$

11. $(e^z)''/2!|_{z=\pi i} = -\frac{1}{2}$ at $z = \pi i$

15. Simple pole at $\frac{1}{4}$ inside C, residue $-1/(2\pi)$. *Answer:* $-i$

17. Simple poles at $\pi/2$, residue $e^{\pi/2}/(-\sin \pi/2)$, and at $-\pi/2$, residue $e^{-\pi/2}/\sin \pi/2 = e^{-\pi/2}$. *Answer:* $-4\pi i \sinh \pi/2$

19. $2\pi i(\sinh \frac{1}{2}i)/2 = -\pi \sin \frac{1}{2}$

21. $z^{-5}\cos \pi z = \cdots + \pi^4/(4!z) - + \cdots$. *Answer:* $2\pi^5 i/24$

23. Residues $\frac{1}{2}$ at $z = \frac{1}{2}$, 2 at $z = \frac{1}{3}$. *Answer:* $5\pi i$

25. Simple poles inside C at $2i, -2i, 3i, -3i$, residues $(2i \cosh 2i)/(4z^3 + 26z)|_{z=2i} = \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$, respectively. *Answer:* $2\pi i \cdot \frac{4}{10}$

Problem Set 16.4, page 733

1. $2\pi/\sqrt{k^2 - 1}$

3. $\pi/\sqrt{2}$

5. $5\pi/12$

7. $2a\pi/\sqrt{a^2 - 1}$

9. 0. Why? (Make a sketch.)

11. $\pi/2$

13. 0. Why?

15. $\pi/3$

17. 0. Why?

19. Simple poles at $\pm 1, i$ (and $-i$); $2\pi i \cdot \frac{1}{4} i + \pi i(-\frac{1}{4} + \frac{1}{4}) = -\frac{1}{2}\pi$

21. Simple poles at 1 and $\pm 2\pi i$, residues i and $-i$. *Answer:* $\dfrac{\pi}{5}(\cos 1 - e^{-2})$

23. $-\pi/2$

25. 0

27. Let $q(z) = (z - a_1)(z - a_2)\cdots(z - a_k)$. Use (4) in Sec. 16.3 to form the sum of the residues $1/q'(a_1) + \cdots + 1/q'(a_k)$ and show that this sum is 0; here $k > 1$.

Chapter 16 Review Questions and Problems, page 733

11. $6\pi i$

13. $2\pi i(-10 - 10)$

15. $2\pi i(25z^2)'|_{z=5} = 500\pi i$

17. 0 (n even), $(-1)^{(n-1)/2} 2\pi i/(n - 1)!$ (n odd)

19. $\pi/6$

21. $\pi/60$

23. 0. Why?

25. $\operatorname*{Res}_{z=i} e^{iz}/(z^2 + 1) = 1/(2ie)$. *Answer:* π/e.

Problem Set 17.1, page 741

5. Only in size

7. $x = c, w = -y + ic; \quad y = k, w = -k + ix$

9. Parallel displacement; each point is moved 2 to the right and 1 up.

11. $|w| \leqq \frac{1}{4}, \quad -\pi/4 < \operatorname{Arg} w < \pi/4$

13. $-5 \leqq \operatorname{Re} z \leqq -2$

15. $u \geqq 1$

17. Annulus $\frac{1}{2} \leqq |w| \leqq 4$

19. $0 < u < \ln 4, \quad \pi/4 < v \leqq 3\pi/4$

21. $z^3 + az^2 + bz + c, \quad z = -\frac{1}{3}(a \pm \sqrt{a^2 - 3b})$

23. $z = (-1 \pm \sqrt{3})/2$

25. $\sinh z = 0$ at $z = 0, \quad \pm\pi i, \pm 2\pi i, \cdots$

29. $M = |z| = 1$ on the unit circle, $J = |z|^2$

31. $|w'| = 1/|z|^2 = 1$ on the unit circle, $J = 1/|z|^4$

33. $M = e^x = 1$ for $x = 0$, the y-axis, $J = e^{2x}$

35. $M = 1/|z| = 1$ on the unit circle, $J = 1/|z|^2$

Problem Set 17.2, page 745

7. $z = \dfrac{w + i}{2w}$

9. $z = \dfrac{4w + i}{-3iw + 1}$

11. $z = 0, \quad 1/(a + ib)$

13. $z = 0, \quad \pm\frac{1}{2}, \pm = \pm i/2$

15. $z = i, 2i$ **17.** $w = \dfrac{az}{cz + a}$ **19.** $w = \dfrac{az + b}{-bz + a}$

Problem Set 17.3, page 750

3. Apply the inverse g of f on both sides of $z_1 = f(z_1)$ to get $g(z_1) = g(f(z_1)) = z_1$.
9. $w = iz$, a rotation. Sketch to see. **11.** $w = (z + i)/(z - i)$
13. $w = 1/z$, almost by inspection **15.** $w = 1/z - 1$
17. $w = (2z - i)/(-iz - 2)$ **19.** $w = (z^4 - i)(-iz^4 + 1)$

Problem Set 17.4, page 754

1. Circle $|w| = e^c$ **3.** Annulus $1/\sqrt{e} \leqq |w| \leqq \sqrt{e}$
5. w-plane without $w = 0$ **7.** $1 < |w| < e, v > 0$
9. $\pm(2n + 1)\pi/2, \quad n = 0, 1, \cdots$
11. $u^2/\cosh^2 2 + v^2/\sinh^2 2 < 1, \quad u > 0, v > 0$
13. Elliptic annulus bounded by $u^2/\cosh^2 1 + v^2/\sinh^2 1 = 1$ and $u^2/\cosh^2 3 + v^2/\sinh^2 3 = 1$
15. $\cosh z = \cos iz = \sin (iz + \frac{1}{2}\pi)$
17. $0 < \operatorname{Im} t < \pi$ is the image of R under $t = z^2/2$. *Answer:* $e^t = e^{z^2/2}$.
19. Hyperbolas $u^2/\cos^2 c - v^2/\sin^2 c = \cosh^2 c - \sinh^2 c = 1$ when $c \neq 0, \pi$, and $u = \pm\cosh y$ (thus $|u| \geqq 1$), $v = 0$ when $c = 0, \pi$.
21. Interior of $u^2/\cosh^2 2 + v^2/\sinh^2 2 = 1$ in the fourth quadrant, or map $\pi/2 < x < \pi, 0 < y < 2$ by $w = \sin z$ (why?).
23. $v < 0$
25. The images of the five points in the figure can be obtained directly from the function w.

Problem Set 17.5, page 756

1. w moves once around the circle $|w| = \frac{1}{2}$.
3. Four sheets, branch point at $z = -1$
5. $-i/4$, three sheets
7. z_0, n sheets
9. $\sqrt{z(z - i)(z + i)}$, $0, \pm i$, two sheets

Chapter 17 Review Questions and Problems, page 756

11. $1 < |w| < 4, |\arg w| < \pi/4$ **13.** Horizontal strip $-8 < v < 8$
15. $u = 1 - \frac{1}{4}v^2$, same (why?) **17.** $|w| > 1$
19. $\frac{1}{3} < |w| < \frac{1}{2}, \quad v < 0$ **21.** $w = 1 + iv, \quad v < 0$
23. $w = \dfrac{10z + 5i}{z + 2i}$ **25.** Rotation $w = iz$
27. $w = 1/z$ **29.** $z = 0$
31. $z = 2 \pm \sqrt{6}$ **33.** $z = 0, \pm i, \pm 3i$
35. $w = e^{4z}$ **37.** $w = iz^2 + 1$
39. $w = z^2/(2c)$

Problem Set 18.1, page 762

1. 2.5 mm = 0.25 cm; $\Phi = \text{Re}\, 110(1 + (\text{Ln}\, z)/\ln 4)$

3. $\Phi = \text{Re}\left(30 - \dfrac{20}{\ln 10}\,\text{Ln}\, z\right)$

5. $\Phi(x) = \text{Re}\,(375 + 25z)$

7. $\Phi(r) = \text{Re}\,(32 - z)$

13. Use Fig. 391 in Sec. 17.4 with the z- and w-planes interchanged and $\cos z = \sin(z + \frac{1}{2}\pi)$.

15. $\Phi = 220(x^3 - 3xy^2) = \text{Re}\,(220z^3)$

Problem Set 18.2, page 766

3. $w = iz^2$ maps R onto the strip $-2 \leqq u \leqq 0$; and $\Phi^* = U_2 + (U_1 - U_2)(1 + \frac{1}{2}u) = U_2 + (U_1 - U_2)(1 - xy)$.

5. **(a)** $\dfrac{(x-2)(2x-1) + 2y^2}{(x-2)^2 + y^2} = c$, **(b)** $x^2 - y^2 = c$, $xy = c$, $e^x \cos y = c$

7. See Fig. 392 in Sec. 17.4. $\Phi = \text{Re}\,(\sin^2 z)$, $\sin^2 x\ (y = 0)$, $\sin^2 x \cosh^2 1 - \cos^2 x \sinh^2 1\ (y = 1)$, $-\sinh^2 y\ (x = 0, \pi)$.

9. $\Phi(x, y) = \cos^2 x \cosh^2 y - \sin^2 x \sinh^2 y$; $\cosh^2 y\ (x = 0)$, $-\sinh y\ (x = \frac{\pi}{2})$, $\cos^2 x\ (y = 0)$, $\cos^2 x \cosh^2 1 - \sin^2 x \sinh^2 1\ (y = 1)$

13. Corresponding rays in the w-plane make equal angles, and the mapping is conformal.

15. Apply $w = z^2$.

17. $z = (2Z - i)/(-iZ - 2)$ by (3) in Sec. 17.3.

19. $\Phi = \dfrac{5}{\pi}\,\text{Arg}\,(z - 2)$, $F = -\dfrac{5i}{\pi}\,\text{Ln}\,(z - 2)$

Problem Set 18.3, page 769

1. $(80/d)y + 20$. Rotate through $\pi/2$.

5. $\dfrac{80}{\pi}\arctan\dfrac{y}{x} = \text{Re}\left(-\dfrac{80i}{\pi}\,\text{Ln}\, z\right)$

7. $T_1 + \dfrac{2}{\pi}(T_2 - T_1)\arctan\dfrac{y}{x} = \text{Re}\left(T_1 - \dfrac{2i}{\pi}(T_2 - T_1)\,\text{Ln}\, z\right)$

9. $\dfrac{T_1}{\pi}\left(\arctan\dfrac{y}{x-b} - \arctan\dfrac{y}{x-a}\right) = \text{Re}\left(\dfrac{iT_1}{\pi}\,\text{Ln}\,\dfrac{z-a}{z-b}\right)$

11. $\dfrac{100}{\pi}(\text{Arg}\,(z-1) - \text{Arg}\,(z+1)) = \text{Re}\left(\dfrac{100i}{\pi}\,\text{Ln}\,\dfrac{z+1}{z-1}\right)$

13. $\dfrac{100}{\pi}[\text{Arg}\,(z^2 - 1) - \text{Arg}\,(z^2 + 1)]$ from $w = z^2$ and Prob. 11.

15. $-20 + (320/\pi)\,\text{Arg}\, z = \text{Re}\left(-20 - \dfrac{320i}{\pi}\,\text{Ln}\, z\right)$

17. $\text{Re}\, F(z) = 100 + (200/\pi)\,\text{Re}\,(\arcsin z)$

Problem Set 18.4, page 776

1. $V(z)$ continuously differentiable.

3. $|F'(iy)| = 1 + 1/y^2$, $|y| \geqq 1$, is maximum at $y = \pm 1$, namely, 2.

5. Calculate or note that ∇^2 = div grad and curl grad is the zero vector; see Sec. 9.8 and Problem Set 9.7.
7. Horizontal parallel flow to the right.
9. $F(z) = z^4$
11. Uniform parallel flow upward, $V = \overline{F'} = iK$, $V_1 = 0$, $V_2 = K$
13. $F(z) = z^3$
15. $F(z) = z/r_0 + r_0/z$
17. Use that $w = \arccos z$ gives $z = \cos w$ and interchanging the roles of the z- and w-planes.
19. $y/(x^2 + y^2) = c$ or $x^2 + (y - k)^2 = k^2$

Problem Set 18.5, page 781

5. $\Phi = \frac{3}{2} r^3 \sin 3\theta$
7. $\Phi = \frac{1}{2} a + \frac{1}{2} a r^8 \cos 8\theta$
9. $\Phi = 3 - 4r^2 \cos 2\theta + r^4 \cos 4\theta$
11. $\Phi = \dfrac{2}{\pi}\left(r \sin\theta - \dfrac{1}{2} r^2 \sin 2\theta + \dfrac{1}{3} r^3 \sin 3\theta - + \cdots\right)$
13. $\Phi = \dfrac{2}{\pi} r \sin\theta + \dfrac{1}{2} r^2 \sin 2\theta - \dfrac{2}{9\pi} r^3 \sin 3\theta - \dfrac{1}{4} r^4 \sin 4\theta + + - \cdots$
15. $\Phi = \dfrac{1}{2} + \dfrac{2}{\pi}\left(r \cos\theta - \dfrac{1}{3} r^3 \cos 3\theta + \dfrac{1}{5} r^5 \cos 5\theta - + \cdots\right)$
17. $\Phi = \dfrac{1}{3} - \dfrac{4}{\pi^2}\left(r \cos\theta - \dfrac{1}{4} r^2 \cos 2\theta + \dfrac{1}{9} r^3 \cos 3\theta - + \cdots\right)$

Problem Set 18.6, page 784

1. Use (2). $F(z_0 + e^{i\alpha}) = (\frac{7}{2} + e^{i\alpha})^3$, etc. $F(\frac{5}{2}) = \frac{343}{8}$
3. Use (2). $F(z_0 + e^{i\alpha}) = (2 + 3e^{i\alpha})^2$, etc. $F(4) = 100$
5. No, because $|z|$ is not analytic.
7.
$$\Phi(2, -2) = -3 = \frac{1}{\pi}\int_0^1\int_0^{2\pi} (1 + r\cos\alpha)(-3 + r\sin\alpha) r\, dr\, d\alpha$$
$$= \frac{1}{\pi}\int_0^1\int_0^{2\pi} (-3r + \cdots)\, dr\, d\alpha = \frac{1}{\pi}\left(-\frac{3}{2}\right)\cdot 2\pi$$
9.
$$\Phi(1, 1) = 3 = \frac{1}{\pi}\int_0^1\int_0^{2\pi} (3 + r\cos\alpha + r\sin\alpha + r^2\cos\alpha\sin\alpha) r\, dr\, d\alpha$$
$$= \frac{1}{\pi}\cdot\frac{3}{2}\cdot 2\pi$$
13. $|F(z)| = [\cos^2 x + \sinh^2 y]^{1/2}$, $z = \pm i$, Max $= [1 + \sinh^2 1]^{1/2} = 1.543$
15. $|F(z)|^2 = \sinh^2 2x \cos^2 2y + \cosh^2 2x \sin^2 2y = \sinh^2 2x + 1 \cdot \sin^2 2y$, $z = 1$, Max $= \sinh 2 = 3.627$
17. $|F(z)|^2 = 4(2 - 2\cos 2\theta)$, $z = \pi/2$, $3\pi/2$, Max $= 4$
19. No. Make up a counterexample.

Chapter 18 Review Questions and Problems, page 785

11. $\Phi = 10(1 - x + y)$, $F = 10 - 10(1 + i)z$

13. $\Phi = \text{Re}\,(220 - 95.54 \text{ Ln } z) = 220 - \dfrac{220}{\ln 10} \ln r = 220 - 95.54 \ln r.$

17. $2(1 - (2/\pi) \text{ Arg } z)$

19. $30(1 - (2/\pi) \text{ Arg } (z - 1))$

21. $\Phi = x + y = \text{const}$, $V = \overline{F'(z)} = 1 - i$, parallel flow

23. $T(x, y) = x(2y + 1) = \text{const}$

25. $\overline{F'(z)} = \bar{z} + 1 = x + 1 - iy$

Problem Set 19.1, page 796

1. $0.84175 \cdot 10^2$, $-0.52868 \cdot 10^3$, $0.92414 \cdot 10^{-3}$, $-0.36201 \cdot 10^6$

3. 6.3698, 6.794, 8.15, impossible

5. Add first, then round.

7. 29.9667, 0.0335; 29.9667, 0.0333704 (6S-exact)

9. 29.97, 0.035; 29.97, 0.03337; 30, 0.0; 30, 0.033

11. $|\epsilon| = |x + y - (\tilde{x} + \tilde{y})| = |(x - \tilde{x}) + (y - \tilde{y})| = |\epsilon_x + \epsilon_y|$
$\leqq |\epsilon_x| + |\epsilon_y| = \beta_x + \beta_y$

13. $\dfrac{a_1}{a_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2 + \epsilon_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2}\left(1 - \dfrac{\epsilon_2}{\tilde{a}_2} + \dfrac{\epsilon_2^2}{\tilde{a}_2^2} - + \cdots\right) \approx \dfrac{\tilde{a}_1}{\tilde{a}_2} + \dfrac{\epsilon_1}{\tilde{a}_2} - \dfrac{\epsilon_2}{\tilde{a}_2} \cdot \dfrac{\tilde{a}_1}{\tilde{a}_2}$,
hence $\left|\left(\dfrac{a_1}{a_2} - \dfrac{\tilde{a}_1}{\tilde{a}_2}\right)\middle/\dfrac{a_1}{a_2}\right| \approx \left|\dfrac{\epsilon_1}{a_1} - \dfrac{\epsilon_2}{a_2}\right| \leqq |\epsilon_{r1}| + |\epsilon_{r2}| \leqq \beta_{r1} + \beta_{r2}$

15. **(a)** $1.38629 - 1.38604 = 0.00025$, **(b)** $\ln 1.00025 = 0.000249969$ is 6S-exact.

19. In the present case, (b) is slightly more accurate than (a) (which may produce nonsensical results; cf. Prob. 20).

21. $c_4 \cdot 2^4 + \cdots + c_0 \cdot 2^0 = (1\ 0\ 1\ 1\ 1.)_2$, NOT $(1\ 1\ 1\ 0\ 1.)_2$

23. The algorithm in Prob. 22 repeats 0011 infinitely often.

25. $n = 26$. The beginning is 0.09375 ($n = 1$).

27. $I_{14} = 0.1812$ (0.1705 4S-exact), $I_{13} = 0.1812$ (0.1820), $I_{12} = 0.1951$ (0.1951), $I_{11} = 0.2102$ (0.2103), etc.

29. $-0.126 \cdot 10^{-2}$, $-0.402 \cdot 10^{-3}$; $-0.266 \cdot 10^{-6}$, $-0.847 \cdot 10^{-7}$

Problem Set 19.2, page 807

3. $g = 0.5 \cos x$, $x = 0.450184$ ($= x_{10}$, exact to 6S)

5. Convergence to 4.7 for all these starting values.

7. $x = x/(e^x \sin x)$; 0.5, 0.63256, $\cdots$ converges to 0.58853 (5S-exact) in 14 steps.

9. $x = x^4 - 0.12$; $x_0 = 0, x_3 = -0.119794$ (6S-exact)

11. $g = 4/x + x^3/16 - x^5/576$; $x_0 = 2, x_n = 2.39165$ ($n \geqq 6$), 2.405 4S-exact

13. This follows from the intermediate value theorem of calculus.

15. $x_3 = 0.450184$

17. Convergence to $x = 4.7, 4.7, 0.8, -0.5$, respectively. Reason seen easily from the graph of f.

19. 0.5, 0.375, 0.377968, 0.377964; (b) $1/\sqrt{7}$

21. 1.834243 ($= x_4$), 0.656620 ($= x_4$), -2.49086 ($= x_4$)

23. $x_0 = 4.5$, $x_4 = 4.73004$ (6S-exact)

25. (a) ALGORITHM BISECT (f, a_0, b_0, ϵ, N) Bisection Method

This algorithm computes the solution c of $f(x) = 0$ (f continuous) within the tolerance ϵ, given an initial interval $[a_0, b_0]$ such that $f(a_0)f(b_0) < 0$.

INPUT: Continuous function f, initial interval $[a_0, b_0]$, tolerance ϵ, maximum number of iterations N.

OUTPUT: A solution c (within the tolerance ϵ), or a message of failure.

For $n = 0, 1, \cdots, N - 1$ do:

$c = \frac{1}{2}(a_n + b_n)$

If $f(c) = 0$ then OUTPUT c Stop. [*Procedure completed*]

Else if $f(a_n)f(b_n) < 0$ then set $a_{n+1} = a_n$ and $b_{n+1} = c$.

Else set $a_{n+1} = c$, and $b_{n+1} = b_n$.

If $|a_{n+1} - b_{n+1}| < \epsilon|c|$ then OUTPUT c. Stop. [*Procedure completed*]

End

OUTPUT $[a_N, b_N]$ and a message "Failure". Stop.

[*Unsuccessful completion; N iterations did not give an interval of length not exceeding the tolerance.*]

End BISECT

Note that $[a_N, b_N]$ gives $(a_N + b_N)/2$ as an approximation of the zero and $(b_N - a_N)/2$ as a corresponding error bound.

(b) 0.739085; **(c)** 1.30980, 0.429494

27. $x_2 = 1.5$, $x_3 = 1.76471, \cdots$, $x_7 = 1.83424$ (6S-exact)

29. 0.904557 (6S-exact)

Problem Set 19.3, page 819

1. $L_0(x) = -2x + 19$, $L_1(x) = 2x - 18$, $p_1(9.3) = L_0(9.3) \cdot f_0 + L_1(9.3) \cdot f_1 = 0.1086 \cdot 9.3 + 1.230 = 2.2297$

3. $$p_2(x) = \frac{(x - 1.02)(x - 1.04)}{(-0.02)(-0.04)} \cdot 1.0000 + \frac{(x - 1)(x - 1.04)}{0.02\,(-0.02)} \cdot 0.9888 + \frac{(x - 1)(x - 1.02)}{0.04 \cdot 0.02} \cdot 0.9784 = x^2 - 2.580x + 2.580;$$ 0.9943, 0.9835

5. 0.8033 (error -0.0245), 0.4872 (error -0.0148); quadratic: 0.7839 (-0.0051), 0.4678 (0.0046)

7. $p_2(x) = 1.1640x - 0.3357x^2$; -0.5089 (error 0.1262), 0.4053 (-0.0226), 0.9053 (0.0186), 0.9911 (-0.0672)

9. $p_2(x) = -0.44304x^2 + 1.30896x - 0.023220$, $p_2(0.75) = 0.70929$ (5S-exact 0.71116)

11. $L_0 = -\frac{1}{6}(x - 1)(x - 2)(x - 3)$, $L_1 = \frac{1}{2}x(x - 2)(x - 3)$, $L_2 = -\frac{1}{2}x(x - 1)(x - 3)$, $L_3 = \frac{1}{6}x(x - 1)(x - 2)$; $p_3(x) = 1 + 0.039740x - 0.335187x^2 + 0.060645x^3$; $p_2(0.5) = 0.943654$, $p_3(1.5) = 0.510116$, $p_3(2.5) = -0.047991$

13. $2x^2 - 4x + 2$

15. $p_3(x) = 2.1972 + (x - 9) \cdot 0.1082 + (x - 9)(x - 9.5) \cdot 0.005235$

17. $r = -1.5$, $p_2(0.3) = 0.6039 + (-1.5) \cdot 0.1755 + \frac{1}{2}(-1.5)(-0.5) \cdot (-0.0302) = 0.3293$

Problem Set 19.4, page 826

9. $[-1.39(x-5)^2 + 0.58(x-5)^3]'' = 0.004$ at $x = 5.8$ (due to roundoff; should be 0).

11. $1 - \frac{5}{4}x^2 + \frac{1}{4}x^4$

13. $1 - x^2$, $\quad -2(x-1) - (x-1)^2 + 2(x-1)^3$, $\quad -1 + 2(x-2) + 5(x-2)^2 - 6(x-2)^3$

15. $4 + x^2 - x^3$, $\quad -8(x-2) - 5(x-2)^2 + 5(x-2)^3$, $4 + 32(x-4) + 25(x-4)^2 - 11(x-4)^3$

17. Use the fact that the third derivative of a cubic polynomial is constant, so that g''' is piecewise constant, hence constant throughout under the present assumption. Now integrate three times.

19. Curvature $f''/(1 + f'^2)^{3/2} \approx f''$ if $|f'|$ is small.

Problem Set 19.5, page 839

1. 0.747131, which is larger than 0.746824. Why?

3. 0.5, $\quad$ 0.375, $\quad$ 0.34375, $\quad$ 0.335 (exact)

5. $\epsilon_{0.5} \approx 0.03452$ ($\epsilon_{0.5} = 0.03307$), $\quad \epsilon_{0.25} \approx 0.00829$ ($\epsilon_{0.25} = 0.00820$)

7. 0.693254 (6S-exact 0.693147)

9. 0.073930 (6S-exact 0.073928)

11. 0.785392 (6S-exact 0.785398)

13. $(0.785398126 - 0.785392156)/15 = 0.39792 \cdot 10^{-6}$

15. **(a)** $M_2 = 2$, $|KM_2| = 2/(12n^2) = 10^{-5}/2$, $n = 183$. **(b)** $f^{\text{iv}} = 24/x^5$, $M_4 = 24$, $|CM_4| = 24/(180 \cdot (2m)^4) = 10^{-5}/2$, $2m = 12.8$, $\quad$ hence 14.

17. 0.94614588, 0.94608693 (8S-exact 0.94608307)

19. 0.9460831 (7S-exact)

21. 0.9774586 (7S-exact 0.9774377)

23. Set $x = \frac{1}{2}(t + 1)$, $\quad$ 0.2642411177 (10S-exact), $\quad 1 - 2/e$

25. $x = \frac{1}{2}(t + 1)$, $\quad dx = \frac{1}{2}dt$, $\quad$ 0.746824127 $\quad$ (9S-exact 0.746824133)

27. 0.08, $\quad$ 0.32, $\quad$ 0.176, $\quad$ 0.256 (exact)

29. $5(0.1040 - \frac{1}{2} \cdot 0.1760 + \frac{1}{3} \cdot 0.1344 - \frac{1}{4} \cdot 0.0384) = 0.256$

Chapter 19 Review Questions and Problems, page 841

17. 4.375, $\quad$ 4.50, $\quad$ 6.0, $\quad$ impossible

19. $44.885 \leqq s \leqq 44.995$

21. The same as that of $\tilde{a}$.

23. $x = 20 \pm \sqrt{398} = 20.00 \pm 19.95$, $\quad x_1 = 39.95$, $\quad x_2 = 0.05$, $\quad x_2 = 2/39.95 = 0.05006$ (error less than 1 unit of the last digit)

25. $x = x^4 - 0.1$, $\quad -0.1$, $\quad -0.999$, $\quad -0.99900399$

27. 0.824

29. $-x + x^3$, $\quad 2(x-1) + 3(x-1)^2 - (x-1)^3$

31. 0.26, $\quad M_2 = 6$, $\quad M_2^* = 0$, $\quad -0.02 \leqq \epsilon \leqq 0$, $\quad$ 0.01

33. 0.90443, $\quad$ 0.90452 (5S-exact 0.90452)

35. (a) $(0.4^3 - 2 \cdot 0.2^3 + 0)/0.04 = 1.2$, (b) $(0.3^3 - 2 \cdot 0.2^3 + 0.1^3)/0.01 = 1.2$ (exact)

Problem Set 20.1, page 851

1. $x_1 = 7.3, \quad x_2 = -3.2$ **3.** No solution **5.** $x_1 = 2, \quad x_2 = 1$

7. $\begin{bmatrix} -3 & 6 & -9 & -46.725 \\ 0 & 9 & -13 & -51.223 \\ 0 & 0 & -2.88889 & -7.38689 \end{bmatrix}$

$x_1 = 3.908, \quad x_2 = -1.998, \quad x_3 = 2.557$

9. $\begin{bmatrix} 13 & -8 & 0 & 178.54 \\ 0 & 6 & 13 & 137.86 \\ 0 & 0 & -16 & -253.12 \end{bmatrix}$

$x_1 = 6.78, \quad x_2 = -11.3, \quad x_3 = 15.82$

11. $\begin{bmatrix} 3.4 & -6.12 & -2.72 & 0 \\ 0 & 0 & 4.32 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

$x_1 = t_1$ arbitrary, $\quad x_2 = (3.4/6.12)t_1, \quad x_3 = 0$

13. $\begin{bmatrix} 5 & 0 & 6 & -0.329193 \\ 0 & -4 & -3.6 & -2.143144 \\ 0 & 0 & 2.3 & -0.4 \end{bmatrix}$

$x_1 = 0.142856, \quad x_2 = 0.692307, \quad x_3 = -0.173912$

15. $\begin{bmatrix} -1 & -3.1 & 2.5 & 0 & -8.7 \\ 0 & 2.2 & 1.5 & -3.3 & -9.3 \\ 0 & 0 & -1.493182 & -0.825 & 1.03773 \\ 0 & 0 & 0 & 6.13826 & 12.2765 \end{bmatrix}$

$x_1 = 4.2, \quad x_2 = 0, \quad x_3 = -1.8, \quad x_4 = 2.0$

Problem Set 20.2, page 857

1. $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 0 & -1 \end{bmatrix}, \quad \begin{matrix} x_1 = -4 \\ x_2 = 6 \end{matrix}$

3. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 5 & 1 \end{bmatrix} \begin{bmatrix} 5 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{matrix} x_1 = 0.4 \\ x_2 = 0.8 \\ x_3 = 1.6 \end{matrix}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 3 & 9 & 1 \end{bmatrix} \begin{bmatrix} 3 & 9 & 6 \\ 0 & -6 & 3 \\ 0 & 0 & -3 \end{bmatrix}, \quad \begin{matrix} x_1 = -\frac{1}{15} \\ x_2 = \frac{4}{15} \\ x_3 = \frac{2}{5} \end{matrix}$

7. $\begin{bmatrix} 3 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & 2 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{matrix} x_1 = 0.6 \\ x_2 = 1.2 \\ x_3 = 0.4 \end{matrix}$

9. $\begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0.3 & 0.2 & 0.1 \end{bmatrix}\begin{bmatrix} 0.1 & 0 & 0.3 \\ 0 & 0.4 & 0.2 \\ 0 & 0 & 0.1 \end{bmatrix}, \quad \begin{matrix} x_1 = 2 \\ x_2 = -11 \\ x_3 = 4 \end{matrix}$

11. $\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 3 & -1 & 3 & 0 \\ 2 & 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} 1 & -1 & 3 & 2 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \quad \begin{matrix} x_1 = 2 \\ x_2 = -3 \\ x_3 = 4 \\ x_4 = -1 \end{matrix}$

13. No, since $\mathbf{x}^\mathsf{T}(-\mathbf{A})\mathbf{x} = -\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} < 0$; yes; yes; no

15. $\begin{bmatrix} -3.5 & 1.25 \\ 3.0 & -1.0 \end{bmatrix}$

17. $\dfrac{1}{36}\begin{bmatrix} 584 & 104 & -66 \\ 104 & 20 & -12 \\ -66 & -12 & 9 \end{bmatrix}$

19. $\dfrac{1}{16}\begin{bmatrix} 21 & -6 & -14 & 6 \\ -6 & 36 & -12 & -4 \\ -14 & -12 & 20 & -4 \\ 6 & -4 & -4 & 4 \end{bmatrix}$

Problem Set 20.3, page 863

5. Exact 0.5, 0.5, 0.5

7. $x_1 = 2, \quad x_2 = -4, \quad x_3 = 8$

9. Exact 2, 1, 4

11. (a) $\mathbf{x}^{(3)\mathsf{T}} = [0.49983 \quad 0.50001 \quad 0.500017]$,
(b) $\mathbf{x}^{(3)\mathsf{T}} = [0.50333 \quad 0.49985 \quad 0.49968]$

13. 8, −16, 43, 86 steps; spectral radius 0.09, 0.35, 0.72, 0.85, approximately

15. $[1.99934 \quad 1.00043 \quad 3.99684]^\mathsf{T}$ (Jacobi, Step 5); $[2.00004 \quad 0.998059 \quad 4.00072]^\mathsf{T}$ (Gauss–Seidel)

19. $\sqrt{306} = 17.49$, 12, 12

Problem Set 20.4, page 871

1. 18, $\sqrt{110} = 10.49$, 8, $[0.125 \quad -0.375 \quad 1 \quad 0 \quad -0.75 \quad 0]$

3. 5.9, $\sqrt{13.81} = 3.716$, 3, $\frac{1}{3}[0.2 \quad 0.6 \quad -2.1 \quad 3.0]$

5. 5, $\sqrt{5}$, 1, $[1 \quad 1 \quad 1 \quad 1 \quad 1]$

7. $ab + bc + ca = 0$

9. $\kappa = 5 \cdot \frac{1}{2} = 2.5$ **11.** $\kappa = (5 + \sqrt{5})(1 + 1/\sqrt{5}) = 6 + 2\sqrt{5}$
13. $\kappa = 19 \cdot 13 = 247$; ill-conditioned
15. $\kappa = 20 \cdot 20 = 400$; ill-conditioned
17. $167 \leqq 21 \cdot 15 = 315$
19. $[-2 \quad 4]^{\mathsf{T}}$, $[-144.0 \quad 184.0]^{\mathsf{T}}$, $\kappa = 25{,}921$, extremely ill-conditioned
21. Small residual $[0.145 \quad 0.120]$, but large deviation of $\tilde{\mathbf{x}}$.
23. 27, 748, 28,375, 943,656, 29,070,279

Problem Set 20.5, page 875

1. $1.846 - 1.038x$ **3.** $1.48 + 0.09x$
5. $s = 90t - 675$, $v_{\text{av}} = 90$ km/hr **9.** $-11.36 + 5.45x - 0.589x^2$
11. $1.89 - 0.739x + 0.207x^2$
13. $2.552 + 16.23x$, $-4.114 + 13.73x + 2.500x^2$, $2.730 + 1.466x$
$- 1.778x^2 + 2.852x^3$

Problem Set 20.7, page 884

1. 5, 0, 7; radii 6, 4, 6. Spectrum $\{-1, 4, 9\}$
3. Centers 0; radii 0.5, 0.7, 0.4. Skew-symmetric, hence $\lambda = i\mu$, $-0.7 \leqq \mu \leqq 0.7$.
5. 2, 3, 8; radii $1 + \sqrt{2}$, 1, $\sqrt{2}$; actually (4S) 1.163, 3.511, 8.326
7. $t_{11} = 100$, $t_{22} = t_{33} = 1$
9. They lie in the intervals with endpoints $a_{jj} \pm (n - 1) \cdot 10^{-5}$. Why?
11. $\rho(\mathbf{A}) \leqq$ Row sum norm $\|\mathbf{A}\|_\infty = \max_j \sum_k |a_{jk}| = \max_j(|a_{jj}| +$ Gerschgorin radius$)$
13. $\sqrt{122} = 11.05$
15. $\sqrt{0.52} = 0.7211$
17. Show that $\mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}} = \bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A}$.
19. 0 lies in no Gerschgorin disk, by (3) with $>$; hence $\det \mathbf{A} = \lambda_1 \cdots \lambda_n \neq 0$.

Problem Set 20.8, page 887

1. $q = 10, 10.9908, 10.9999$; $|\epsilon| \leqq 3, 0.3028, 0.0275$
3. $q \pm \delta = 4 \pm 1.633$, 4.786 ± 0.619, 4.917 ± 0.398
5. Same answer as in Prob. 3, possibly except for small roundoff errors.
7. $q = 5.5, 5.5738, 5.6018$; $|\epsilon| \leqq 0.5, 0.3115, 0.1899$; eigenvalues (4S) 1.697,
3.382, 5.303, 5.618
9. $\mathbf{y} = \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{y}^{\mathsf{T}}\mathbf{x} = \lambda\mathbf{x}^{\mathsf{T}}\mathbf{x}$, $\mathbf{y}^{\mathsf{T}}\mathbf{y} = \lambda^2\mathbf{x}^{\mathsf{T}}\mathbf{x}$,
$\epsilon^2 \leqq \mathbf{y}^{\mathsf{T}}\mathbf{y}/\mathbf{x}^{\mathsf{T}}\mathbf{x} - (\mathbf{y}^{\mathsf{T}}\mathbf{x}/\mathbf{x}^{\mathsf{T}}\mathbf{x})^2 = \lambda^2 - \lambda^2 = 0$
11. $q = 1, \cdots, -2.8993$ approximates -3 (0 of the given matrix),
$|\epsilon| \leqq 1.633, \cdots, 0.7024$ (Step 8)

Problem Set 20.9, page 896

1. $$\begin{bmatrix} 0.98 & -0.4418 & 0 \\ -0.4418 & 0.8702 & 0.3718 \\ 0 & 0.3718 & 0.4898 \end{bmatrix}$$

3. $\begin{bmatrix} 7 & -3.6056 & 0 \\ -3.6056 & 13.462 & 3.6923 \\ 0 & 3.6923 & 3.5385 \end{bmatrix}$

5. $\begin{bmatrix} 3 & -67.59 & 0 & 0 \\ -67.59 & 143.5 & 45.35 & 0 \\ 0 & 45.35 & 23.34 & 3.126 \\ 0 & 0 & 3.126 & -33.87 \end{bmatrix}$

7. Eigenvalues 16, 6, 2

$\begin{bmatrix} 11.2903 & -5.0173 & 0 \\ -5.0173 & 10.6144 & 0.7499 \\ 0 & 0.7499 & 2.0952 \end{bmatrix}, \begin{bmatrix} 14.9028 & -3.1265 & 0 \\ -3.1265 & 7.0883 & 0.1966 \\ 0 & 0.1966 & 2.0089 \end{bmatrix}, \begin{bmatrix} 15.8299 & -1.2932 & 0 \\ -1.2932 & 6.1692 & 0.0625 \\ 0 & 0.0625 & 2.0010 \end{bmatrix}$

9. Eigenvalues (4S) 141.4, 68.64, -30.04

$\begin{bmatrix} 141.1 & 4.926 & 0 \\ 4.926 & 68.97 & 0.8691 \\ 0 & 0.8691 & -30.03 \end{bmatrix}, \begin{bmatrix} 141.3 & 2.400 & 0 \\ 2.400 & 68.72 & 0.3797 \\ 0 & 0.3797 & -30.04 \end{bmatrix}, \begin{bmatrix} 141.4 & 1.166 & 0 \\ 1.166 & 68.66 & 0.1661 \\ 0 & 0.1661 & -30.04 \end{bmatrix}$

Chapter 20 Review Questions and Problems, page 896

15. $[3.9 \quad 4.3 \quad 1.8]^T$

17. $[-2 \quad 0 \quad 5]^T$

19. $\begin{bmatrix} 0.28193 & -0.15904 & -0.00482 \\ -0.15904 & 0.12048 & -0.00241 \\ -0.00482 & -0.00241 & 0.01205 \end{bmatrix}$

21. $\begin{bmatrix} 5.750 \\ 3.600 \\ 0.838 \end{bmatrix}, \begin{bmatrix} 6.400 \\ 3.559 \\ 1.000 \end{bmatrix}, \begin{bmatrix} 6.390 \\ 3.600 \\ 0.997 \end{bmatrix}$

Exact: $[6.4 \quad 3.6 \quad 1.0]^T$

23. $\begin{bmatrix} 1.700 \\ 1.180 \\ 4.043 \end{bmatrix}, \begin{bmatrix} 1.986 \\ 0.999 \\ 4.002 \end{bmatrix}, \begin{bmatrix} 2.000 \\ 1.000 \\ 4.000 \end{bmatrix}$

Exact: $[2 \quad 1 \quad 4]^T$

25. 42, $\sqrt{674} = 25.96$, 21

27. 30

29. 5

31. $115 \cdot 0.4458 = 51.27$

33. $5 \cdot \frac{21}{63} = \frac{5}{3}$

35. $1.514 + 1.129x - 0.214x^2$

37. Centers 15, 35, 90; radii 30, 35, 25, respectively. Eigenvalues (3S) 2.63, 40.8, 96.6

39. Centers 0, -1, -4; radii 9, 6, 7, respectively; eigenvalues 0, 4.446, -9.446

Problem Set 21.1, page 910

1. $y = 5e^{-0.2x}$, 0.00458, 0.00830 (errors of y_5, y_{10})
3. $y = x - \tanh x$ (set $y - x = u$), 0.00929, 0.01885 (errors of y_5, y_{10})
5. $y = e^x$, 0.0013, 0.0042 (errors of y_5, y_{10})
7. $y = 1/(1 - x^2/2)$, 0.00029, 0.01187 (errors of y_5, y_{10})
9. Errors 0.03547 and 0.28715 of y_5 and y_{10} much larger
11. $y = 1/(1 - x^2/2)$; error -10^{-8}, $-4 \cdot 10^{-8}$, $\cdots$, $-6 \cdot 10^{-7}$, $+9 \cdot 10^{-6}$; $\epsilon = 0.0002/15 = 1.3 \cdot 10^{-5}$ (use RK with $h = 0.2$)
13. $y = \tan x$; error $0.83 \cdot 10^{-7}, 0.16 \cdot 10^{-6}, \cdots, -0.56 \cdot 10^{-6}, +0.13 \cdot 10^{-5}$
15. $y = 3 \cos x - 2 \cos^2 x$; error $\cdot\, 10^7$: 0.18, 0.74, 1.73, 3.28, 5.59, 9.04, 14.3, 22.8, 36.8, 61.4
17. $y' = 1/(2 - x^4)$; error $\cdot\, 10^9$: 0.2, 3.1, 10.7, 23.2, 28.5, −32.3, −376, −1656, −3489, +80444
19. Errors for Euler–Cauchy 0.02002, 0.06286, 0.05074; for improved Euler–Cauchy −0.000455, 0.012086, 0.009601; for Runge–Kutta. 0.0000011, 0.000016, 0.000536

Problem Set 21.2, page 915

1. $y = e^x$, $y_5^* = 1.648717$, $y_5 = 1.648722$, $\epsilon_5 = -3.8 \cdot 10^{-8}$, $y_{10}^* = 2.718276$, $y_{10} = 2.718284$, $\epsilon_{10} = -1.8 \cdot 10^{-6}$
3. $y = \tan x$, $y_4, \cdots, y_{10}$ (error $\cdot\, 10^5$) 0.422798 (−0.49), 0.546315 (−1.2), 0.684161 (−2.4), 0.842332 (−4.4), 1.029714 (−7.5), 1.260288 (−13), 1.557626 (−22)
5. RK error smaller in absolute value, error $\cdot\, 10^5 = 0.4, 0.3, 0.2, 5.6$ (for $x = 0.4, 0.6, 0.8, 1.0$)
7. $y = 1/(4 + e^{-3x})$, $y_4, \cdots, y_{10}$ (error $\cdot\, 10^5$) 0.232490 (0.34), 0.236787 (0.44), 0.240075 (0.42), 0.242570 (0.35), 0.244453 (0.25), 0.245867 (0.16), 0.246926 (0.09)
9. $y = \exp(x^3) - 1$, $y_4, \cdots, y_{10}$ (error $\cdot\, 10^7$) 0.008032 (−4), 0.015749 (−10), 0.027370 (−17), 0.043810 (−26), 0.066096 (−39), 0.095411 (−54), 0.133156 (−74)
13. $y = \exp(x^2)$. Errors $\cdot\, 10^5$ from $x = 0.3$ to 0.7: −5, −11, −19, −31, −41
15. **(a)** 0, 0.02, 0.0884, 0.215848, $y_4 = 0.417818$, $y_5 = 0.708887$ (poor)
(b) By 30–50%

Problem Set 21.3, page 922

1. $y_1 = -e^{-2x} + 4e^x$, $y_2 = -e^{-2x} + e^x$; errors of y_1 (of y_2) from 0.002 to 0.5 (from −0.01 to 0.1), monotone
3. $y_1' = y_2$, $y_2' = -\frac{1}{4}y_1$, $y = y_1 = 1$, 0.99, 0.97, 0.94, 0.9005, error −0.005, −0.01, −0.015, −0.02, −0.0229; exact $y = \cos \frac{1}{2}x$
5. $y_1' = y_2$, $y_2' = y_1 + x$, $y_1(0) = 1$, $y_2(0) = -2$, $y = y_1 = e^{-x} - x$, $y = 0.8$ (error 0.005), 0.61 (0.01), 0.429 (0.012), 0.2561 (0.0142), 0.0905 (0.0160)
7. By about a factor 10^5. $\epsilon_n(y_1) \cdot 10^6 = -0.082, \cdots, -0.27$, $\epsilon_n(y_2) \cdot 10^6 = 0.08, \cdots, 0.27$
9. Errors of y_1 (of y_2) from $0.3 \cdot 10^{-5}$ to $1.3 \cdot 10^{-5}$ (from $0.3 \cdot 10^{-5}$ to $0.6 \cdot 10^{-5}$)
11. $(y_1, y_2) = (0, 1), (0.20, 0.98), (0.39, 0.92), \cdots, (-0.23, -0.97), (-0.42, -0.91), (-0.59), (-0.81)$; continuation will give an "ellipse."

Problem Set 21.4, page 930

3. $-3u_{11} + u_{12} = -200, \quad u_{11} - 3u_{12} = -100$

5. 105, 155, 105, 115; Step 5: 104.94, 154.97, 104.97, 114.98

7. 0, 0, 0, 0. All equipotential lines meet at the corners (why?).
Step 5: 0.29298, 0.14649, 0.14649, 0.073245

9. 0.108253, 0.108253, 0.324760, 0.324760; Step 10: 0.108538, 0.108396, 0.324902, 0.324831

11. **(a)** $u_{11} = -u_{12} = -66$. **(b)** Reduce to 4 equations by symmetry.
$u_{11} = u_{31} = -u_{15} = -u_{35} = -92.92$, $u_{21} = -u_{25} = -87.45$,
$u_{12} = u_{32} = -u_{14} = -u_{34} = -64.22$, $u_{22} = -u_{24} = -53.98$,
$u_{13} = u_{23} = u_{33} = 0$

13. $u_{12} = u_{32} = 31.25, \quad u_{21} = u_{23} = 18.75, \quad u_{jk} = 25$ at the others

15. $u_{21} = u_{23} = 0.25, \quad u_{12} = u_{32} = -0.25, \quad u_{jk} = 0$ otherwise

17. $\sqrt{3}$, $u_{11} = u_{21} = 0.0849$, $u_{12} = u_{22} = 0.3170$. (0.1083, 0.3248 are 4S-values of the solution of the linear system of the problem.)

Problem Set 21.5, page 935

5. $u_{11} = 0.766, \quad u_{21} = 1.109, \quad u_{12} = 1.957, u_{22} = 3.293$

7. **A**, as in Example 1, right sides $-220, -220, -220, -220$.
Solution $u_{11} = u_{21} = 125.7$, $u_{21} = u_{22} = 157.1$

13. $-4u_{11} + u_{21} + u_{12} = -3$, $u_{11} - 4u_{21} + u_{22} = -12$, $u_{11} - 4u_{12} + u_{22} = 0$,
$2u_{21} + 2u_{12} - 12u_{22} = -14$, $u_{11} = u_{22} = 2$, $u_{21} = 4$, $u_{12} = 1$.
Here $-\frac{14}{3} = -\frac{4}{3}(1 + 2.5)$ with $\frac{4}{3}$ from the stencil.

15. $\mathbf{b} = [-200, -100, -100, 0]^{\mathsf{T}}; \quad u_{11} = 73.68$, $u_{21} = u_{12} = 47.37$, $u_{22} = 15.79$ (4S)

Problem Set 21.6, page 941

5. 0, 0.6625, 1.25, 1.7125, 2, 2.1, 2, 1.7125, 1.25, 0.6625, 0

7. Substantially less accurate, 0.15, 0.25 ($t = 0.04$), 0.100, 0.163 ($t = 0.08$)

9. Step 5 gives 0, 0.06279, 0.09336, 0.08364, 0.04707, 0.

11. Step 2: 0 (exact 0), 0.0453 (0.0422), 0.0672 (0.0658), 0.0671 (0.0628), 0.0394 (0.0373), 0 (0)

13. 0.3301, 0.5706, 0.4522, 0.2380 ($t = 0.04$), 0.06538, 0.10603, 0.10565, 0.6543 ($t = 0.20$)

15. 0.1018, 0.1673, 0.1673, 0.1018 ($t = 0.04$), 0.0219, 0.0355, $\cdots$ ($t = 0.20$)

Problem Set 21.7, page 944

1. $u(x, 1) = 0, -0.05, -0.10, -0.15, -0.20, 0$

3. For $x = 0.2, 0.4$ we obtain 0.24, 0.40 ($t = 0.2$), 0.08, 0.16 ($t = 0.4$), $-0.08, -0.16$ ($t = 0.6$), etc.

5. 0, 0.354, 0.766, 1.271, 1.679, 1.834, $\cdots$ ($t = 0.1$); 0, 0.575, 0.935, 1.135, 1.296, 1.357, $\cdots$ ($t = 0.2$)

7. 0.190, 0.308, 0.308, 0.190, (3S-exact: 0.178, 0.288, 0.288, 0.178)

Chapter 21 Review Questions and Problems, page 945

17. $y = e^x$, 0.038, 0.125 (errors of y_5 and y_{10})
19. $y = \tan x$; 0 (0), 0.10050 (−0.00017), 0.20304 (−0.00033), 0.30981 (−0.00048), 0.42341 (−0.00062), 0.54702 (−0.00072), 0.68490 (−0.00076), 0.84295 (−0.00066), 1.0299 (−0.0002), 1.2593 (0.0009), 1.5538 (0.0036)
21. 0.1003346 $(0.8 \cdot 10^{-7})$ 0.2027099 $(1.6 \cdot 10^{-7})$, 0.3093360 $(2.1 \cdot 10^{-7})$, 0.4227930 $(2.3 \cdot 10^{-7})$, 0.5463023 $(1.8 \cdot 10^{-7})$
23. $y = \sin x$, $y_{0.8} = 0.717366$, $y_{1.0} = 0.841496$ (errors $-1.0 \cdot 10^{-5}$, $-2.5 \cdot 10^{-5}$)
25. $y_1' = y_2$, $y_2' = x^2 y_1$, $y = y_1 = 1, 1, 1, 1.0001, 1.0006, 1.002$
27. $y_1' = y_2$, $y_2' = 2e^x - y_1$, $y = e^x - \cos x$, $y = y_1 = 0, 0.241, 0.571, \cdots$; errors between 10^{-6} and 10^{-5}
29. 3.93, 15.71, 58.93
31. 0, 0.04, 0.08, 0.12, 0.15, 0.16, 0.15, 0.12, 0.08, 0.04, 0 ($t = 0.3$. 3 time steps)
33. $u(P_{11}) = u(P_{31}) = 270$, $u(P_{21}) = u(P_{13}) = u(P_{23}) = u(P_{33}) = 30$, $u(P_{12}) = u(P_{32}) = 90$, $u(P_{22}) = 60$
35. 0.043330, 0.077321, 0.089952, 0.058488 ($t = 0.04$), 0.010956, 0.017720, 0.017747, 0.010964 ($t = 0.20$)

Problem Set 22.1, page 953

3. $f(\mathbf{x}) = 2(x_1 - 1)^2 + (x_2 + 2)^2 - 6$; Step 3: (1.037, −1.926), value −5.992
9. Step 5: (0.11247, −0.00012), value 0.000016

Problem Set 22.2, page 957

7. No
9. x_3, x_4 is the unused time on M_1, M_2, respectively.
11. $f(2.5, 2.5) = 100$
13. $f(-\frac{11}{3}, \frac{26}{3}) = 198\frac{1}{3}$
15. $f(9, 6) = 360$
17. $0.5x_1 + 0.75x_2 \leqq 45$ (copper), $0.5x_1 + 0.25x_2 \leqq 30$, $f = 120x_1 + 100x_2$, $f_{\max} = f(45, 30) = 8400$
19. $f = x_1 + x_2$, $2x_1 + 3x_2 \leqq 1200$, $4x_1 + 2x_2 \leqq 1600$, $f_{\max} = f(300, 200) = 500$
21. $x_1/3 + x_2/2 \leqq 100$, $x_1/3 + x_2/6 \leqq 80$, $f = 150x_1 + 100x_2$, $f_{\max} = f(210, 60) =$ 37,500

Problem Set 22.3, page 961

3. $f(120/11, 60/11) = 480/11$
5. Eliminate in Column 3, so that 20 goes. $f_{\min} = f(0, \frac{1}{2}) = -10$.
7. $f_{\max} = f(\frac{60}{21}, 0, \frac{1500}{105}, 0) = \frac{2200}{7}$
9. $f_{\max} = 6$ on the segment from (3, 0, 0) to (0, 0, 2)
11. We minimize! The augmented matrix is

$$\mathbf{T}_0 = \begin{bmatrix} 1 & 1.8 & 2.1 & 0 & 0 & 0 \\ 0 & 15 & 30 & 1 & 0 & 150 \\ 0 & 600 & 500 & 0 & 1 & 3900 \end{bmatrix}.$$

The pivot is 600. The calculation gives

$$\mathbf{T}_1 = \begin{bmatrix} 1 & 0 & \frac{6}{10} & 0 & -\frac{3}{1000} & -\frac{117}{10} \\ 0 & 0 & \frac{35}{2} & 1 & -\frac{1}{40} & \frac{105}{2} \\ 0 & 600 & 500 & 0 & 1 & 3900 \end{bmatrix} \begin{array}{l} \text{Row 1} - \frac{1.8}{600}\,\text{Row 3} \\ \text{Row 2} - \frac{15}{600}\,\text{Row 3} \\ \text{Row 3} \end{array}$$

The next pivot is $\frac{35}{2}$. The calculation gives

$$\mathbf{T}_2 = \begin{bmatrix} 1 & 0 & 0 & -\frac{6}{175} & -\frac{3}{1400} & -\frac{27}{2} \\ 0 & 0 & \frac{35}{2} & 1 & -\frac{1}{40} & \frac{105}{2} \\ 0 & 600 & 0 & -\frac{200}{7} & \frac{12}{7} & 2400 \end{bmatrix} \begin{array}{l} \text{Row 1} - \frac{1.2}{35}\,\text{Row 2} \\ \text{Row 2} \\ \text{Row 3} - \frac{1000}{35}\,\text{Row 2} \end{array}$$

Hence $-f$ has the maximum value -13.5, so that f has the minimum value 13.5, at the point

$$(x_1, x_2) = \left(\frac{2400}{600}, \frac{105/2}{35/2}\right) = (4, 3).$$

13. $f_{\max} = f(5, 4, 6) = 478$

Problem Set 22.4, page 968

1. $f(6, 3) = 84$
3. $f(20, 20) = 40$
5. $f(10, 5) = 5500$
7. $f(1, 1, 0) = 13$
9. $f(4, 0, \frac{1}{2}) = 9$

Chapter 22 Review Questions and Problems, page 968

9. Step 5: $[0.353 \quad -0.028]^{\mathsf{T}}$. Slower. Why?
11. Of course! Step 5: $[-1.003 \quad 1.897]^{\mathsf{T}}$
17. $f(2, 4) = 100$
19. $f(3, 6) = -54$

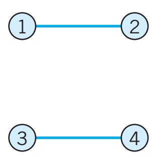
Problem Set 23.1, page 974

9. $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

13. $\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

15.



17. If G is complete.

19.
$$
\text{Vertex}\quad
\begin{array}{c|c}
 & \text{Edge} \\
 & \begin{array}{cccc} e_1 & e_2 & e_3 & e_4 \end{array} \\
\hline
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} &
\begin{bmatrix}
-1 & -1 & 1 & -1 \\
1 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

Problem Set 23.2, page 979

1. 5

3. 4

5. The idea is to go backward. There is a v_{k-1} adjacent to v_k and labeled $k-1$, etc. Now the only vertex labeled 0 is s. Hence $\lambda(v_0) = 0$ implies $v_0 = s$, so that $v_0 - v_1 - \cdots - v_{k-1} - v_k$ is a path $s \to v_k$ that has length k.

15. Delete the edge (2, 4).

17. No

Problem Set 23.3, page 983

1. (1, 2), (2, 4), (4, 3); $L_2 = 12, L_3 = 36, L_4 = 28$

5. (1, 2), (2, 4), (3, 4), (3, 5); $L_2 = 2, L_3 = 4, L_4 = 3, L_5 = 6$

7. (1, 2), (2, 4), (3, 4); $L_2 = 10, L_3 = 15, L_4 = 13$

9. (1, 5), (2, 3), (2, 6), (3, 4), (3, 5); $L_2 = 9, L_3 = 7, L_4 = 8, L_5 = 4, L_6 = 14$

Problem Set 23.4, page 987

1.
$$
\begin{array}{l} 2 \diagdown \\ \;\; 4 - 3 - 5 \\ 1 \diagup \end{array} \qquad L = 10
$$

3.
$$
5 - 3 - 6 \begin{array}{l} \diagup\, 1 \\ \\ \diagdown\, 2 - 4 \end{array} \qquad L = 17
$$

5.
$$
1 \begin{array}{l} \diagup\, 2 \\ \\ \diagdown\, 4 \begin{array}{l} \diagup\, 3 \\ \diagdown\, 5 \end{array} \end{array} \qquad L = 12
$$

9. Yes

11.
$$
1 - 3 - 4 \begin{array}{l} \diagup\, 2 \\ \\ \diagdown\, 5 - 6 \end{array} \qquad L = 38
$$

13. New York–Washington–Chicago–Dalles–Denver–Los Angeles

15. G is connected. If G were not a tree, it would have a cycle, but this cycle would provide two paths between any pair of its vertices, contradicting the uniqueness.

19. If we add an edge (u, v) to T, then since T is connected, there is a path $u \to v$ in T which, together with (u, v), forms a cycle.

Problem Set 23.5, page 990

1. If G is a tree.
3. A shortest spanning tree of the largest connected graph that contains vertex 1.
7. (1, 4), (1, 3), (1, 2), (2, 6), (3, 5); $L = 32$
9. (1, 4), (4, 3), (4, 2), (3, 5); $L = 20$
11. (1, 4), (4, 3), (4, 5), (1, 2); $L = 12$

Problem Set 23.6, page 997

1. $\{3, 6\}$, $11 + 3 = 14$
3. $\{4, 5, 6\}$, $10 + 5 + 13 = 28$
5. $\{3, 6, 7\}$, $8 + 4 + 4 = 16$
7. $S = \{1, 4\}$, $8 + 6 = 14$
9. One is interested in flows *from s to t,* not in the opposite direction.
13. $\Delta_{12} = 5, \Delta_{24} = 8, \Delta_{45} = 2$; $\Delta_{12} = 5, \Delta_{25} = 3$; $\Delta_{13} = 4, \Delta_{35} = 9$
P_1: $1 - 2 - 4 - 5, \Delta f = 2$; P_2: $1 - 2 - 5, \Delta f = 3$; P_3: $1 - 3 - 5, \Delta f = 4$
15. $1 - 2 - 5, \Delta f = 2$; $1 - 4 - 2 - 5, \Delta f = 2$, etc.
17. $f_{13} = f_{35} = 8$, $f_{14} = f_{45} = 5$, $f_{12} = f_{24} = f_{46} = 4$, $f_{56} = 13$, $f = 4 + 13 = 17$, $f = 17$ is unique.
19. For instance, $f_{12} = 10$, $f_{24} = f_{45} = 7$, $f_{13} = f_{25} = 5$, $f_{35} = 3$, $f_{32} = 2$, $f = 3 + 5 + 7 = 15$, $f = 15$ is unique.

Problem Set 23.7, page 1000

3. (2, 3) and (5, 6)
5. By considering only edges with one labeled end and one unlabeled end
7. $1 - 2 - 5, \Delta_t = 2$; $1 - 4 - 2 - 5, \Delta_t = 1$; $f = 6 + 2 + 1 = 9$, where 6 is the given flow
9. $1 - 2 - 4 - 6, \Delta_t = 2$; $1 - 3 - 5 - 6, \Delta_t = 1$; $f = 4 + 2 + 1 = 7$, where 4 is the given flow
15. $S = \{1, 2, 4, 5\}$, $T = \{3, 6\}$, $\operatorname{cap}(S, T) = 14$

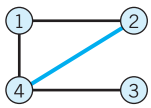
Problem Set 23.8, page 1005

1. No
3. No
5. Yes, $S = \{1, 4, 5, 8\}$
7. Yes, $S = \{1, 3, 5\}$
11. $1 - 2 - 3 - 7 - 5 - 4$
13. $1 - 2 - 3 - 7 - 5 - 4$ is augmenting and gives $1 - 2 - 3 - 7 - 5 - 4$ and (1, 2), (3, 7), (5, 4) is of maximum cardinality.
15. $1 - 4 - 3 - 6 - 7 - 8$ is augmenting and gives $1 - 4 - 3 - 6 - 7 - 8$ and (1, 4), (3, 6), (7, 8) is of maximum cardinality.
19. 3
21. 2
23. 3
25. K_4

Chapter 23 Review Questions and Problems, page 1006

11. $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$

13.
$$\begin{array}{ll} & \text{To vertex} \quad 1 \quad\ 2 \quad\ 3 \quad\ 4 \\ \text{From vertex} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

15. 

17.

Vertex	Incident Edges
1	(1, 2), (1, 4)
2	(2, 1), (2, 4)
3	(3, 4)
4	(4, 1), (4, 2), (4, 3)

19. (1, 2), (1, 4), (2, 3); $L_2 = 2, L_3 = 5, L_4 = 5$

23. (1, 6), (4, 5), (2, 3), (7, 8)

Problem Set 24.1, page 1015

1. $q_L = 19, q_M = 20, q_U = 20.5$

3. $q_L = 138, q_M = 144, q_U = 154$

5. $q_L = 199, q_M = 201, q_U = 201$

7. $q_L = 1.3, q_M = 1.4, q_U = 1.45$

9. $q_L = 89.9, q_M = 91.0, q_U = 91.8$

11. $\bar{x} = 19.875, s = 0.835$, IQR $= 1.5$

13. $\bar{x} = 144.67, s = 8.9735$, IQR $= 16$

15. $\bar{x} = 1.355, s = 0.136$, IQR $= 0.15$

17. 3.54, 1.29

Problem Set 24.2, page 1017

1. 2^3 outcomes: *RRR, RRL, RLR, LRR, RLL, LRL, LLR, LLL*

3. $6^2 = 36$ outcomes (1, 1), (1, 2), $\cdots$, (6, 6), first number (second number) referring to the first die (second die)

5. Infinitely many outcomes *H TH TTH TTTH* $\cdots$ (*H* = *Head*, *T* = *Tail*)

7. The space of ordered pairs of numbers

9. 10 outcomes: *D ND NND* $\cdots$ *NNNNNNNNND*

11. Yes

17. $A \cup B = B$ implies $A \subseteq B$ by the definition of union. Conversely. $A \subseteq B$ implies that $A \cup B = B$ because always $B \subseteq A \cup B$, and if $A \subseteq B$, we must have equality in the previous relation.

Problem Set 24.3, page 1024

1. $1 - 4/216 = 98.15\%$, by Theorem 1
3. **(a)** $0.9^3 = 72.9\%$, **(b)** $\frac{90}{100} \cdot \frac{89}{99} \cdot \frac{88}{98} = 72.65\%$
5. $\frac{8}{9}$
7. Small sample from a large population containing *many* items in each class we are interested in (defectives and nondefectives, etc.)
9. $\frac{498}{500} \cdot \frac{497}{499} \cdot \frac{496}{498} \cdot \frac{495}{497} \cdot \frac{494}{496} \approx 0.98008$
11. **(a)** $\frac{100}{200} \cdot \frac{99}{199} = 24.874\%$, **(b)** $\frac{100}{200} \cdot \frac{100}{199} + \frac{100}{200} \cdot \frac{100}{199} = 50.25\%$, **(c)** same as (a). (a) + (b) + (c) = 1. Why?
13. $1 - 0.96^3 = 11.5\%$
15. $1 - 0.875^4 = 0.4138 < 1 - 0.75^2 = 0.4375 < 0.5 \quad (c < b < a)$
17. $A = B \cup (A \cap B^c)$, hence $P(A) = P(B) + P(A \cap B^c) \geqq P(B)$ by disjointedness of B and $A \cap B^c$

Problem Set 24.4, page 1028

1. In $10! = 3{,}628{,}800$ ways
3. $\frac{2}{6} \cdot \frac{1}{5} \cdot \frac{4}{4} \cdot \frac{3}{3} \cdot \frac{2}{2} \cdot \frac{1}{1} = \frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} \cdot \frac{1}{1} = \frac{4!2!}{6!} = \frac{2}{6} \cdot \frac{1}{5} = \frac{1}{15}$
5. $\binom{10}{3}\binom{5}{2}\binom{6}{2} = 18{,}000$
7. 210, 70, 112, 28
9. In $6!/6 = 120$ ways
11. $9 \cdot 8 = 72$
13. **(b)** $1/(12n)$
15. P (*No two people have a birthday in common*) $= 365 \cdot 364 \cdots 346/365^{20} = 0.59$. *Answer:* 41%, which is surprisingly large.

Problem Set 24.5, page 1034

1. $k = \frac{1}{55}$ by (6)
3. $k = \frac{1}{4}$ by (10), $P(0 \leqq X \leqq 2) = \frac{1}{2}$
5. No, because of (6)
7. $k = \frac{1}{100}$ because of (6) and $1 + 8 + 27 + 64 = 100$
9. $k = 5$; 50%
11. $0.5^3 = 12.5\%$
13. $F(x) = 0$ if $x < -1$, $F(x) = \frac{1}{2}(x + 1)^2$ if $-1 \leqq x < 0$
$F(x) = 1 - \frac{1}{2}(x - 1)^2$ if $0 \leqq x < 1$, $F(x) = 1$ if $x \leqq 1$
Answer: 500 cans, $P = 0.125, 0$
15. $X > b, X \geqq b, X < c, X \leqq c$, etc.

Problem Set 24.6, page 1038

1. $k = \frac{1}{2}, \mu = \frac{4}{3}, \sigma^2 = \frac{2}{9}$
3. $\mu = \pi, \sigma^2 = \pi^2/3$; cf. Example 2
5. $\mu = \frac{1}{4}, \sigma^2 = \frac{1}{16}$
7. $C = \frac{1}{2}, \mu = 2, \sigma^2 = 4$
9. 750, 1, 0.002
11. $c = 0.073$
13. \$643.50
15. $\frac{1}{2}, \frac{1}{20}, (X - \frac{1}{2})\sqrt{20}$
17. $X =$ *Product of the 2 numbers.* $E(X) = 12.25$, 12 cents
19. $(0 + 1 \cdot 3 + 3 \cdot 8 + 1 \cdot 27)/8 = 54/8 = 6 \cdot 75$

Problem Set 24.7, page 1044

3. 38%
5. $\binom{5}{x}\ 0.5^5$, 0.03125, 0.15625, $1 - f(0) = 0.96875$, 0.96875
7. 0.265
9. $f(x) = 0.5^x e^{-0.5}/x!$, $f(0) + f(1) = e^{-0.5}(1.0 + 0.5) = 0.91$. *Answer:* 9%
11. $13\frac{1}{4}$ %
13. 42%, 47.2%, 10.5%, 0.3%
15. $1 - e^{-0.2} = 18\%$

Problem Set 24.8, page 1050

1. 0.1587, 0.5, 0.6915, 0.6247
3. 45.065, 56.978, 2.022
5. 15.9%
7. 31.1%, 95.4%
9. About 58%
11. $t = 1084$ hours
13. About 683 (Fig. 521a)

Problem Set 24.9, page 1059

1. $\frac{1}{8}, \frac{3}{16}, \frac{3}{8}$
3. $\frac{2}{9}, \frac{1}{9}, \frac{1}{2}$
5. $f_2(y) = 1/(\beta_2 - \alpha_2)$ if $\alpha_2 < y < \beta_2$
7. 27.45 mm, 0.38 mm
11. 25.26 cm, 0.0078 cm
13. 50%
15. The distributions in Prob. 17 and Example 1
17. No

Chapter 24 Review Questions and Problems, page 1060

11. $Q_L = 110$, $Q_M = 112$, $Q_U = 115$
13. $\bar{x} = 111.9$, $s = 4.0125$, $s^2 = 16.1$
21. $x_{\min} \leqq x_j \leqq x_{\max}$. Sum over j from 1.
17. $\bar{x} = 6$, $s = 3.65$
19. $f(x) = \binom{50}{x}0.03^x 0.97^{50-x} \approx 1.5^x e^{-1.5}/x!$
21. $f(x) = 2^{-x}$, $x = 1, 2, \cdots$
23. $1, \frac{1}{2}$
25. 0.1587, 0.6306, 0.5, 0.4950

Problem Set 25.2, page 1067

1. In Example 1, $\mu = 0$ so $\sum_{j=1}^{n} x_j = 0$. $\partial \ln \ell/\partial \ell = 0$ and $\tilde{\sigma}^2$ is as before.
3. $\ell = e^{-n\mu}\mu^{(x_1 + \cdots + x_n)}/(x_1! \cdots x_n!)$, $\partial \ln \ell/\partial \mu = -n + (x_1 + \cdots + x_n)/\mu = 0$, $n\hat{\mu} = n\bar{x}$, $\hat{\mu} = \bar{x} = 15.3$
5. $l = p^k(1 - p)^{n-k}$, $\hat{p} = k/n$, $k =$ number of successes in n trails
7. 7/12
9. $l = f = p(1 - p)^{x-1}$, etc., $\hat{p} = 1/x$
11. $\hat{\theta} = n/\Sigma x_j = 1/\bar{x}$
13. $\hat{\theta} = 1$
15. Variability larger than perhaps expected

Problem Set 25.3, page 1077

3. Shorter by a factor $\sqrt{2}$ **5.** 4, 16

7. $c = 1.96, \bar{x} = 126, s^2 = 126 \cdot 674/800 = 106.155, k = cs/\sqrt{n} = 0.714$, $\text{CONF}_{0.95}\{125.3 \leqq \mu \leqq 126.7\}$, $\text{CONF}_{0.95}\{0.1566 \leqq p \leqq 0.1583\}$

9. $\text{CONF}_{0.99}\{63.72 \leqq \mu \leqq 66.28\}$

11. $n - 1 = 5, F(c) = 0.995, c = 4.03, \bar{x} = 9533.33, s^2 = 49{,}666.67$, $k = 366.66$ (Table 25.2), $\text{CONF}_{0.99}\{9166.7 \leqq \mu \leqq 9900\}$

13. $\text{CONF}_{0.95}\{0.023 \leqq \sigma^2 \leqq 0.085\}$

15. $n - 1 = 99$ degrees of freedom. $F(c_1) = 0.025$, $c_1 = 74.2$, $F(c_2) = 0.975$, $c_2 = 129.6$. Hence $k_1 = 12.41$, $k_2 = 7.10$. $\text{CONF}_{0.95}\{7.10 \leqq \sigma^2 \leqq 12.41\}$.

17. $\text{CONF}_{0.95}\{0.74 \leqq \sigma^2 \leqq 5.19\}$

19. $Z = X + Y$ is normal with mean 105 and variance 1.25. *Answer:* $P(104 \leqq Z \leqq 106) = 63\%$

Problem Set 25.4, page 1086

3. $t = (0.286 - 0)/(4.31/\sqrt{7}) = 0.18 < c = 1.94$; accept the hypothesis.

5. $c = 6090 > 6019$: do not reject the hypothesis.

7. $\sigma^2/n = 1.8$, $c = 57.8$, accept the hypothesis.

9. $\mu < 58.69$ or $\mu > 61.31$

11. Alternative $\mu \neq 5000$, $t = (4990 - 5000)/(20/\sqrt{50}) = -3.54 < c = -2.01$ (Table A9, Appendix 5). Reject the hypothesis $\mu = 5000$ g.

13. Two-sided. $t = (0.55 - 0)/\sqrt{0.546/8} = 2.11 < c = 2.37$ (Table A9, Appendix 5), no difference

15. $19 \cdot 1.0^2/0.8^2 = 29.69 < c = 30.14$ (Table A10. Appendix 5), accept the hypothesis

17. By (12), $t_0 = \sqrt{16}(20.2 - 19.6)/\sqrt{0.16 + 0.36} > c = 1.70$. Assert that B is better.

Problem Set 25.5, page 1091

1. $\text{LCL} = 1 - 2.58 \cdot 0.02/2 = 0.974$, $\text{UCL} = 1.026$

3. 27

5. Choose 4 times the original sample size

9. $2.58\sqrt{0.0004}/\sqrt{2} = 0.036$, $\text{LCL} = 3.464$, $\text{UCL} = 3.536$

11. $\text{LCL} = np - 3\sqrt{np(1 - p)}$, $\text{CL} = np$, $\text{UCL} = np + 3\sqrt{np(1 - p)}$

13. In about 30% (5%) of the cases

15. $\text{LCL} = \mu - 3\sqrt{\mu}$ is negative in (b) and we set $\text{LCL} = 0$, $\text{CL} = \mu = 3.6$, $\text{UCL} = \mu + 3\sqrt{\mu} = 9.3$.

Problem Set 25.6, page 1095

1. 0.9825, 0.9384, 0.4060 **3.** 0.8187, 0.6703, 0.1353

5. $e^{-25\theta}(1 + 25\theta)$, $P(A; 1.5) = 94.5$, $\alpha = 5.5\%$ **7.** 19.5%, 14.7%

9. $(1 - \theta)^n + n\theta(1 - \theta)^{n-1}$ **11.** $(1 - \frac{1}{2})^3 + 3 \cdot \frac{1}{2}(1 - \frac{1}{2})^2 = \frac{1}{2}$

13. $\displaystyle\sum_{x=0}^{9} \binom{100}{x} 0.12^x\, 0.88^{100-x} = 22\%$ (by the normal approximation)

15. $(1 - \theta)^5$, $[\theta(1 - \theta)^{5-1}]' = 0$, $\theta = \frac{1}{6}$, $\text{AOQL} = 6.7\%$

Problem Set 25.7, page 1099

3. $\chi_0^2 = (40 - 50)^2/50 + (60 - 50)^2/50 = 4 > c = 3.84$; no
5. $\chi_0^2 = \frac{16}{10} > 11.07$; yes
7. $\chi_0^2 = 10.264 < 11.07$; yes
9. 42 even digits, accept.
13. $\chi_0^2 = \dfrac{(355 - 358.5)^2}{358.5} + \dfrac{(123 - 119.5)^2}{119.5} = 0.137 < c = 3.84$ (1 degree of freedom, 95%)
15. Combining the last three nonzero values, we have $K - r - 1 = 9$ ($r = 1$ since we estimated the mean, $\frac{10{,}094}{2608} \approx 3.87$). $\chi_0^2 = 12.8 < c = 16.92$. Accept the hypothesis.

Problem Set 25.8, page 1102

3. $(\frac{1}{2})^8 + 8 \cdot (\frac{1}{2})^8 = 3.5\%$ is the probability that 7 cases in 8 trials favor A under the hypothesis that A and B are equally good. Reject.
5. $(\frac{1}{2})^{18}(1 + 18 + 153 + 816) = 0.0038$
7. $\bar{x} = 9.67$, $s = 11.87$. $t_0 = 9.67/(11.87/\sqrt{15}) = 3.16 > c = 1.76$ ($\alpha = 5\%$). Hypothesis rejected.
9. Hypothesis $\tilde{\mu} = 0$. Alternative $\tilde{\mu} > 0$, $\bar{x} = 1.58$, $t = \sqrt{10} \cdot 1.58/1.23 = 4.06 > c = 1.83$ ($\alpha = 5\%$). Hypothesis rejected.
11. Consider $y_j = x_j - \tilde{\mu}_0$.
13. $n = 8$; 4 transpositions, $P(T \leqq 4) = 0.007$. Assert that fertilizing increases yield.
15. $P(T \leqq 2) = 2.8\%$. Assert that there is an increase.

Problem Set 25.9, page 1111

1. $y = 0.98 + 0.495x$
3. $y = -11{,}457.9 + 43.2x$
5. $y = -10 + 0.55x$
7. $y = 0.5932 + 0.1138x$, $R = 1/0.1138$
9. $y = 0.32923 + 0.00032x$, $y(66) = 0.35035$
13. $c = 3.18$ (Table A9), $k_1 = 43.2$, $q_0 = 54{,}878$, $K = 1.502$, $\text{CONF}_{0.95}\{41.7 \leqq \kappa_1 \leqq 44.7\}$.
15. $y - 1.875 = 0.067(x - 25)$, $3s_x^2 = 500$, $q_0 = 0.023$, $K = 0.021$, $\text{CONF}_{0.95}\{0.046 \leqq \kappa_1 \leqq 0.088\}$

Chapter 25 Review Questions and Problems, page 1111

15. $\hat{\mu} = 20.325$, $\hat{\sigma}^2 = (\frac{7}{8})s^2 = 3.982$
17. $\text{CONF}_{0.99}\{27.94 \leqq \mu \leqq 34.81\}$
19. $c = 14.74 > 14.5$, reject μ_0; $\Phi((14.74 - 14.50)/\sqrt{0.025}) = 0.9353$
21. $2.58 \cdot \sqrt{0.00024}/\sqrt{2} = 0.028$, LCL $= 2.722$, UCL $= 2.778$
23. $\alpha = 1 - (1 - \theta)^6 = 5.85\%$, when $\theta = 0.01$. For $\theta = 15\%$ we obtain $\beta = (1 - \theta)^6 = 37.7\%$. If n increases, so does α, whereas β decreases.
25. $y = 3.4 - 1.85x$

APPENDIX 3

Auxiliary Material

A3.1 Formulas for Special Functions

For tables of numeric values, see Appendix 5.

Exponential function e^x (Fig. 545)

$$e = 2.71828\ 18284\ 59045\ 23536\ 02874\ 71353$$

$$\text{(1)} \qquad e^x e^y = e^{x+y}, \qquad e^x/e^y = e^{x-y}, \qquad (e^x)^y = e^{xy}$$

Natural logarithm (Fig. 546)

$$\text{(2)} \qquad \ln (xy) = \ln x + \ln y, \qquad \ln (x/y) = \ln x - \ln y, \qquad \ln (x^a) = a \ln x$$

$\ln x$ is the inverse of e^x, and $e^{\ln x} = x$, $e^{-\ln x} = e^{\ln (1/x)} = 1/x$.

Logarithm of base ten $\log_{10} x$ or simply $\log x$

$$\text{(3)} \qquad \log x = M \ln x, \qquad M = \log e = 0.43429\ 44819\ 03251\ 82765\ 11289\ 18917$$

$$\text{(4)} \qquad \ln x = \frac{1}{M} \log x, \qquad \frac{1}{M} = \ln 10 = 2.30258\ 50929\ 94045\ 68401\ 79914\ 54684$$

$\log x$ is the inverse of 10^x, and $10^{\log x} = x$, $10^{-\log x} = 1/x$.

Sine and cosine functions (Figs. 547, 548). In calculus, angles are measured in radians, so that $\sin x$ and $\cos x$ have period 2π.

$\sin x$ is odd, $\sin (-x) = -\sin x$, and $\cos x$ is even, $\cos (-x) = \cos x$.

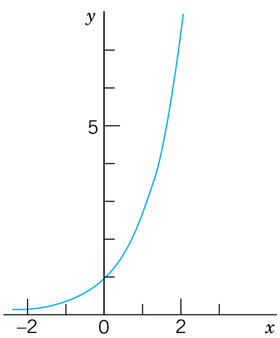


Fig. 545. Exponential function e^x

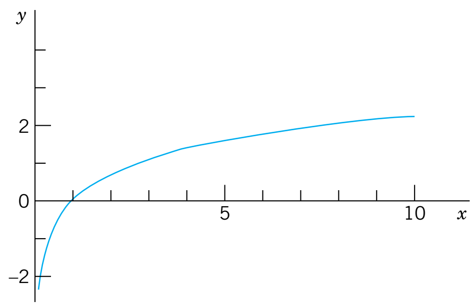


Fig. 546. Natural logarithm $\ln x$

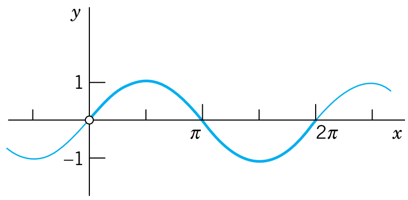


Fig. 547. sin x

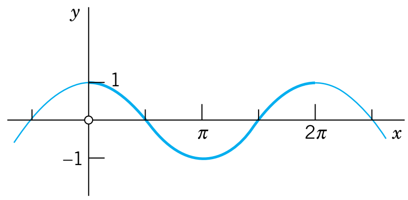


Fig. 548. cos x

$$1° = 0.01745\ 32925\ 19943 \text{ radian}$$

$$1 \text{ radian} = 57°\ 17'\ 44.80625'' = 57.29577\ 95131°$$

(5) $$\sin^2 x + \cos^2 x = 1$$

(6) $$\begin{cases} \sin(x+y) = \sin x \cos y + \cos x \sin y \\ \sin(x-y) = \sin x \cos y - \cos x \sin y \\ \cos(x+y) = \cos x \cos y - \sin x \sin y \\ \cos(x-y) = \cos x \cos y + \sin x \sin y \end{cases}$$

(7) $$\sin 2x = 2 \sin x \cos x, \qquad \cos 2x = \cos^2 x - \sin^2 x$$

(8) $$\begin{cases} \sin x = \cos\left(x - \dfrac{\pi}{2}\right) = \cos\left(\dfrac{\pi}{2} - x\right) \\ \cos x = \sin\left(x + \dfrac{\pi}{2}\right) = \sin\left(\dfrac{\pi}{2} - x\right) \end{cases}$$

(9) $$\sin(\pi - x) = \sin x, \qquad \cos(\pi - x) = -\cos x$$

(10) $$\cos^2 x = \tfrac{1}{2}(1 + \cos 2x), \qquad \sin^2 x = \tfrac{1}{2}(1 - \cos 2x)$$

(11) $$\begin{cases} \sin x \sin y = \tfrac{1}{2}[-\cos(x+y) + \cos(x-y)] \\ \cos x \cos y = \tfrac{1}{2}[\cos(x+y) + \cos(x-y)] \\ \sin x \cos y = \tfrac{1}{2}[\sin(x+y) + \sin(x-y)] \end{cases}$$

(12) $$\begin{cases} \sin u + \sin v = 2 \sin \dfrac{u+v}{2} \cos \dfrac{u-v}{2} \\ \cos u + \cos v = 2 \cos \dfrac{u+v}{2} \cos \dfrac{u-v}{2} \\ \cos v - \cos u = 2 \sin \dfrac{u+v}{2} \sin \dfrac{u-v}{2} \end{cases}$$

(13) $$A \cos x + B \sin x = \sqrt{A^2 + B^2} \cos(x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \mp \frac{B}{A}$$

(14) $$A \cos x + B \sin x = \sqrt{A^2 + B^2} \sin(x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \pm \frac{A}{B}$$

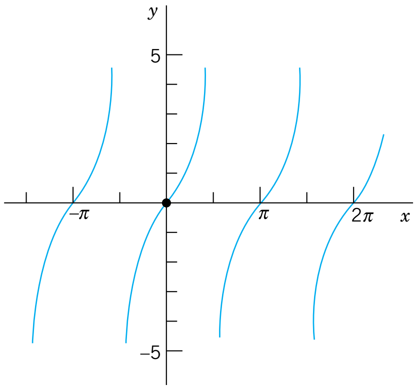


Fig. 549. tan x

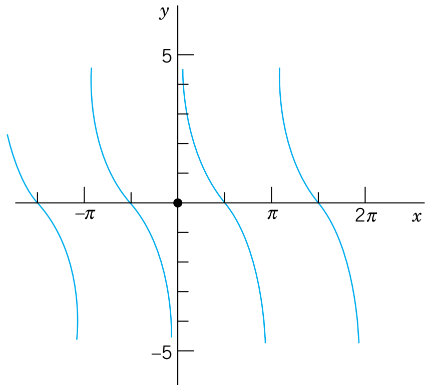


Fig. 550. cot x

Tangent, cotangent, secant, cosecant (Figs. 549, 550)

$$\text{(15)} \quad \tan x = \frac{\sin x}{\cos x}, \qquad \cot x = \frac{\cos x}{\sin x}, \qquad \sec x = \frac{1}{\cos x}, \qquad \csc x = \frac{1}{\sin x}$$

$$\text{(16)} \quad \tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}, \qquad \tan(x - y) = \frac{\tan x - \tan y}{1 + \tan x \tan y}$$

Hyperbolic functions (hyperbolic sine sinh x, etc.; Figs. 551, 552)

$$\text{(17)} \quad \sinh x = \tfrac{1}{2}(e^x - e^{-x}), \qquad \cosh x = \tfrac{1}{2}(e^x + e^{-x})$$

$$\text{(18)} \quad \tanh x = \frac{\sinh x}{\cosh x}, \qquad \coth x = \frac{\cosh x}{\sinh x}$$

$$\text{(19)} \quad \cosh x + \sinh x = e^x, \qquad \cosh x - \sinh x = e^{-x}$$

$$\text{(20)} \quad \cosh^2 x - \sinh^2 x = 1$$

$$\text{(21)} \quad \sinh^2 x = \tfrac{1}{2}(\cosh 2x - 1), \qquad \cosh^2 x = \tfrac{1}{2}(\cosh 2x + 1)$$

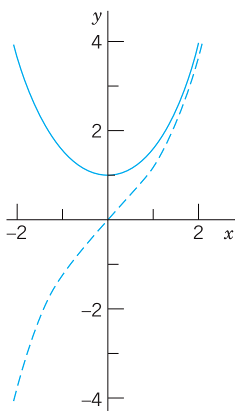


Fig. 551. sinh x (dashed) and cosh x

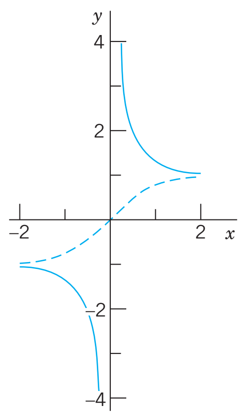


Fig. 552. tanh x (dashed) and coth x

$$\left\{\begin{aligned} \sinh(x \pm y) &= \sinh x \cosh y \pm \cosh x \sinh y \\ \cosh(x \pm y) &= \cosh x \cosh y \pm \sinh x \sinh y \end{aligned}\right. \tag{22}$$

$$\tanh(x \pm y) = \frac{\tanh x \pm \tanh y}{1 \pm \tanh x \tanh y} \tag{23}$$

Gamma function (Fig. 553 and Table A2 in App. 5). The gamma function $\Gamma(\alpha)$ is defined by the integral

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, dt \qquad (\alpha > 0), \tag{24}$$

which is meaningful only if $\alpha > 0$ (or, if we consider complex α, for those α whose real part is positive). Integration by parts gives the important *functional relation of the gamma function,*

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha). \tag{25}$$

From (24) we readily have $\Gamma(1) = 1$; hence if α is a positive integer, say k, then by repeated application of (25) we obtain

$$\Gamma(k + 1) = k! \qquad (k = 0, 1, \cdots). \tag{26}$$

This shows that *the gamma function can be regarded as a generalization of the elementary factorial function.* [Sometimes the notation $(\alpha - 1)!$ is used for $\Gamma(\alpha)$, even for noninteger values of α, and the gamma function is also known as the **factorial function**.]

By repeated application of (25) we obtain

$$\Gamma(\alpha) = \frac{\Gamma(\alpha+1)}{\alpha} = \frac{\Gamma(\alpha+2)}{\alpha(\alpha+1)} = \cdots = \frac{\Gamma(\alpha+k+1)}{\alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k)}$$

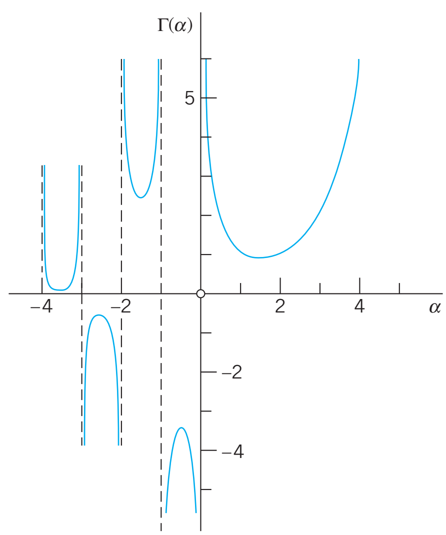


Fig. 553. Gamma function

and we may use this relation

$$\Gamma(\alpha) = \frac{\Gamma(\alpha + k + 1)}{\alpha(\alpha + 1) \cdots (\alpha + k)} \qquad (\alpha \neq 0, -1, -2, \cdots), \tag{27}$$

for defining the gamma function for negative α ($\neq -1, -2, \cdots$), choosing for k the smallest integer such that $\alpha + k + 1 > 0$. *Together with* (24), *this then gives a definition of* $\Gamma(\alpha)$ *for all* α *not equal to zero or a negative integer* (Fig. 553).

It can be shown that the gamma function may also be represented as the limit of a product, namely, by the formula

$$\Gamma(\alpha) = \lim_{n\to\infty} \frac{n!\, n^{\alpha}}{\alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + n)} \qquad (\alpha \neq 0, -1, \cdots). \tag{28}$$

From (27) or (28) we see that, for complex α, the gamma function $\Gamma(\alpha)$ is a meromorphic function with simple poles at $\alpha = 0, -1, -2, \cdots$.

An approximation of the gamma function for large positive α is given by the **Stirling formula**

$$\Gamma(\alpha + 1) \approx \sqrt{2\pi\alpha}\left(\frac{\alpha}{e}\right)^{\alpha} \tag{29}$$

where e is the base of the natural logarithm. We finally mention the special value

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}. \tag{30}$$

Incomplete gamma functions

$$P(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1}\, dt, \qquad Q(\alpha, x) = \int_x^{\infty} e^{-t} t^{\alpha-1}\, dt \qquad (\alpha > 0) \tag{31}$$

$$\Gamma(\alpha) = P(\alpha, x) + Q(\alpha, x) \tag{32}$$

Beta function

$$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt \qquad (x > 0, y > 0) \tag{33}$$

Representation in terms of gamma functions:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \tag{34}$$

Error function (Fig. 554 and Table A4 in App. 5)

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{35}$$

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}}\left(x - \frac{x^3}{1!3} + \frac{x^5}{2!5} - \frac{x^7}{3!7} + - \cdots\right) \tag{36}$$

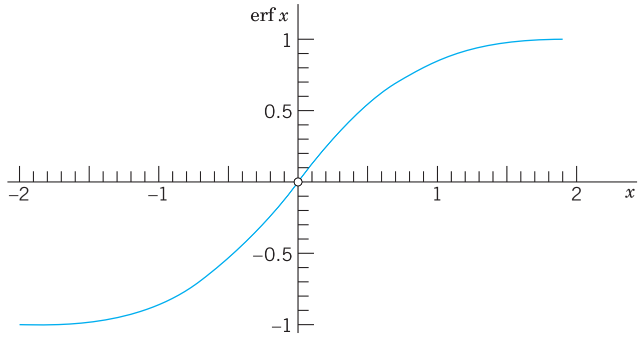


Fig. 554. Error function

erf $(\infty) = 1$, *complementary error function*

$$\operatorname{erfc} x = 1 - \operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2}\, dt \tag{37}$$

Fresnel integrals[1] (Fig. 555)

$$\mathrm{C}(x) = \int_0^x \cos(t^2)\, dt, \qquad \mathrm{S}(x) = \int_0^x \sin(t^2)\, dt \tag{38}$$

$\mathrm{C}(\infty) = \sqrt{\pi/8}$, $\mathrm{S}(\infty) = \sqrt{\pi/8}$, *complementary functions*

$$\begin{aligned} \mathrm{c}(x) &= \sqrt{\frac{\pi}{8}} - \mathrm{C}(x) = \int_x^{\infty} \cos(t^2)\, dt \\ \mathrm{s}(x) &= \sqrt{\frac{\pi}{8}} - \mathrm{S}(x) = \int_x^{\infty} \sin(t^2)\, dt \end{aligned} \tag{39}$$

Sine integral (Fig. 556 and Table A4 in App. 5)

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt \tag{40}$$

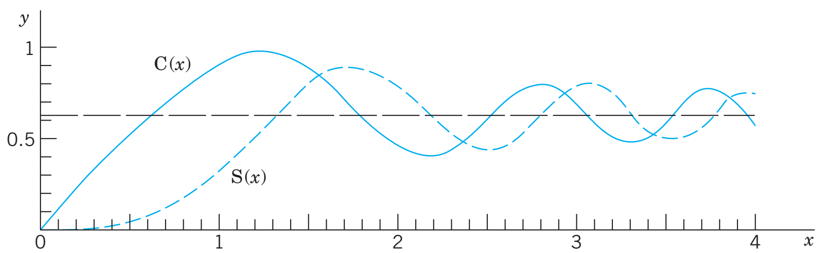


Fig. 555. Fresnel integrals

[1]AUGUSTIN FRESNEL (1788–1827), French physicist and mathematician. For tables see Ref. [GenRef1].

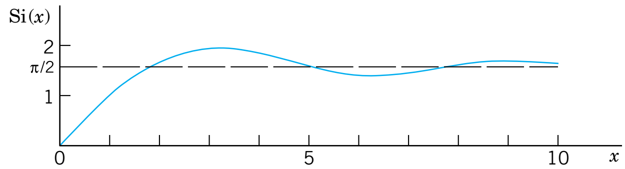


Fig. 556. Sine integral

$\mathrm{Si}(\infty) = \pi/2$, *complementary function*

$$\mathrm{si}(x) = \frac{\pi}{2} - \mathrm{Si}(x) = \int_x^\infty \frac{\sin t}{t}\, dt \tag{41}$$

Cosine integral (Table A4 in App. 5)

$$\mathrm{ci}(x) = \int_x^\infty \frac{\cos t}{t}\, dt \qquad (x > 0) \tag{42}$$

Exponential integral

$$\mathrm{Ei}(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt \qquad (x > 0) \tag{43}$$

Logarithmic integral

$$\mathrm{li}(x) = \int_0^x \frac{dt}{\ln t} \tag{44}$$

A3.2 Partial Derivatives

For differentiation formulas, see inside of front cover.

Let $z = f(x, y)$ be a real function of two independent real variables, x and y. If we keep y constant, say, $y = y_1$, and think of x as a variable, then $f(x, y_1)$ depends on x alone. If the derivative of $f(x, y_1)$ with respect to x for a value $x = x_1$ exists, then the value of this derivative is called the **partial derivative** of $f(x, y)$ *with respect to x at the point* (x_1, y_1) and is denoted by

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1, y_1)} \qquad \text{or by} \qquad \left.\frac{\partial z}{\partial x}\right|_{(x_1, y_1)}.$$

Other notations are

$$f_x(x_1, y_1) \qquad \text{and} \qquad z_x(x_1, y_1);$$

these may be used when subscripts are not used for another purpose and there is no danger of confusion.

We thus have, by the definition of the derivative,

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1,y_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1) - f(x_1, y_1)}{\Delta x}. \tag{1}$$

The partial derivative of $z = f(x, y)$ with respect to y is defined similarly; we now keep x constant, say, equal to x_1, and differentiate $f(x_1, y)$ with respect to y. Thus

$$\left.\frac{\partial f}{\partial y}\right|_{(x_1,y_1)} = \left.\frac{\partial z}{\partial y}\right|_{(x_1,y_1)} = \lim_{\Delta y \to 0} \frac{f(x_1, y_1 + \Delta y) - f(x_1, y_1)}{\Delta y}. \tag{2}$$

Other notations are $f_y(x_1, y_1)$ and $z_y(x_1, y_1)$.

It is clear that the values of those two partial derivatives will in general depend on the point (x_1, y_1). Hence the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ at a variable point (x, y) are functions of x and y. The function $\partial z/\partial x$ is obtained as in ordinary calculus by differentiating $z = f(x, y)$ with respect to x, ***treating y as a constant***, and $\partial z/\partial y$ is obtained by differentiating z with respect to y, ***treating x as a constant***.

EXAMPLE 1 Let $z = f(x, y) = x^2 y + x \sin y$. Then

$$\frac{\partial f}{\partial x} = 2xy + \sin y, \qquad \frac{\partial f}{\partial y} = x^2 + x \cos y. \qquad \blacksquare$$

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ of a function $z = f(x, y)$ have a very simple **geometric interpretation**. The function $z = f(x, y)$ can be represented by a surface in space. The equation $y = y_1$ then represents a vertical plane intersecting the surface in a curve, and the partial derivative $\partial z/\partial x$ at a point (x_1, y_1) is the slope of the tangent (that is, $\tan \alpha$ where α is the angle shown in Fig. 557) to the curve. Similarly, the partial derivative $\partial z/\partial y$ at (x_1, y_1) is the slope of the tangent to the curve $x = x_1$ on the surface $z = f(x, y)$ at (x_1, y_1).

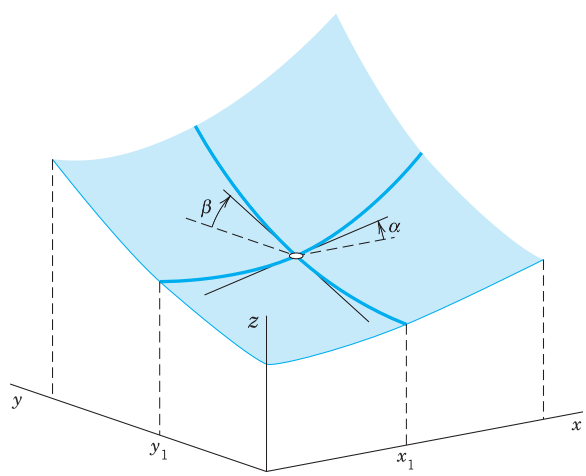


Fig. 557. Geometrical interpretation of first partial derivatives

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ are called *first partial derivatives* or *partial derivatives of first order.* By differentiating these derivatives once more, we obtain the four *second partial derivatives* (or *partial derivatives of second order*)[2]

$$
(3) \qquad
\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = f_{xx} \\
\frac{\partial^2 f}{\partial x\,\partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = f_{yx} \\
\frac{\partial^2 f}{\partial y\,\partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = f_{xy} \\
\frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = f_{yy}.
\end{aligned}
$$

It can be shown that if all the derivatives concerned are continuous, then the two mixed partial derivatives are equal, so that the order of differentiation does not matter (see Ref. [GenRef4] in App. 1), that is,

$$
(4) \qquad \frac{\partial^2 z}{\partial x\,\partial y} = \frac{\partial^2 z}{\partial y\,\partial x}\,.
$$

EXAMPLE 2 For the function in Example 1.

$$
f_{xx} = 2y, \qquad f_{xy} = 2x + \cos y = f_{yx}, \qquad f_{yy} = -x \sin y. \qquad \blacksquare
$$

By differentiating the second partial derivatives again with respect to x and y, respectively, we obtain the *third partial derivatives* or *partial derivatives of the third order* of f, etc.

If we consider a function $f(x, y, z)$ of ***three independent variables***, then we have the three first partial derivatives $f_x(x, y, z)$, $f_y(x, y, z)$, and $f_z(x, y, z)$. Here f_x is obtained by differentiating f with respect to x, ***treating both y and z as constants***. Thus, analogous to (1), we now have

$$
\left.\frac{\partial f}{\partial x}\right|_{(x_1, y_1, z_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1, z_1) - f(x_1, y_1, z_1)}{\Delta x}\,,
$$

etc. By differentiating f_x, f_y, f_z again in this fashion we obtain the second partial derivatives of f, etc.

EXAMPLE 3 Let $f(x, y, z) = x^2 + y^2 + z^2 + xy\,e^z$. Then

$$
\begin{aligned}
&f_x = 2x + y\,e^z, && f_y = 2y + x\,e^z, && f_z = 2z + xy\,e^z, \\
&f_{xx} = 2, && f_{xy} = f_{yx} = e^z, && f_{xz} = f_{zx} = y\,e^z, \\
&f_{yy} = 2, && f_{yz} = f_{zy} = x\,e^z, && f_{zz} = 2 + xy\,e^z.
\end{aligned} \qquad \blacksquare
$$

[2] **CAUTION!** In the subscript notation, the subscripts are written in the order in which we differentiate, whereas in the "∂" notation the order is opposite.

A3.3 Sequences and Series

See also Chap. 15.

Monotone Real Sequences

We call a real sequence $x_1, x_2, \cdots, x_n, \cdots$ a **monotone sequence** if it is either **monotone increasing**, that is,

$$x_1 \leqq x_2 \leqq x_3 \leqq \cdots$$

or **monotone decreasing**, that is,

$$x_1 \geqq x_2 \geqq x_3 \geqq \cdots.$$

We call $x_1, x_2, \cdots$ a **bounded sequence** if there is a positive constant K such that $|x_n| < K$ for all n.

THEOREM 1 *If a real sequence is bounded and monotone, it converges.*

PROOF Let $x_1, x_2, \cdots$ be a bounded monotone increasing sequence. Then its terms are smaller than some number B and, since $x_1 \leqq x_n$ for all n, they lie in the interval $x_1 \leqq x_n \leqq B$, which will be denoted by I_0. We bisect I_0; that is, we subdivide it into two parts of equal length. If the right half (together with its endpoints) contains terms of the sequence, we denote it by I_1. If it does not contain terms of the sequence, then the left half of I_0 (together with its endpoints) is called I_1. This is the first step.

In the second step we bisect I_1, select one half by the same rule, and call it I_2, and so on (see Fig. 558).

In this way we obtain shorter and shorter intervals $I_0, I_1, I_2, \cdots$ with the following properties. Each I_m contains all I_n for $n > m$. No term of the sequence lies to the right of I_m, and, since the sequence is monotone increasing, all x_n with n greater than some number N lie in I_m; of course, N will depend on m, in general. The lengths of the I_m approach zero as m approaches infinity. Hence there is precisely one number, call it L, that lies in all those intervals,[3] and we may now easily prove that the sequence is convergent with the limit L.

In fact, given an $\epsilon > 0$, we choose an m such that the length of I_m is less than ϵ. Then L and all the x_n with $n > N(m)$ lie in I_m, and, therefore, $|x_n - L| < \epsilon$ for all those n. This completes the proof for an increasing sequence. For a decreasing sequence the proof is the same, except for a suitable interchange of "left" and "right" in the construction of those intervals. ■

---

[3]This statement seems to be obvious, but actually it is not; it may be regarded as an axiom of the real number system in the following form. Let $J_1, J_2, \cdots$ be closed intervals such that each J_m contains all J_n with $n > m$, and the lengths of the J_m approach zero as m approaches infinity. Then there is precisely one real number that is contained in all those intervals. This is the so-called **Cantor–Dedekind axiom**, named after the German mathematicians GEORG CANTOR (1845–1918), the creator of set theory, and RICHARD DEDEKIND (1831–1916), known for his fundamental work in number theory. For further details see Ref. [GenRef2] in App. 1. (An interval I is said to be **closed** if its two endpoints are regarded as points belonging to I. It is said to be **open** if the endpoints are not regarded as points of I.)

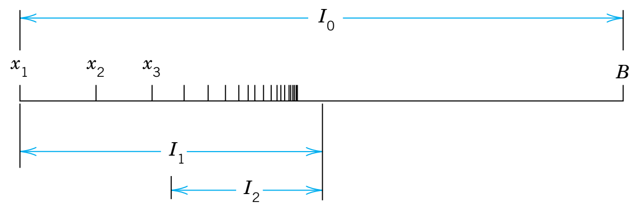


Fig. 558. Proof of Theorem 1

Real Series

THEOREM 2

Leibniz Test for Real Series

Let $x_1, x_2, \cdots$ be real and monotone decreasing to zero, that is,

$$\text{(1)} \qquad \text{(a)} \quad x_1 \geqq x_2 \geqq x_3 \geqq \cdots, \qquad \text{(b)} \quad \lim_{m \to \infty} x_m = 0.$$

Then the series with terms of alternating signs

$$x_1 - x_2 + x_3 - x_4 + - \cdots$$

converges, and for the remainder R_n after the nth term we have the estimate

$$\text{(2)} \qquad |R_n| \leqq x_{n+1}.$$

PROOF Let s_n be the nth partial sum of the series. Then, because of (1a),

$$s_1 = x_1, \qquad s_2 = x_1 - x_2 \leqq s_1,$$

$$s_3 = s_2 + x_3 \geqq s_2, \qquad s_3 = s_1 - (x_2 - x_3) \leqq s_1,$$

so that $s_2 \leqq s_3 \leqq s_1$. Proceeding in this fashion, we conclude that (Fig. 559)

$$\text{(3)} \qquad s_1 \geqq s_3 \geqq s_5 \geqq \cdots \geqq s_6 \geqq s_4 \geqq s_2$$

which shows that the odd partial sums form a bounded monotone sequence, and so do the even partial sums. Hence, by Theorem 1, both sequences converge, say,

$$\lim_{n \to \infty} s_{2n+1} = s, \qquad \lim_{n \to \infty} s_{2n} = s^*.$$

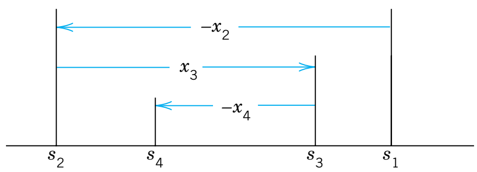


Fig. 559. Proof of the Leibniz test

Now, since $s_{2n+1} - s_{2n} = x_{2n+1}$, we readily see that (lb) implies

$$s - s^* = \lim_{n\to\infty} s_{2n+1} - \lim_{n\to\infty} s_{2n} = \lim_{n\to\infty} (s_{2n+1} - s_{2n}) = \lim_{n\to\infty} x_{2n+1} = 0.$$

Hence $s^* = s$, and the series converges with the sum s.

We prove the estimate (2) for the remainder. Since $s_n \to s$, it follows from (3) that

$$s_{2n+1} \geqq s \geqq s_{2n} \qquad \text{and also} \qquad s_{2n-1} \geqq s \geqq s_{2n}.$$

By subtracting s_{2n} and s_{2n-1}, respectively, we obtain

$$s_{2n+1} - s_{2n} \geqq s - s_{2n} \geqq 0, \qquad 0 \geqq s - s_{2n-1} \geqq s_{2n} - s_{2n-1}.$$

In these inequalities, the first expression is equal to x_{2n+1}, the last is equal to $-x_{2n}$, and the expressions between the inequality signs are the remainders R_{2n} and R_{2n-1}. Thus the inequalities may be written

$$x_{2n+1} \geqq R_{2n} \geqq 0, \qquad 0 \geqq R_{2n-1} \geqq -x_{2n}$$

and we see that they imply (2). This completes the proof. ■

A3.4 Grad, Div, Curl, ∇^2 in Curvilinear Coordinates

To simplify formulas, we write Cartesian coordinates $x = x_1$, $y = x_2$, $z = x_3$. We denote curvilinear coordinates by q_1, q_2, q_3. Through each point P there pass three coordinate surfaces $q_1 = \text{const}$, $q_2 = \text{const}$, $q_3 = \text{const}$. They intersect along coordinate curves. We assume the three coordinate curves through P to be **orthogonal** (perpendicular to each other). We write coordinate transformations as

$$(1) \qquad x_1 = x_1(q_1, q_2, q_3), \qquad x_2 = x_2(q_1, q_2, q_3), \qquad x_3 = x_3(q_1, q_2, q_3).$$

Corresponding transformations of grad, div, curl, and ∇^2 can all be written by using

$$(2) \qquad h_j^2 = \sum_{k=1}^{3} \left(\frac{\partial x_k}{\partial q_j}\right)^2.$$

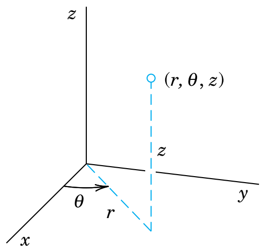
Next to Cartesian coordinates, most important are **cylindrical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = z$ (Fig. 560a) defined by

$$(3) \qquad x_1 = q_1 \cos q_2 = r\cos\theta, \qquad x_2 = q_1 \sin q_2 = r\sin\theta, \qquad x_3 = q_3 = z$$

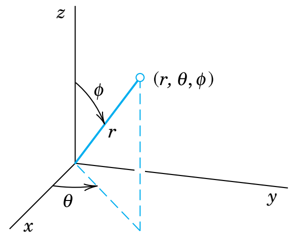
and **spherical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = \phi$ (Fig. 560b) defined by[4]

$$(4) \qquad \begin{gathered} x_1 = q_1 \cos q_2 \sin q_3 = r\cos\theta\sin\phi, \qquad x_2 = q_1 \sin q_2 \sin q_3 = r\sin\theta\sin\phi \\ x_3 = q_1 \cos q_3 = r\cos\phi. \end{gathered}$$

[4]This is the notation used in calculus and in many other books. It is logical since in it, θ plays the same role as in polar coordinates. **CAUTION!** Some books interchange the roles of θ and ϕ.



(*a*) Cylindrical coordinates



(*b*) Spherical coordinates

Fig. 560. Special curvilinear coordinates

In addition to the general formulas for any orthogonal coordinates q_1, q_2, q_3, we shall give additional formulas for these important special cases.

Linear Element *ds*. In Cartesian coordinates,

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 \qquad \text{(Sec. 9.5).}$$

For the q-coordinates,

$$ds^2 = h_1^2\, dq_1^2 + h_2^2\, dq_2^2 + h_3^2\, dq_3^2. \tag{5}$$

$$ds^2 = dr^2 + r^2\, d\theta^2 + dz^2 \qquad \text{(Cylindrical coordinates).} \tag{5$'$}$$

For polar coordinates set $dz^2 = 0$.

$$ds^2 = dr^2 + r^2 \sin^2 \phi\, d\theta^2 + r^2\, d\phi^2 \qquad \text{(Spherical coordinates).} \tag{5$''$}$$

Gradient. $\operatorname{grad} f = \nabla f = [f_{x_1}, \; f_{x_2}, \; f_{x_3}]$ (partial derivatives; Sec. 9.7). In the q-system, with **u, v, w** denoting unit vectors in the positive directions of the q_1, q_2, q_3 coordinate curves, respectively,

$$\operatorname{grad} f = \nabla f = \frac{1}{h_1}\frac{\partial f}{\partial q_1}\mathbf{u} + \frac{1}{h_2}\frac{\partial f}{\partial q_2}\mathbf{v} + \frac{1}{h_3}\frac{\partial f}{\partial q_3}\mathbf{w} \tag{6}$$

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{\partial f}{\partial z}\mathbf{w} \qquad \text{(Cylindrical coordinates)} \tag{6$'$}$$

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r \sin \phi}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{1}{r}\frac{\partial f}{\partial \phi}\mathbf{w} \qquad \text{(Spherical coordinates).} \tag{6$''$}$$

Divergence $\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = (F_1)_{x_1} + (F_2)_{x_2} + (F_3)_{x_3}$ ($\mathbf{F} = [F_1, F_2, F_3]$, Sec. 9.8);

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}(h_2 h_3 F_1) + \frac{\partial}{\partial q_2}(h_3 h_1 F_2) + \frac{\partial}{\partial q_3}(h_1 h_2 F_3)\right] \tag{7}$$

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{r}\frac{\partial}{\partial r}(rF_1) + \frac{1}{r}\frac{\partial F_2}{\partial \theta} + \frac{\partial F_3}{\partial z} \qquad \text{(Cylindrical coordinates)} \tag{7$'$}$$

$$(7'') \qquad \operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 F_1) + \frac{1}{r \sin \phi}\frac{\partial F_2}{\partial \theta} + \frac{1}{r \sin \phi}\frac{\partial}{\partial \phi}(\sin \phi \, F_3)$$

(Spherical coordinates).

Laplacian $\nabla^2 f = \nabla \cdot \nabla f = \operatorname{div}(\operatorname{grad} f) = f_{x_1 x_1} + f_{x_2 x_2} + f_{x_3 x_3}$ (Sec. 9.8):

$$(8) \qquad \nabla^2 f = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}\left(\frac{h_2 h_3}{h_1}\frac{\partial f}{\partial q_1}\right) + \frac{\partial}{\partial q_2}\left(\frac{h_3 h_1}{h_2}\frac{\partial f}{\partial q_2}\right) + \frac{\partial}{\partial q_3}\left(\frac{h_1 h_2}{h_3}\frac{\partial f}{\partial q_3}\right)\right]$$

$$(8') \qquad \nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2} \qquad \text{(Cylindrical coordinates)}$$

$$(8'') \qquad \nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{2}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2 \sin^2 \phi}\frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\cot \phi}{r^2}\frac{\partial f}{\partial \phi}$$

(Spherical coordinates).

Curl (Sec. 9.9):

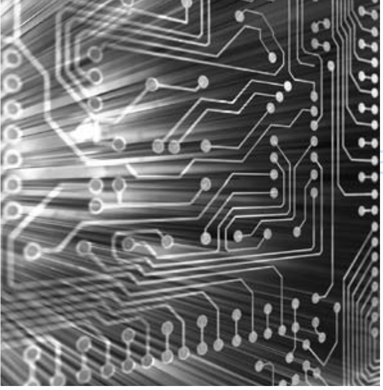
$$(9) \qquad \operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F} = \frac{1}{h_1 h_2 h_3}\begin{vmatrix} h_1 \mathbf{u} & h_2 \mathbf{v} & h_3 \mathbf{w} \\ \dfrac{\partial}{\partial q_1} & \dfrac{\partial}{\partial q_2} & \dfrac{\partial}{\partial q_3} \\ h_1 F_1 & h_2 F_2 & h_3 F_3 \end{vmatrix}.$$

For cylindrical coordinates we have in (9) (as in the previous formulas)

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 = r, \qquad h_3 = h_z = 1$$

and for spherical coordinates we have

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 \sin q_3 = r \sin \phi, \qquad h_3 = h_\phi = q_1 = r.$$



APPENDIX 4

Additional Proofs

Section 2.6, page 74

PROOF OF THEOREM 1 **Uniqueness**[1]

Assuming that the problem consisting of the ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

and the two initial conditions

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \tag{2}$$

has two solutions $y_1(x)$ and $y_2(x)$ on the interval I in the theorem, we show that their difference

$$y(x) = y_1(x) - y_2(x)$$

is identically zero on I; then $y_1 \equiv y_2$ on I, which implies uniqueness.

Since (1) is homogeneous and linear, y is a solution of that ODE on I, and since y_1 and y_2 satisfy the same initial conditions, y satisfies the conditions

$$y(x_0) = 0, \qquad y'(x_0) = 0. \tag{11}$$

We consider the function

$$z(x) = y(x)^2 + y'(x)^2$$

and its derivative

$$z' = 2yy' + 2y'y''.$$

From the ODE we have

$$y'' = -py' - qy.$$

By substituting this in the expression for z' we obtain

$$z' = 2yy' - 2py'^2 - 2qyy'. \tag{12}$$

Now, since y and y' are real,

$$(y \pm y')^2 = y^2 \pm 2yy' + y'^2 \geqq 0.$$

[1]This proof was suggested by my colleague, Prof. A. D. Ziebur. In this proof, we use some formula numbers that have not yet been used in Sec. 2.6.

From this and the definition of z we obtain the two inequalities

$$(13) \qquad \text{(a)} \quad 2yy' \leqq y^2 + y'^2 = z, \qquad \text{(b)} \quad -2yy' \leqq y^2 + y'^2 = z.$$

From (13b) we have $2yy' \geqq -z$. Together, $|2yy'| \leqq z$. For the last term in (12) we now obtain

$$-2qyy' \leqq |-2qyy'| = |q||2yy'| \leqq |q|z.$$

Using this result as well as $-p \leqq |p|$ and applying (13a) to the term $2yy'$ in (12), we find

$$z' \leqq z + 2|p|y'^2 + |q|z.$$

Since $y'^2 \leqq y^2 + y'^2 = z$, from this we obtain

$$z' \leqq (1 + 2|p| + |q|)z$$

or, denoting the function in parentheses by h,

$$(14a) \qquad z' \leqq hz \qquad \text{for all } x \text{ on } I.$$

Similarly, from (12) and (13) it follows that

$$(14b) \qquad \begin{aligned} -z' &= -2yy' + 2py'^2 + 2qyy' \\ &\leqq z + 2|p|z + |q|z = hz. \end{aligned}$$

The inequalities (14a) and (14b) are equivalent to the inequalities

$$(15) \qquad z' - hz \leqq 0, \qquad z' + hz \geqq 0.$$

Integrating factors for the two expressions on the left are

$$F_1 = e^{-\int h(x)\,dx} \qquad \text{and} \qquad F_2 = e^{\int h(x)\,dx}.$$

The integrals in the exponents exist because h is continuous. Since F_1 and F_2 are positive, we thus have from (15)

$$F_1(z' - hz) = (F_1 z)' \leqq 0 \qquad \text{and} \qquad F_2(z' + hz) = (F_2 z)' \geqq 0.$$

This means that $F_1 z$ is nonincreasing and $F_2 z$ is nondecreasing on I. Since $z(x_0) = 0$ by (11), when $x \leqq x_0$ we thus obtain

$$F_1 z \geqq (F_1 z)_{x_0} = 0, \qquad F_2 z \leqq (F_2 z)_{x_0} = 0$$

and similarly, when $x \geqq x_0$,

$$F_1 z \leqq 0, \qquad F_2 z \geqq 0.$$

Dividing by F_1 and F_2 and noting that these functions are positive, we altogether have

$$z \leqq 0, \qquad z \geqq 0 \qquad \text{for all } x \text{ on } I.$$

This implies that $z = y^2 + y'^2 \equiv 0$ on I. Hence $y \equiv 0$ or $y_1 \equiv y_2$ on I. ■

Section 5.3, page 182

PROOF OF THEOREM 2 Frobenius Method. Basis of Solutions. Three Cases

The formula numbers in this proof are the same as in the text of Sec. 5.3. An additional formula not appearing in Sec. 5.3 will be called (A) (see below).

The ODE in Theorem 2 is

$$y'' + \frac{b(x)}{x} y' + \frac{c(x)}{x^2} y = 0, \tag{1}$$

where $b(x)$ and $c(x)$ are analytic functions. We can write it

$$x^2 y'' + x b(x) y' + c(x) y = 0. \tag{1$'$}$$

The indicial equation of (1) is

$$r(r-1) + b_0 r + c_0 = 0. \tag{4}$$

The roots r_1, r_2 of this quadratic equation determine the general form of a basis of solutions of (1), and there are three possible cases as follows.

Case 1. Distinct Roots Not Differing by an Integer. A first solution of (1) is of the form

$$y_1(x) = x^{r_1}(a_0 + a_1 x + a_2 x^2 + \cdots) \tag{5}$$

and can be determined as in the power series method. For a proof that in this case, the ODE (1) has a second independent solution of the form

$$y_2(x) = x^{r_2}(A_0 + A_1 x + A_2 x^2 + \cdots), \tag{6}$$

see Ref. [A11] listed in App. 1.

Case 2. Double Root. The indicial equation (4) has a double root r if and only if $(b_0 - 1)^2 - 4c_0 = 0$, and then $r = \frac{1}{2}(1 - b_0)$. A first solution

$$y_1(x) = x^r(a_0 + a_1 x + a_2 x^2 + \cdots), \qquad r = \tfrac{1}{2}(1 - b_0), \tag{7}$$

can be determined as in Case 1. We show that a second independent solution is of the form

$$y_2(x) = y_1(x) \ln x + x^r(A_1 x + A_2 x^2 + \cdots) \qquad (x > 0). \tag{8}$$

We use the method of reduction of order (see Sec. 2.1), that is, we determine $u(x)$ such that $y_2(x) = u(x) y_1(x)$ is a solution of (1). By inserting this and the derivatives

$$y_2' = u' y_1 + u y_1', \qquad y_2'' = u'' y_1 + 2u' y_1' + u y_1''$$

into the ODE (1$'$) we obtain

$$x^2(u'' y_1 + 2u' y_1' + u y_1'') + x b(u' y_1 + u y_1') + c u y_1 = 0.$$

Since y_1 is a solution of (1′), the sum of the terms involving u is zero, and this equation reduces to

$$x^2 y_1 u'' + 2x^2 y_1' u' + xby_1 u' = 0.$$

By dividing by $x^2 y_1$ and inserting the power series for b we obtain

$$u'' + \left(2\frac{y_1'}{y_1} + \frac{b_0}{x} + \cdots\right) u' = 0.$$

Here, and in the following, the dots designate terms that are constant or involve positive powers of x. Now, from (7), it follows that

$$\begin{aligned}\frac{y_1'}{y_1} &= \frac{x^{r-1}[ra_0 + (r+1)a_1 x + \cdots]}{x^r[a_0 + a_1 x + \cdots]} \\ &= \frac{1}{x}\left(\frac{ra_0 + (r+1)a_1 x + \cdots}{a_0 + a_1 x + \cdots}\right) = \frac{r}{x} + \cdots.\end{aligned}$$

Hence the previous equation can be written

$$\text{(A)} \qquad u'' + \left(\frac{2r + b_0}{x} + \cdots\right) u' = 0.$$

Since $r = (1 - b_0)/2$, the term $(2r + b_0)/x$ equals $1/x$, and by dividing by u' we thus have

$$\frac{u''}{u'} = -\frac{1}{x} + \cdots.$$

By integration we obtain $\ln u' = -\ln x + \cdots$, hence $u' = (1/x)e^{(\cdots)}$. Expanding the exponential function in powers of x and integrating once more, we see that u is of the form

$$u = \ln x + k_1 x + k_2 x^2 + \cdots.$$

Inserting this into $y_2 = uy_1$, we obtain for y_2 a representation of the form (8).

Case 3. Roots Differing by an Integer. We write $r_1 = r$ and $r_2 = r - p$ where p is a ***positive*** integer. A first solution

$$\text{(9)} \qquad y_1(x) = x^{r_1}(a_0 + a_1 x + a_2 x^2 + \cdots)$$

can be determined as in Cases 1 and 2. We show that a second independent solution is of the form

$$\text{(10)} \qquad y_2(x) = ky_1(x)\ln x + x^{r_2}(A_0 + A_1 x + A_2 x^2 + \cdots)$$

where we may have $k \neq 0$ or $k = 0$. As in Case 2 we set $y_2 = uy_1$. The first steps are literally as in Case 2 and give Eq. (A),

$$u'' + \left(\frac{2r + b_0}{x} + \cdots\right) u' = 0.$$

Now by elementary algebra, the coefficient $b_0 - 1$ of r in (4) equals minus the sum of the roots,

$$b_0 - 1 = -(r_1 + r_2) = -(r + r - p) = -2r + p.$$

Hence $2r + b_0 = p + 1$, and division by u' gives

$$\frac{u''}{u'} = -\left(\frac{p+1}{x} + \cdots\right).$$

The further steps are as in Case 2. Integrating, we find

$$\ln u' = -(p+1)\ln x + \cdots, \qquad \text{thus} \qquad u' = x^{-(p+1)}e^{(\cdots)}$$

where dots stand for some series of nonnegative integer powers of x. By expanding the exponential function as before we obtain a series of the form

$$u' = \frac{1}{x^{p+1}} + \frac{k_1}{x^p} + \cdots + \frac{k_{p-1}}{x^2} + \frac{k_p}{x} + k_{p+1} + k_{p+2}x + \cdots.$$

We integrate once more. Writing the resulting logarithmic term first, we get

$$u = k_p \ln x + \left(-\frac{1}{px^p} - \cdots - \frac{k_{p-1}}{x} + k_{p+1}x + \cdots\right).$$

Hence, by (9) we get for $y_2 = uy_1$ the formula

$$y_2 = k_p y_1 \ln x + x^{r_1 - p}\left(-\frac{1}{p} - \cdots - k_{p-1}x^{p-1} + \cdots\right)(a_0 + a_1 x + \cdots).$$

But this is of the form (10) with $k = k_p$ since $r_1 - p = r_2$ and the product of the two series involves nonnegative integer powers of x only. ■

Section 7.7, page 293

THEOREM

Determinants

The definition of a determinant

$$(7) \qquad D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

as given in Sec. 7.7 is unambiguous, that is, it yields the same value of D no matter which rows or columns we choose in the development.

PROOF In this proof we shall use formula numbers not yet used in Sec. 7.7.

We shall prove first that *the same value is obtained no matter which* **row** *is chosen.*

The proof is by induction. The statement is true for a second-order determinant, for which the developments by the first row $a_{11}a_{22} + a_{12}(-a_{21})$ and by the second row $a_{21}(-a_{12}) + a_{22}a_{11}$ give the same value $a_{11}a_{22} - a_{12}a_{21}$. Assuming the statement to be true for an $(n-1)$st-order determinant, we prove that it is true for an nth-order determinant.

For this purpose we expand D in terms of each of two arbitrary rows, say, the ith and the jth, and compare the results. Without loss of generality let us assume $i < j$.

First Expansion. We expand D by the ith row. A typical term in this expansion is

$$a_{ik}C_{ik} = a_{ik} \cdot (-1)^{i+k}M_{ik}. \tag{19}$$

The minor M_{ik} of a_{ik} in D is an $(n-1)$st-order determinant. By the induction hypothesis we may expand it by any row. We expand it by the row corresponding to the jth row of D. This row contains the entries a_{jl} $(l \neq k)$. It is the $(j-1)$st row of M_{ik}, because M_{ik} does not contain entries of the ith row of D, and $i < j$. We have to distinguish between two cases as follows.

Case I. If $l < k$, then the entry a_{jl} belongs to the lth column of M_{ik} (see Fig. 561). Hence the term involving a_{jl} in this expansion is

$$a_{jl} \cdot (\text{cofactor of } a_{jl} \text{ in } M_{ik}) = a_{jl} \cdot (-1)^{(j-1)+l}M_{ikjl} \tag{20}$$

where M_{ikjl} is the minor of a_{jl} in M_{ik}. Since this minor is obtained from M_{ik} by deleting the row and column of a_{jl}, it is obtained from D by deleting the ith and jth rows and the kth and lth columns of D. We insert the expansions of the M_{ik} into that of D. Then it follows from (19) and (20) that the terms of the resulting representation of D are of the form

$$a_{ik}a_{jl} \cdot (-1)^{b}M_{ikjl} \qquad (l < k) \tag{21a}$$

where

$$b = i + k + j + l - 1.$$

Case II. If $l > k$, the only difference is that then a_{jl} belongs to the $(l-1)$st column of M_{ik}, because M_{ik} does not contain entries of the kth column of D, and $k < l$. This causes an additional minus sign in (20), and, instead of (21a), we therefore obtain

$$-a_{ik}a_{jl} \cdot (-1)^{b}M_{ikjl} \qquad (l > k) \tag{21b}$$

where b is the same as before.

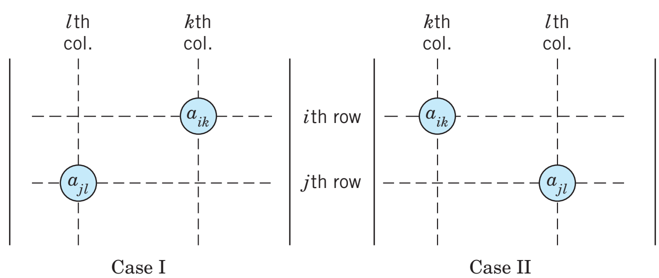


Fig. 561. Cases I and II of the two expansions of D

Second Expansion. We now expand D at first by the jth row. A typical term in this expansion is

$$a_{jl}C_{jl} = a_{jl} \cdot (-1)^{j+l}M_{jl}. \tag{22}$$

By the induction hypothesis we may expand the minor M_{jl} of a_{jl} in D by its ith row, which corresponds to the ith row of D, since $j > i$.

Case I. If $k > l$, the entry a_{ik} in that row belongs to the $(k-1)$st column of M_{jl}, because M_{jl} does not contain entries of the lth column of D, and $l < k$ (see Fig. 561). Hence the term involving a_{ik} in this expansion is

$$a_{ik} \cdot (\text{cofactor of } a_{ik} \text{ in } M_{jl}) = a_{ik} \cdot (-1)^{i+(k-1)}M_{ikjl}, \tag{23}$$

where the minor M_{ikjl} of a_{ik} in M_{jl} is obtained by deleting the ith and jth rows and the kth and lth columns of D [and is, therefore, identical with M_{ikjl} in (20), so that our notation is consistent]. We insert the expansions of the M_{jl} into that of D. It follows from (22) and (23) that this yields a representation whose terms are identical with those given by (21a) when $l < k$.

Case II. If $k < l$, then a_{ik} belongs to the kth column of M_{jl}, we obtain an additional minus sign, and the result agrees with that characterized by (21b).

We have shown that the two expansions of D consist of the same terms, and this proves our statement concerning rows.

The proof of the statement concerning ***columns*** is quite similar; if we expand D in terms of two arbitrary columns, say, the kth and the lth, we find that the general term involving $a_{jl}a_{ik}$ is exactly the same as before. This proves that not only all column expansions of D yield the same value, but also that their common value is equal to the common value of the row expansions of D.

This completes the proof and shows that *our definition of an nth-order determinant is unambiguous.* ■

Section 9.3, page 368

PROOF OF FORMULA (2)

We prove that in right-handed Cartesian coordinates, the vector product

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = [a_1, \quad a_2, \quad a_3] \times [b_1, \quad b_2, \quad b_3]$$

has the components

$$v_1 = a_2b_3 - a_3b_2, \qquad v_2 = a_3b_1 - a_1b_3, \qquad v_3 = a_1b_2 - a_2b_1. \tag{2}$$

We need only consider the case $\mathbf{v} \neq \mathbf{0}$. Since $\mathbf{v}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, Theorem 1 in Sec. 9.2 gives $\mathbf{a} \bullet \mathbf{v} = 0$ and $\mathbf{b} \bullet \mathbf{v} = 0$; in components [see (2), Sec. 9.2],

$$\begin{aligned} a_1v_1 + a_2v_2 + a_3v_3 &= 0 \\ b_1v_1 + b_2v_2 + b_3v_3 &= 0. \end{aligned} \tag{3}$$

Multiplying the first equation by b_3, the last by a_3, and subtracting, we obtain

$$(a_3b_1 - a_1b_3)v_1 = (a_2b_3 - a_3b_2)v_2.$$

Multiplying the first equation by b_1, the last by a_1, and subtracting, we obtain

$$(a_1b_2 - a_2b_1)v_2 = (a_3b_1 - a_1b_3)v_3.$$

We can easily verify that these two equations are satisfied by

$$(4) \qquad v_1 = c(a_2b_3 - a_3b_2), \qquad v_2 = c(a_3b_1 - a_1b_3), \qquad v_3 = c(a_1b_2 - a_2b_1)$$

where c is a constant. The reader may verify, by inserting, that (4) also satisfies (3). Now each of the equations in (3) represents a plane through the origin in $v_1v_2v_3$-space. The vectors **a** and **b** are normal vectors of these planes (see Example 6 in Sec. 9.2). Since $\mathbf{v} \neq \mathbf{0}$, these vectors are not parallel and the two planes do not coincide. Hence their intersection is a straight line L through the origin. Since (4) is a solution of (3) and, for varying c, represents a straight line, we conclude that (4) represents L, and every solution of (3) must be of the form (4). In particular, the components of **v** must be of this form, where c is to be determined. From (4) we obtain

$$|\mathbf{v}|^2 = v_1^2 + v_2^2 + v_3^2 = c^2\big[(a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2\big].$$

This can be written

$$|\mathbf{v}|^2 = c^2\big[(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1b_1 + a_2b_2 + a_3b_3)^2\big],$$

as can be verified by performing the indicated multiplications in both formulas and comparing. Using (2) in Sec. 9.2, we thus have

$$|\mathbf{v}|^2 = c^2\big[(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2\big].$$

By comparing this with formula (12) in Prob. 4 of Problem Set 9.3 we conclude that $c = \pm 1$.

We show that $c = +1$. This can be done as follows.

If we change the lengths and directions of **a** and **b** continuously and so that at the end $\mathbf{a} = \mathbf{i}$ and $\mathbf{b} = \mathbf{j}$ (Fig. 188a in Sec. 9.3), then **v** will change its length and direction continuously, and at the end, $\mathbf{v} = \mathbf{i} \times \mathbf{j} = \mathbf{k}$. Obviously we may effect the change so that both **a** and **b** remain different from the zero vector and are not parallel at any instant. Then **v** is never equal to the zero vector, and since the change is continuous and c can only assume the values $+1$ or -1, it follows that at the end c must have the same value as before. Now at the end $\mathbf{a} = \mathbf{i}$, $\mathbf{b} = \mathbf{j}$, $\mathbf{v} = \mathbf{k}$ and, therefore, $a_1 = 1$, $b_2 = 1$, $v_3 = 1$, and the other components in (4) are zero. Hence from (4) we see that $v_3 = c = +1$. This proves Theorem 1.

For a left-handed coordinate system, $\mathbf{i} \times \mathbf{j} = -\mathbf{k}$ (see Fig. 188b in Sec. 9.3), resulting in $c = -1$. This proves the statement right after formula (2). ■

Section 9.9, page 408

PROOF OF THE INVARIANCE OF THE CURL

This proof will follow from two theorems (A and B), which we prove first.

THEOREM A

Transformation Law for Vector Components

For any vector **v** *the components* v_1, v_2, v_3 *and* v_1^*, v_2^*, v_3^* *in any two systems of Cartesian coordinates* x_1, x_2, x_3 *and* x_1^*, x_2^*, x_3^*, *respectively, are related by*

$$(1) \qquad \begin{aligned} v_1^* &= c_{11}v_1 + c_{12}v_2 + c_{13}v_3 \\ v_2^* &= c_{21}v_1 + c_{22}v_2 + c_{23}v_3 \\ v_3^* &= c_{31}v_1 + c_{32}v_2 + c_{33}v_3, \end{aligned}$$

and conversely

$$(2) \qquad \begin{aligned} v_1 &= c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^* \\ v_2 &= c_{12}v_1^* + c_{22}v_2^* + c_{32}v_3^* \\ v_3 &= c_{13}v_1^* + c_{23}v_2^* + c_{33}v_3^* \end{aligned}$$

with coefficients

$$(3) \qquad \begin{array}{lll} c_{11} = \mathbf{i}^* \cdot \mathbf{i} & c_{12} = \mathbf{i}^* \cdot \mathbf{j} & c_{13} = \mathbf{i}^* \cdot \mathbf{k} \\ c_{21} = \mathbf{j}^* \cdot \mathbf{i} & c_{22} = \mathbf{j}^* \cdot \mathbf{j} & c_{23} = \mathbf{j}^* \cdot \mathbf{k} \\ c_{31} = \mathbf{k}^* \cdot \mathbf{i} & c_{32} = \mathbf{k}^* \cdot \mathbf{j} & c_{33} = \mathbf{k}^* \cdot \mathbf{k} \end{array}$$

satisfying

$$(4) \qquad \sum_{j=1}^{3} c_{kj}c_{mj} = \delta_{km} \qquad (k, m = 1, 2, 3),$$

where the **Kronecker delta**[2] *is given by*

$$\delta_{km} = \begin{cases} 0 & (k \neq m) \\ 1 & (k = m) \end{cases}$$

and **i, j, k** *and* **i*, j*, k*** *denote the unit vectors in the positive* x_1-, x_2-, x_3- *and* x_1^*-, x_2^*-, x_3^*-*directions, respectively.*

---

[2]LEOPOLD KRONECKER (1823–1891), German mathematician at Berlin, who made important contributions to algebra, group theory, and number theory.

We shall keep our discussion completely independent of Chap. 7, but readers familiar with matrices should recognize that we are dealing with **orthogonal transformations and matrices** and that our present theorem follows from Theorem 2 in Sec. 8.3.

PROOF The representation of $\mathbf{v}$ in the two systems are

$$(5) \qquad \text{(a)} \quad \mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k} \qquad \text{(b)} \quad \mathbf{v} = v_1^*\mathbf{i}^* + v_2^*\mathbf{j}^* + v_3^*\mathbf{k}^*.$$

Since $\mathbf{i}^* \cdot \mathbf{i}^* = 1$, $\mathbf{i}^* \cdot \mathbf{j}^* = 0$, $\mathbf{i}^* \cdot \mathbf{k}^* = 0$, we get from (5b) simply $\mathbf{i}^* \cdot \mathbf{v} = v_1^*$ and from this and (5a)

$$v_1^* = \mathbf{i}^* \cdot \mathbf{v} = \mathbf{i}^* \cdot v_1\mathbf{i} + \mathbf{i}^* \cdot v_2\mathbf{j} + \mathbf{i}^* \cdot v_3\mathbf{k} = v_1\mathbf{i}^* \cdot \mathbf{i} + v_2\mathbf{i}^* \cdot \mathbf{j} + v_3\mathbf{i}^* \cdot \mathbf{k}.$$

Because of (3), this is the first formula in (1), and the other two formulas are obtained similarly, by considering $\mathbf{j}^* \cdot \mathbf{v}$ and then $\mathbf{k}^* \cdot \mathbf{v}$. Formula (2) follows by the same idea, taking $\mathbf{i} \cdot \mathbf{v} = v_1$ from (5a) and then from (5b) and (3)

$$v_1 = \mathbf{i} \cdot \mathbf{v} = v_1^*\mathbf{i} \cdot \mathbf{i}^* + v_2^*\mathbf{i} \cdot \mathbf{j}^* + v_3^*\mathbf{i} \cdot \mathbf{k}^* = c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^*,$$

and similarly for the other two components.

We prove (4). We can write (1) and (2) briefly as

$$(6) \qquad \text{(a)} \quad v_j = \sum_{m=1}^{3} c_{mj}v_m^*, \qquad \text{(b)} \quad v_k^* = \sum_{j=1}^{3} c_{kj}v_j.$$

Substituting v_j into v_k^*, we get

$$v_k^* = \sum_{j=1}^{3} c_{kj} \sum_{m=1}^{3} c_{mj}v_m^* = \sum_{m=1}^{3} v_m^* \left(\sum_{j=1}^{3} c_{kj}c_{mj} \right),$$

where $k = 1, 2, 3$. Taking $k = 1$, we have

$$v_1^* = v_1^* \left(\sum_{j=1}^{3} c_{1j}c_{1j} \right) + v_2^* \left(\sum_{j=1}^{3} c_{1j}c_{2j} \right) + v_3^* \left(\sum_{j=1}^{3} c_{1j}c_{3j} \right).$$

For this to hold for *every* vector $\mathbf{v}$, the first sum must be 1 and the other two sums 0. This proves (4) with $k = 1$ for $m = 1, 2, 3$. Taking $k = 2$ and then $k = 3$, we obtain (4) with $k = 2$ and 3, for $m = 1, 2, 3$. ■

THEOREM B

Transformation Law for Cartesian Coordinates

The transformation of any Cartesian $x_1x_2x_3$-coordinate system into any other Cartesian $x_1^*x_2^*x_3^*$-coordinate system is of the form

$$(7) \qquad x_m^* = \sum_{j=1}^{3} c_{mj}x_j + b_m, \qquad m = 1, 2, 3,$$

with coefficients (3) *and constants b_1, b_2, b_3; conversely,*

$$(8) \qquad x_k = \sum_{n=1}^{3} c_{nk}x_n^* + \tilde{b}_k, \qquad k = 1, 2, 3.$$

Theorem B follows from Theorem A by noting that the most general transformation of a Cartesian coordinate system into another such system may be decomposed into a transformation of the type just considered and a translation; and under a translation, corresponding coordinates differ merely by a constant.

PROOF OF THE INVARIANCE OF THE CURL

We write again x_1, x_2, x_3 instead of x, y, z, and similarly x_1^*, x_2^*, x_3^* for other Cartesian coordinates, assuming that both systems are right-handed. Let a_1, a_2, a_3 denote the components of curl $\mathbf{v}$ in the $x_1x_2x_3$-coordinates, as given by (1), Sec. 9.9, with

$$x = x_1, \qquad y = x_2, \qquad z = x_3.$$

Similarly, let a_1^*, a_2^*, a_3^* denote the components of curl $\mathbf{v}$ in the $x_1^*x_2^*x_3^*$-coordinate system. We prove that the length and direction of curl $\mathbf{v}$ are independent of the particular choice of Cartesian coordinates, as asserted. We do this by showing that the components of curl $\mathbf{v}$ satisfy the transformation law (2), which is characteristic of vector components. We consider a_1. We use (6a), and then the chain rule for functions of several variables (Sec. 9.6). This gives

$$\begin{aligned} a_1 &= \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} = \sum_{m=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_3} \right) \\ &= \sum_{m=1}^{3} \sum_{j=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_3} \right). \end{aligned}$$

From this and (7) we obtain

$$\begin{aligned} a_1 &= \sum_{m=1}^{3} \sum_{j=1}^{3} (c_{m3}c_{j2} - c_{m2}c_{j3}) \frac{\partial v_m^*}{\partial x_j^*} \\ &= (c_{33}c_{22} - c_{32}c_{23}) \left(\frac{\partial v_3^*}{\partial x_2^*} - \frac{\partial v_2^*}{\partial x_3^*} \right) + \cdots \\ &= (c_{33}c_{22} - c_{32}c_{23})a_1^* + (c_{13}c_{32} - c_{12}c_{33})a_2^* + (c_{23}c_{12} - c_{22}c_{13})a_3^*. \end{aligned}$$

Note what we did. The double sum had $3 \times 3 = 9$ terms, 3 of which were zero (when $m = j$), and the remaining 6 terms we combined in pairs as we needed them in getting a_1^*, a_2^*, a_3^*.

We now use (3), Lagrange's identity (see Formula (15) in Team Project 24 in Problem Set 9.3) and $\mathbf{k}^* \times \mathbf{j}^* = -\mathbf{i}^*$ and $\mathbf{k} \times \mathbf{j} = -\mathbf{i}$. Then

$$\begin{aligned} c_{33}c_{22} - c_{32}c_{23} &= (\mathbf{k}^* \cdot \mathbf{k})(\mathbf{j}^* \cdot \mathbf{j}) - (\mathbf{k}^* \cdot \mathbf{j})(\mathbf{j}^* \cdot \mathbf{k}) \\ &= (\mathbf{k}^* \times \mathbf{j}^*) \cdot (\mathbf{k} \times \mathbf{j}) = \mathbf{i}^* \cdot \mathbf{i} = c_{11}, \qquad \text{etc.} \end{aligned}$$

Hence $a_1 = c_{11}a_1^* + c_{21}a_2^* + c_{31}a_3^*$. This is of the form of the first formula in (2) in Theorem A, and the other two formulas of the form (2) are obtained similarly. This proves the theorem for right-handed systems. If the $x_1x_2x_3$-coordinates are left-handed, then $\mathbf{k} \times \mathbf{j} = +\mathbf{i}$, but then there is a minus sign in front of the determinant in (1), Sec. 9.9. ■

Section 10.2, page 420

PROOF OF THEOREM 1, PART (b) We prove that if

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) \tag{1}$$

with continuous F_1, F_2, F_3 in a domain D is independent of path in D, then $F = \operatorname{grad} f$ in D for some f; in components

$$F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}. \tag{2'}$$

We choose any fixed A: (x_0, y_0, z_0) in D and any B: (x, y, z) in D and define f by

$$f(x, y, z) = f_0 + \int_A^B (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*) \tag{3}$$

with any constant f_0 and any path from A to B in D. Since A is fixed and we have independence of path, the integral depends only on the coordinates x, y, z, so that (3) defines a function $f(x, y, z)$ in D. We show that $\mathbf{F} = \operatorname{grad} f$ with this f, beginning with the first of the three relations (2′). Because of independence of path we may integrate from A to B_1: (x_1, y, z) and then parallel to the x-axis along the segment B_1B in Fig. 562 with B_1 chosen so that the whole segment lies in D. Then

$$f(x, y, z) = f_0 + \int_A^{B_1} (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*) + \int_{B_1}^{B} (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*).$$

We now take the partial derivative with respect to x on both sides. On the left we get $\partial f/\partial x$. We show that on the right we get F_1. The derivative of the first integral is zero because A: (x_0, y_0, z_0) and B_1: (x_1, y, z) do not depend on x. We consider the second integral. Since on the segment B_1B, both y and z are constant, the terms $F_2\,dy^*$ and

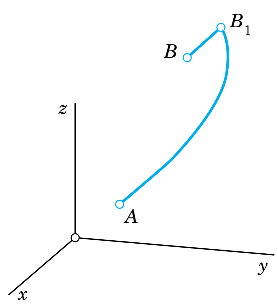


Fig. 562. Proof of Theorem 1

$F_3\, dz^*$ do not contribute to the derivative of the integral. The remaining part can be written as a definite integral,

$$\int_{B_1}^{B} F_1\, dx^* = \int_{x_1}^{x} F_1(x^*, y, z)\, dx^*.$$

Hence its partial derivative with respect to x is $F_1(x, y, z)$, and the first of the relations $(2')$ is proved. The other two formulas in $(2')$ follow by the same argument. ■

Section 11.5, page 500

THEOREM **Reality of Eigenvalues**

If p, q, r, and p' in the Sturm–Liouville equation (1) *of Sec.* 11.5 *are real-valued and continuous on the interval $a \leqq x \leqq b$ and $r(x) > 0$ throughout that interval (or $r(x) < 0$ throughout that interval), then all the eigenvalues of the Sturm–Liouville problem* (1), (2), *Sec.* 11.5, *are real.*

PROOF Let $\lambda = \alpha + i\beta$ be an eigenvalue of the problem and let

$$y(x) = u(x) + iv(x)$$

be a corresponding eigenfunction; here α, β, u, and v are real. Substituting this into (1), Sec. 11.5, we have

$$(pu' + ipv')' + (q + \alpha r + i\beta r)(u + iv) = 0.$$

This complex equation is equivalent to the following pair of equations for the real and the imaginary parts:

$$(pu')' + (q + \alpha r)u - \beta r v = 0$$

$$(pv')' + (q + \alpha r)v + \beta r u = 0.$$

Multiplying the first equation by v, the second by $-u$ and adding, we get

$$\begin{aligned} -\beta(u^2 + v^2)r &= u(pv')' - v(pu')' \\ &= [(pv')u - (pu')v]'. \end{aligned}$$

The expression in brackets is continuous on $a \leqq x \leqq b$, for reasons similar to those in the proof of Theorem 1, Sec. 11.5. Integrating over x from a to b, we thus obtain

$$-\beta \int_a^b (u^2 + v^2) r\, dx = \Big[p(uv' - u'v) \Big]_a^b .$$

Because of the boundary conditions, the right side is zero; this is as in that proof. Since y is an eigenfunction, $u^2 + v^2 \not\equiv 0$. Since y and r are continuous and $r > 0$ (or $r < 0$) on the interval $a \leqq x \leqq b$, the integral on the left is not zero. Hence, $\beta = 0$, which means that $\lambda = \alpha$ is real. This completes the proof. ■

Section 13.4, page 627

PROOF OF THEOREM 2 Cauchy–Riemann Equations

We prove that Cauchy–Riemann equations

$$(1) \qquad u_x = v_y, \qquad u_y = -v_x$$

are sufficient for a complex function $f(z) = u(x, y) + iv(x, y)$ to be analytic; precisely, *if the real part u and the imaginary part v of $f(z)$ satisfy* (1) *in a domain D in the complex plane and if the partial derivatives in* (1) *are* ***continuous*** *in D, then $f(z)$ is analytic in D.*

In this proof we write $\Delta z = \Delta x + i\Delta y$ and $\Delta f = f(z + \Delta z) - f(z)$. The idea of proof is as follows.

(a) We express Δf in terms of first partial derivatives of u and v, by applying the mean value theorem of Sec. 9.6.

(b) We get rid of partial derivatives with respect to y by applying the Cauchy–Riemann equations.

(c) We let Δz approach zero and show that then $\Delta f/\Delta z$, as obtained, approaches a limit, which is equal to $u_x + iv_x$, the right side of (4) in Sec. 13.4, regardless of the way of approach to zero.

(a) Let P: (x, y) be any fixed point in D. Since D is a domain, it contains a neighborhood of P. We can choose a point Q: $(x + \Delta x, y + \Delta y)$ in this neighborhood such that the straight-line segment PQ is in D. Because of our continuity assumptions we may apply the mean value theorem in Sec. 9.6. This yields

$$u(x + \Delta x, y + \Delta y) - u(x, y) = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1)$$
$$v(x + \Delta x, y + \Delta y) - v(x, y) = (\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)$$

where M_1 and M_2 ($\neq M_1$ in general!) are suitable points on that segment. The first line is Re Δf and the second is Im Δf, so that

$$\Delta f = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)].$$

(b) $u_y = -v_x$ and $v_y = u_x$ by the Cauchy–Riemann equations, so that

$$\Delta f = (\Delta x)u_x(M_1) - (\Delta y)v_x(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)].$$

Also $\Delta z = \Delta x + i\Delta y$, so that we can write $\Delta x = \Delta z - i\Delta y$ in the first term and $\Delta y = (\Delta z - \Delta x)/i = -i(\Delta z - \Delta x)$ in the second term. This gives

$$\Delta f = (\Delta z - i\Delta y)u_x(M_1) + i(\Delta z - \Delta x)v_x(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)].$$

By performing the multiplications and reordering we obtain

$$\Delta f = (\Delta z)u_x(M_1) - i\Delta y\{u_x(M_1) - u_x(M_2)\}$$
$$\qquad + i[(\Delta z)v_x(M_1) - \Delta x\{v_x(M_1) - v_x(M_2)\}].$$

Division by Δz now yields

$$\text{(A)} \quad \frac{\Delta f}{\Delta z} = u_x(M_1) + iv_x(M_1) - \frac{i\Delta y}{\Delta z}\{u_x(M_1) - u_x(M_2)\} - \frac{i\Delta x}{\Delta z}\{v_x(M_1) - v_x(M_2)\}.$$

(c) We finally let Δz approach zero and note that $|\Delta y/\Delta z| \leqq 1$ and $|\Delta x/\Delta z| \leqq 1$ in (A). Then Q: $(x + \Delta x, y + \Delta y)$ approaches P: (x, y), so that M_1 and M_2 must approach P. Also, since the partial derivatives in (A) are assumed to be continuous, they approach their value at P. In particular, the differences in the braces $\{\cdot\cdot\cdot\}$ in (A) approach zero. Hence the limit of the right side of (A) exists and is independent of the path along which $\Delta z \to 0$. We see that this limit equals the right side of (4) in Sec. 13.4. This means that $f(z)$ is analytic at every point z in D, and the proof is complete. ■

Section 14.2, pages 653–654

GOURSAT'S PROOF OF CAUCHY'S INTEGRAL THEOREM Goursat proved Cauchy's integral theorem without assuming that $f'(z)$ is continuous, as follows.

We start with the case when C is the boundary of a triangle. We orient C counterclockwise. By joining the midpoints of the sides we subdivide the triangle into four congruent triangles (Fig. 563). Let $C_{\text{I}}, C_{\text{II}}, C_{\text{III}}, C_{\text{IV}}$ denote their boundaries. We claim that (see Fig. 563).

$$\text{(1)} \qquad \oint_C f\,dz = \oint_{C_{\text{I}}} f\,dz + \oint_{C_{\text{II}}} f\,dz + \oint_{C_{\text{III}}} f\,dz + \oint_{C_{\text{IV}}} f\,dz.$$

Indeed, on the right we integrate along each of the three segments of subdivision in both possible directions (Fig. 563), so that the corresponding integrals cancel out in pairs, and the sum of the integrals on the right equals the integral on the left. We now pick an integral on the right that is biggest in absolute value and call its path C_1. Then, by the triangle inequality (Sec. 13.2),

$$\left|\oint_C f\,dz\right| \leqq \left|\oint_{C_{\text{I}}} f\,dz\right| + \left|\oint_{C_{\text{II}}} f\,dz\right| + \left|\oint_{C_{\text{III}}} f\,dz\right| + \left|\oint_{C_{\text{IV}}} f\,dz\right| \leqq 4\left|\oint_{C_1} f\,dz\right|.$$

We now subdivide the triangle bounded by C_1 as before and select a triangle of subdivision with boundary C_2 for which

$$\left|\oint_{C_1} f\,dz\right| \leqq 4\left|\oint_{C_2} f\,dz\right|. \qquad \text{Then} \qquad \left|\oint_C f\,dz\right| \leqq 4^2\left|\oint_{C_2} f\,dz\right|.$$

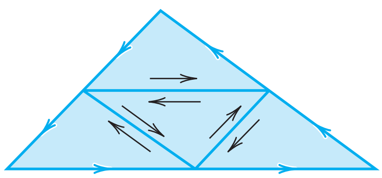


Fig. 563. Proof of Cauchy's integral theorem

Continuing in this fashion, we obtain a sequence of triangles $T_1, T_2, \cdots$ with boundaries $C_1, C_2, \cdots$ that are similar and such that T_n lies in T_m when $n > m$, and

$$\left|\oint_C f\,dz\right| \leqq 4^n \left|\oint_{C_n} f\,dz\right|, \qquad n = 1, 2, \cdots. \tag{2}$$

Let z_0 be the point that belongs to all these triangles. Since f is differentiable at $z = z_0$, the derivative $f'(z_0)$ exists. Let

$$h(z) = \frac{f(z) - f(z_0)}{z - z_0} - f'(z_0). \tag{3}$$

Solving this algebraically for $f(z)$ we have

$$f(z) = f(z_0) + (z - z_0)f'(z_0) + h(z)(z - z_0).$$

Integrating this over the boundary C_n of the triangle T_n gives

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} f(z_0)\,dz + \oint_{C_n} (z - z_0)f'(z_0)\,dz + \oint_{C_n} h(z)(z - z_0)dz.$$

Since $f(z_0)$ and $f'(z_0)$ are constants and C_n is a closed path, the first two integrals on the right are zero, as follows from Cauchy's proof, which is applicable because the integrands do have continuous derivatives (0 and const, respectively). We thus have

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} h(z)(z - z_0)\,dz.$$

Since $f'(z_0)$ is the limit of the difference quotient in (3), for given $\epsilon > 0$ we can find a $\delta > 0$ such that

$$|h(z)| < \epsilon \qquad \text{when} \qquad |z - z_0| < \delta. \tag{4}$$

We may now take n so large that the triangle T_n lies in the disk $|z - z_0| < \delta$. Let L_n be the length of C_n. Then $|z - z_0| < L_n$ for all z on C_n and z_0 in T_n. From this and (4) we have $|h(z)(z - z_0)| < \epsilon L_n$. The ML-inequality in Sec. 14.1 now gives

$$\left|\oint_{C_n} f(z)\,dz\right| = \left|\oint_{C_n} h(z)(z - z_0)\,dz\right| \leqq \epsilon L_n \cdot L_n = \epsilon L_n^2. \tag{5}$$

Now denote the length of C by L. Then the path C_1 has the length $L_1 = L/2$, the path C_2 has the length $L_2 = L_1/2 = L/4$, etc., and C_n has the length $L_n = L/2^n$. Hence $L_n^2 = L^2/4^n$. From (2) and (5) we thus obtain

$$\left|\oint_C f\,dz\right| \leqq 4^n \left|\oint_{C_n} f\,dz\right| \leqq 4^n \epsilon L_n^2 = 4^n \epsilon \frac{L^2}{4^n} = \epsilon L^2.$$

By choosing ϵ (> 0) sufficiently small we can make the expression on the right as small as we please, while the expression on the left is the definite value of an integral. Consequently, this value must be zero, and the proof is complete.

The proof for *the case in which C is the boundary of a polygon* follows from the previous proof by subdividing the polygon into triangles (Fig. 564). The integral corresponding to each such triangle is zero. The sum of these integrals is equal to the integral over C, because we integrate along each segment of subdivision in both directions, the corresponding integrals cancel out in pairs, and we are left with the integral over C.

The case of a general simple closed path C can be reduced to the preceding one by inscribing in C a closed polygon P of chords, which approximates C "sufficiently accurately," and it can be shown that there is a polygon P such that the integral over P differs from that over C by less than any preassigned positive real number $\tilde{\epsilon}$, no matter how small. The details of this proof are somewhat involved and can be found in Ref. [D6] listed in App. 1. ■

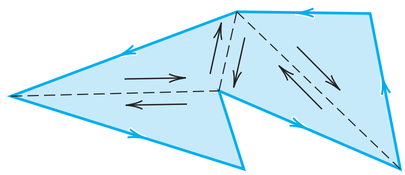


Fig. 564. Proof of Cauchy's integral theorem for a polygon

Section 15.1, page 674

PROOF OF THEOREM 4 Cauchy's Convergence Principle for Series

(a) In this proof we need two concepts and a theorem, which we list first.

1. A **bounded sequence** $s_1, s_2, \cdots$ is a sequence whose terms all lie in a disk of (sufficiently large, finite) radius K with center at the origin; thus $|s_n| < K$ for all n.

2. A **limit point** a of a sequence $s_1, s_2, \cdots$ is a point such that, given an $\epsilon > 0$, there are infinitely many terms satisfying $|s_n - a| < \epsilon$. (Note that this does *not* imply convergence, since there may still be infinitely many terms that do not lie within that circle of radius ϵ and center a.)

Example: $\frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{7}{8}, \frac{1}{16}, \frac{15}{16}, \cdots$ has the limit points 0 and 1 and diverges.

3. A bounded sequence in the complex plane has at least one limit point. (Bolzano–Weierstrass theorem; proof below. Recall that "sequence" always means *infinite* sequence.)

(b) We now turn to the actual proof that $z_1 + z_2 + \cdots$ converges if and only if, for every $\epsilon > 0$, we can find an N such that

$$|z_{n+1} + \cdots + z_{n+p}| < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots. \tag{1}$$

Here, by the definition of partial sums,

$$s_{n+p} - s_n = z_{n+1} + \cdots + z_{n+p}.$$

Writing $n + p = r$, we see from this that (1) is equivalent to

$$|s_r - s_n| < \epsilon \qquad \text{for all } r > N \text{ and } n > N. \tag{1*}$$

Suppose that $s_1, s_2, \cdots$ converges. Denote its limit by s. Then for a given $\epsilon > 0$ we can find an N such that

$$|s_n - s| < \frac{\epsilon}{2} \qquad \text{for every } n > N.$$

Hence, if $r > N$ and $n > N$, then by the triangle inequality (Sec. 13.2),

$$|s_r - s_n| = |(s_r - s) - (s_n - s)| \leqq |s_r - s| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

that is, (1*) holds.

(c) Conversely, assume that $s_1, s_2, \cdots$ satisfies (1*). We first prove that then the sequence must be bounded. Indeed, choose a fixed ϵ and a fixed $n = n_0 > N$ in (1*). Then (1*) implies that all s_r with $r > N$ lie in the disk of radius ϵ and center s_{n_0} and only *finitely many terms* $s_1, \cdots, s_N$ may not lie in this disk. Clearly, we can now find a circle so large that this disk and these finitely many terms all lie within this new circle. Hence the sequence is bounded. By the Bolzano–Weierstrass theorem, it has at least one limit point, call it s.

We now show that the sequence is convergent with the limit s. Let $\epsilon > 0$ be given. Then there is an N^* such that $|s_r - s_n| < \epsilon/2$ for all $r > N^*$ and $n > N^*$, by (1*). Also, by the definition of a limit point, $|s_n - s| < \epsilon/2$ for *infinitely many n*, so that we can find and fix an $n > N^*$ such that $|s_n - s| < \epsilon/2$. Together, for *every* $r > N^*$,

$$|s_r - s| = |(s_r - s_n) + (s_n - s)| \leqq |s_r - s_n| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon;$$

that is, the sequence $s_1, s_2, \cdots$ is convergent with the limit s. ■

THEOREM

Bolzano–Weierstrass Theorem[3]

A bounded infinite sequence $z_1, z_2, z_3, \cdots$ in the complex plane has at least one limit point.

PROOF It is obvious that we need both conditions: a finite sequence cannot have a limit point, and the sequence $1, 2, 3, \cdots$, which is infinite but not bounded, has no limit point. To prove the theorem, consider a bounded infinite sequence $z_1, z_2, \cdots$ and let K be such that $|z_n| < K$ for all n. If only finitely many values of the z_n are different, then, since the sequence is infinite, some number z must occur infinitely many times in the sequence, and, by definition, this number is a limit point of the sequence.

We may now turn to the case when the sequence contains infinitely many *different* terms. We draw a large square Q_0 that contains all z_n. We subdivide Q_0 into four congruent squares, which we number 1, 2, 3, 4. Clearly, at least one of these squares (each taken with its complete boundary) must contain infinitely many terms of the sequence. The square of this type with the lowest number (1, 2, 3, or 4) will be denoted by Q_1. This is

[3]BERNARD BOLZANO (1781–1848), Austrian mathematician and professor of religious studies, was a pioneer in the study of point sets, the foundation of analysis, and mathematical logic.

For Weierstrass, see Sec. 15.5.

the first step. In the next step we subdivide Q_1 into four congruent squares and select a square Q_2 by the same rule, and so on. This yields an infinite sequence of squares Q_0, $Q_1, Q_2, \cdots, Q_n, \cdots$ with the property that the side of Q_n approaches zero as n approaches infinity, and Q_m contains all Q_n with $n > m$. It is not difficult to see that the number which belongs to all these squares,[4] call it $z = a$, is a limit point of the sequence. In fact, given an $\epsilon > 0$, we can choose an N so large that the side of the square Q_N is less than ϵ and, since Q_N contains infinitely many z_n, we have $|z_n - a| < \epsilon$ for infinitely many n. This completes the proof. ■

Section 15.3, pages 688–689

PART (b) OF THE PROOF OF THEOREM 5

We have to show that

$$\sum_{n=2}^{\infty} a_n \left[\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1} \right]$$

$$= \sum_{n=2}^{\infty} a_n \, \Delta z\big[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n - 1)z^{n-2}\big],$$

thus,

$$\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1}$$

$$= \Delta z\big[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n - 1)z^{n-2}\big].$$

If we set $z + \Delta z = b$ and $z = a$, thus $\Delta z = b - a$, this becomes simply

$$\text{(7a)} \qquad \frac{b^n - a^n}{b - a} - na^{n-1} = (b - a)A_n \qquad (n = 2, 3, \cdots),$$

where A_n is the expression in the brackets on the right,

$$\text{(7b)} \qquad A_n = b^{n-2} + 2ab^{n-3} + 3a^2b^{n-4} + \cdots + (n - 1)a^{n-2};$$

thus, $A_2 = 1$, $A_3 = b + 2a$, etc. We prove (7) by induction. When $n = 2$, then (7) holds, since then

$$\frac{b^2 - a^2}{b - a} - 2a = \frac{(b + a)(b - a)}{b - a} - 2a = b - a = (b - a)A_2.$$

Assuming that (7) holds for $n = k$, we show that it holds for $n = k + 1$. By adding and subtracting a term in the numerator and then dividing we first obtain

$$\frac{b^{k+1} - a^{k+1}}{b - a} = \frac{b^{k+1} - ba^k + ba^k - a^{k+1}}{b - a} = b\,\frac{b^k - a^k}{b - a} + a^k.$$

---

[4] The fact that such a unique number $z = a$ exists seems to be obvious, but it actually follows from an axiom of the real number system, the so-called *Cantor–Dedekind axiom:* see footnote 3 in App. A3.3.

By the induction hypothesis, the right side equals $b\big[(b - a)A_k + ka^{k-1}\big] + a^k$. Direct calculation shows that this is equal to

$$(b - a)\{bA_k + ka^{k-1}\} + aka^{k-1} + a^k.$$

From (7b) with $n = k$ we see that the expression in the braces $\{\cdots\}$ equals

$$b^{k-1} + 2ab^{k-2} + \cdots + (k - 1)ba^{k-2} + ka^{k-1} = A_{k+1}.$$

Hence our result is

$$\frac{b^{k+1} - a^{k+1}}{b - a} = (b - a)A_{k+1} + (k + 1)a^k.$$

Taking the last term to the left, we obtain (7) with $n = k + 1$. This proves (7) for any integer $n \geqq 2$ and completes the proof. ■

Section 18.2, page 763

ANOTHER PROOF OF THEOREM 1 ***without the use of a harmonic conjugate***

We show that if $w = u + iv = f(z)$ is analytic and maps a domain D conformally onto a domain D^* and $\Phi^*(u, v)$ is harmonic in D^*, then

$$\Phi(x, y) = \Phi^*(u(x, y), v(x, y)) \tag{1}$$

is harmonic in D, that is, $\nabla^2\Phi = 0$ in D. We make no use of a harmonic conjugate of Φ^*, but use straightforward differentiation. By the chain rule,

$$\Phi_x = \Phi^*_u u_x + \Phi^*_v v_x.$$

We apply the chain rule again, underscoring the terms that will drop out when we form $\nabla^2\Phi$:

$$\Phi_{xx} = \underline{\Phi^*_u u_{xx}} + (\Phi^*_{uu}u_x + \underline{\Phi^*_{uv}v_x})u_x$$
$$+ \underline{\Phi^*_v v_{xx}} + (\underline{\Phi^*_{vu}u_x} + \Phi^*_{vv}v_x)v_x.$$

Φ_{yy} is the same with each x replaced by y. We form the sum $\nabla^2\Phi$. In it, $\Phi^*_{vu} = \Phi^*_{uv}$ is multiplied by

$$u_x v_x + u_y v_y$$

which is 0 by the Cauchy–Riemann equations. Also $\nabla^2 u = 0$ and $\nabla^2 v = 0$. There remains

$$\nabla^2\Phi = \Phi^*_{uu}(u_x^2 + u_y^2) + \Phi^*_{vv}(v_x^2 + v_y^2).$$

By the Cauchy–Riemann equations this becomes

$$\nabla^2\Phi = (\Phi^*_{uu} + \Phi^*_{vv})(u_x^2 + v_x^2)$$

and is 0 since Φ^* is harmonic. ■

APPENDIX 5

Tables

For Tables of Laplace Transforms see Secs. 6.8 and 6.9.
For Tables of Fourier Transforms see Sec. 11.10.

If you have a Computer Algebra System (CAS), you may not need the present tables, but you may still find them convenient from time to time.

Table A1 Bessel Functions

For more extensive tables see Ref. [GenRef1] in App. 1.

x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$
0.0	1.0000	0.0000	3.0	−0.2601	0.3391	6.0	0.1506	−0.2767
0.1	0.9975	0.0499	3.1	−0.2921	0.3009	6.1	0.1773	−0.2559
0.2	0.9900	0.0995	3.2	−0.3202	0.2613	6.2	0.2017	−0.2329
0.3	0.9776	0.1483	3.3	−0.3443	0.2207	6.3	0.2238	−0.2081
0.4	0.9604	0.1960	3.4	−0.3643	0.1792	6.4	0.2433	−0.1816
0.5	0.9385	0.2423	3.5	−0.3801	0.1374	6.5	0.2601	−0.1538
0.6	0.9120	0.2867	3.6	−0.3918	0.0955	6.6	0.2740	−0.1250
0.7	0.8812	0.3290	3.7	−0.3992	0.0538	6.7	0.2851	−0.0953
0.8	0.8463	0.3688	3.8	−0.4026	0.0128	6.8	0.2931	−0.0652
0.9	0.8075	0.4059	3.9	−0.4018	−0.0272	6.9	0.2981	−0.0349
1.0	0.7652	0.4401	4.0	−0.3971	−0.0660	7.0	0.3001	−0.0047
1.1	0.7196	0.4709	4.1	−0.3887	−0.1033	7.1	0.2991	0.0252
1.2	0.6711	0.4983	4.2	−0.3766	−0.1386	7.2	0.2951	0.0543
1.3	0.6201	0.5220	4.3	−0.3610	−0.1719	7.3	0.2882	0.0826
1.4	0.5669	0.5419	4.4	−0.3423	−0.2028	7.4	0.2786	0.1096
1.5	0.5118	0.5579	4.5	−0.3205	−0.2311	7.5	0.2663	0.1352
1.6	0.4554	0.5699	4.6	−0.2961	−0.2566	7.6	0.2516	0.1592
1.7	0.3980	0.5778	4.7	−0.2693	−0.2791	7.7	0.2346	0.1813
1.8	0.3400	0.5815	4.8	−0.2404	−0.2985	7.8	0.2154	0.2014
1.9	0.2818	0.5812	4.9	−0.2097	−0.3147	7.9	0.1944	0.2192
2.0	0.2239	0.5767	5.0	−0.1776	−0.3276	8.0	0.1717	0.2346
2.1	0.1666	0.5683	5.1	−0.1443	−0.3371	8.1	0.1475	0.2476
2.2	0.1104	0.5560	5.2	−0.1103	−0.3432	8.2	0.1222	0.2580
2.3	0.0555	0.5399	5.3	−0.0758	−0.3460	8.3	0.0960	0.2657
2.4	0.0025	0.5202	5.4	−0.0412	−0.3453	8.4	0.0692	0.2708
2.5	−0.0484	0.4971	5.5	−0.0068	−0.3414	8.5	0.0419	0.2731
2.6	−0.0968	0.4708	5.6	0.0270	−0.3343	8.6	0.0146	0.2728
2.7	−0.1424	0.4416	5.7	0.0599	−0.3241	8.7	−0.0125	0.2697
2.8	−0.1850	0.4097	5.8	0.0917	−0.3110	8.8	−0.0392	0.2641
2.9	−0.2243	0.3754	5.9	0.1220	−0.2951	8.9	−0.0653	0.2559

$J_0(x) = 0$ for $x = 2.40483, 5.52008, 8.65373, 11.7915, 14.9309, 18.0711, 21.2116, 24.3525, 27.4935, 30.6346$
$J_1(x) = 0$ for $x = 3.83171, 7.01559, 10.1735, 13.3237, 16.4706, 19.6159, 22.7601, 25.9037, 29.0468, 32.1897$

Table A1 *(continued)*

x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$
0.0	$(-\infty)$	$(-\infty)$	2.5	0.498	0.146	5.0	−0.309	0.148
0.5	−0.445	−1.471	3.0	0.377	0.325	5.5	−0.339	−0.024
1.0	0.088	−0.781	3.5	0.189	0.410	6.0	−0.288	−0.175
1.5	0.382	−0.412	4.0	−0.017	0.398	6.5	−0.173	−0.274
2.0	0.510	−0.107	4.5	−0.195	0.301	7.0	−0.026	−0.303

Table A2 **Gamma Function** [see (24) in App. A3.1]

α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$
1.00	1.000 000	1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384
1.02	0.988 844	1.22	0.913 106	1.42	0.886 356	1.62	0.895 924	1.82	0.936 845
1.04	0.978 438	1.24	0.908 521	1.44	0.885 805	1.64	0.898 642	1.84	0.942 612
1.06	0.968 744	1.26	0.904 397	1.46	0.885 604	1.66	0.901 668	1.86	0.948 687
1.08	0.959 725	1.28	0.900 718	1.48	0.885 747	1.68	0.905 001	1.88	0.955 071
1.10	0.951 351	1.30	0.897 471	1.50	0.886 227	1.70	0.908 639	1.90	0.961 766
1.12	0.943 590	1.32	0.894 640	1.52	0.887 039	1.72	0.912 581	1.92	0.968 774
1.14	0.936 416	1.34	0.892 216	1.54	0.888 178	1.74	0.916 826	1.94	0.976 099
1.16	0.929 803	1.36	0.890 185	1.56	0.889 639	1.76	0.921 375	1.96	0.983 743
1.18	0.923 728	1.38	0.888 537	1.58	0.891 420	1.78	0.926 227	1.98	0.991 708
1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384	2.00	1.000 000

Table A3 **Factorial Function and Its Logarithm with Base 10**

n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$
1	1	0.000 000	6	720	2.857 332	11	39 916 800	7.601 156
2	2	0.301 030	7	5 040	3.702 431	12	479 001 600	8.680 337
3	6	0.778 151	8	40 320	4.605 521	13	6 227 020 800	9.794 280
4	24	1.380 211	9	362 880	5.559 763	14	87 178 291 200	10.940 408
5	120	2.079 181	10	3 628 800	6.559 763	15	1 307 674 368 000	12.116 500

Table A4 **Error Function, Sine and Cosine Integrals** [see (35), (40), (42) in App. A3.1]

x	erf x	Si(x)	ci(x)	x	erf x	Si(x)	ci(x)
0.0	0.0000	0.0000	∞	2.0	0.9953	1.6054	−0.4230
0.2	0.2227	0.1996	1.0422	2.2	0.9981	1.6876	−0.3751
0.4	0.4284	0.3965	0.3788	2.4	0.9993	1.7525	−0.3173
0.6	0.6039	0.5881	0.0223	2.6	0.9998	1.8004	−0.2533
0.8	0.7421	0.7721	−0.1983	2.8	0.9999	1.8321	−0.1865
1.0	0.8427	0.9461	−0.3374	3.0	1.0000	1.8487	−0.1196
1.2	0.9103	1.1080	−0.4205	3.2	1.0000	1.8514	−0.0553
1.4	0.9523	1.2562	−0.4620	3.4	1.0000	1.8419	0.0045
1.6	0.9763	1.3892	−0.4717	3.6	1.0000	1.8219	0.0580
1.8	0.9891	1.5058	−0.4568	3.8	1.0000	1.7934	0.1038
2.0	0.9953	1.6054	−0.4230	4.0	1.0000	1.7582	0.1410

Table A5 Binomial Distribution

Probability function $f(x)$ [see (2), Sec. 24.7] and distribution function $F(x)$

		$p = 0.1$		$p = 0.2$		$p = 0.3$		$p = 0.4$		$p = 0.5$	
n	x	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
		0.		**0.**		**0.**		**0.**		**0.**	
1	0	9000	0.9000	8000	0.8000	7000	0.7000	6000	0.6000	5000	0.5000
	1	1000	1.0000	2000	1.0000	3000	1.0000	4000	1.0000	5000	1.0000
2	0	8100	0.8100	6400	0.6400	4900	0.4900	3600	0.3600	2500	0.2500
	1	1800	0.9900	3200	0.9600	4200	0.9100	4800	0.8400	5000	0.7500
	2	0100	1.0000	0400	1.0000	0900	1.0000	1600	1.0000	2500	1.0000
3	0	7290	0.7290	5120	0.5120	3430	0.3430	2160	0.2160	1250	0.1250
	1	2430	0.9720	3840	0.8960	4410	0.7840	4320	0.6480	3750	0.5000
	2	0270	0.9990	0960	0.9920	1890	0.9730	2880	0.9360	3750	0.8750
	3	0010	1.0000	0080	1.0000	0270	1.0000	0640	1.0000	1250	1.0000
4	0	6561	0.6561	4096	0.4096	2401	0.2401	1296	0.1296	0625	0.0625
	1	2916	0.9477	4096	0.8192	4116	0.6517	3456	0.4752	2500	0.3125
	2	0486	0.9963	1536	0.9728	2646	0.9163	3456	0.8208	3750	0.6875
	3	0036	0.9999	0256	0.9984	0756	0.9919	1536	0.9744	2500	0.9375
	4	0001	1.0000	0016	1.0000	0081	1.0000	0256	1.0000	0625	1.0000
5	0	5905	0.5905	3277	0.3277	1681	0.1681	0778	0.0778	0313	0.0313
	1	3281	0.9185	4096	0.7373	3602	0.5282	2592	0.3370	1563	0.1875
	2	0729	0.9914	2048	0.9421	3087	0.8369	3456	0.6826	3125	0.5000
	3	0081	0.9995	0512	0.9933	1323	0.9692	2304	0.9130	3125	0.8125
	4	0005	1.0000	0064	0.9997	0284	0.9976	0768	0.9898	1563	0.9688
	5	0000	1.0000	0003	1.0000	0024	1.0000	0102	1.0000	0313	1.0000
6	0	5314	0.5314	2621	0.2621	1176	0.1176	0467	0.0467	0156	0.0156
	1	3543	0.8857	3932	0.6554	3025	0.4202	1866	0.2333	0938	0.1094
	2	0984	0.9841	2458	0.9011	3241	0.7443	3110	0.5443	2344	0.3438
	3	0146	0.9987	0819	0.9830	1852	0.9295	2765	0.8208	3125	0.6563
	4	0012	0.9999	0154	0.9984	0595	0.9891	1382	0.9590	2344	0.8906
	5	0001	1.0000	0015	0.9999	0102	0.9993	0369	0.9959	0938	0.9844
	6	0000	1.0000	0001	1.0000	0007	1.0000	0041	1.0000	0156	1.0000
7	0	4783	0.4783	2097	0.2097	0824	0.0824	0280	0.0280	0078	0.0078
	1	3720	0.8503	3670	0.5767	2471	0.3294	1306	0.1586	0547	0.0625
	2	1240	0.9743	2753	0.8520	3177	0.6471	2613	0.4199	1641	0.2266
	3	0230	0.9973	1147	0.9667	2269	0.8740	2903	0.7102	2734	0.5000
	4	0026	0.9998	0287	0.9953	0972	0.9712	1935	0.9037	2734	0.7734
	5	0002	1.0000	0043	0.9996	0250	0.9962	0774	0.9812	1641	0.9375
	6	0000	1.0000	0004	1.0000	0036	0.9998	0172	0.9984	0547	0.9922
	7	0000	1.0000	0000	1.0000	0002	1.0000	0016	1.0000	0078	1.0000
8	0	4305	0.4305	1678	0.1678	0576	0.0576	0168	0.0168	0039	0.0039
	1	3826	0.8131	3355	0.5033	1977	0.2553	0896	0.1064	0313	0.0352
	2	1488	0.9619	2936	0.7969	2965	0.5518	2090	0.3154	1094	0.1445
	3	0331	0.9950	1468	0.9437	2541	0.8059	2787	0.5941	2188	0.3633
	4	0046	0.9996	0459	0.9896	1361	0.9420	2322	0.8263	2734	0.6367
	5	0004	1.0000	0092	0.9988	0467	0.9887	1239	0.9502	2188	0.8555
	6	0000	1.0000	0011	0.9999	0100	0.9987	0413	0.9915	1094	0.9648
	7	0000	1.0000	0001	1.0000	0012	0.9999	0079	0.9993	0313	0.9961
	8	0000	1.0000	0000	1.0000	0001	1.0000	0007	1.0000	0039	1.0000

Table A6 Poisson Distribution

Probability function $f(x)$ [see (5), Sec. 24.7] and distribution function $F(x)$

x	$\mu = 0.1$		$\mu = 0.2$		$\mu = 0.3$		$\mu = 0.4$		$\mu = 0.5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	9048	0.9048	8187	0.8187	7408	0.7408	6703	0.6703	6065	0.6065
1	0905	0.9953	1637	0.9825	2222	0.9631	2681	0.9384	3033	0.9098
2	0045	0.9998	0164	0.9989	0333	0.9964	0536	0.9921	0758	0.9856
3	0002	1.0000	0011	0.9999	0033	0.9997	0072	0.9992	0126	0.9982
4	0000	1.0000	0001	1.0000	0003	1.0000	0007	0.9999	0016	0.9998
5							0001	1.0000	0002	1.0000

x	$\mu = 0.6$		$\mu = 0.7$		$\mu = 0.8$		$\mu = 0.9$		$\mu = 1$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	5488	0.5488	4966	0.4966	4493	0.4493	4066	0.4066	3679	0.3679
1	3293	0.8781	3476	0.8442	3595	0.8088	3659	0.7725	3679	0.7358
2	0988	0.9769	1217	0.9659	1438	0.9526	1647	0.9371	1839	0.9197
3	0198	0.9966	0284	0.9942	0383	0.9909	0494	0.9865	0613	0.9810
4	0030	0.9996	0050	0.9992	0077	0.9986	0111	0.9977	0153	0.9963
5	0004	1.0000	0007	0.9999	0012	0.9998	0020	0.9997	0031	0.9994
6			0001	1.0000	0002	1.0000	0003	1.0000	0005	0.9999
7									0001	1.0000

x	$\mu = 1.5$		$\mu = 2$		$\mu = 3$		$\mu = 4$		$\mu = 5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	2231	0.2231	1353	0.1353	0498	0.0498	0183	0.0183	0067	0.0067
1	3347	0.5578	2707	0.4060	1494	0.1991	0733	0.0916	0337	0.0404
2	2510	0.8088	2707	0.6767	2240	0.4232	1465	0.2381	0842	0.1247
3	1255	0.9344	1804	0.8571	2240	0.6472	1954	0.4335	1404	0.2650
4	0471	0.9814	0902	0.9473	1680	0.8153	1954	0.6288	1755	0.4405
5	0141	0.9955	0361	0.9834	1008	0.9161	1563	0.7851	1755	0.6160
6	0035	0.9991	0120	0.9955	0504	0.9665	1042	0.8893	1462	0.7622
7	0008	0.9998	0034	0.9989	0216	0.9881	0595	0.9489	1044	0.8666
8	0001	1.0000	0009	0.9998	0081	0.9962	0298	0.9786	0653	0.9319
9			0002	1.0000	0027	0.9989	0132	0.9919	0363	0.9682
10					0008	0.9997	0053	0.9972	0181	0.9863
11					0002	0.9999	0019	0.9991	0082	0.9945
12					0001	1.0000	0006	0.9997	0034	0.9980
13							0002	0.9999	0013	0.9993
14							0001	1.0000	0005	0.9998
15									0002	0.9999
16									0000	1.0000

Table A7 Normal Distribution

Values of the distribution function $\Phi(z)$ [see (3), Sec. 24.8]. $\Phi(-z) = 1 - \Phi(z)$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
	0.		**0.**		**0.**		**0.**		**0.**		**0.**
0.01	5040	0.51	6950	1.01	8438	1.51	9345	2.01	9778	2.51	9940
0.02	5080	0.52	6985	1.02	8461	1.52	9357	2.02	9783	2.52	9941
0.03	5120	0.53	7019	1.03	8485	1.53	9370	2.03	9788	2.53	9943
0.04	5160	0.54	7054	1.04	8508	1.54	9382	2.04	9793	2.54	9945
0.05	5199	0.55	7088	1.05	8531	1.55	9394	2.05	9798	2.55	9946
0.06	5239	0.56	7123	1.06	8554	1.56	9406	2.06	9803	2.56	9948
0.07	5279	0.57	7157	1.07	8577	1.57	9418	2.07	9808	2.57	9949
0.08	5319	0.58	7190	1.08	8599	1.58	9429	2.08	9812	2.58	9951
0.09	5359	0.59	7224	1.09	8621	1.59	9441	2.09	9817	2.59	9952
0.10	5398	0.60	7257	1.10	8643	1.60	9452	2.10	9821	2.60	9953
0.11	5438	0.61	7291	1.11	8665	1.61	9463	2.11	9826	2.61	9955
0.12	5478	0.62	7324	1.12	8686	1.62	9474	2.12	9830	2.62	9956
0.13	5517	0.63	7357	1.13	8708	1.63	9484	2.13	9834	2.63	9957
0.14	5557	0.64	7389	1.14	8729	1.64	9495	2.14	9838	2.64	9959
0.15	5596	0.65	7422	1.15	8749	1.65	9505	2.15	9842	2.65	9960
0.16	5636	0.66	7454	1.16	8770	1.66	9515	2.16	9846	2.66	9961
0.17	5675	0.67	7486	1.17	8790	1.67	9525	2.17	9850	2.67	9962
0.18	5714	0.68	7517	1.18	8810	1.68	9535	2.18	9854	2.68	9963
0.19	5753	0.69	7549	1.19	8830	1.69	9545	2.19	9857	2.69	9964
0.20	5793	0.70	7580	1.20	8849	1.70	9554	2.20	9861	2.70	9965
0.21	5832	0.71	7611	1.21	8869	1.71	9564	2.21	9864	2.71	9966
0.22	5871	0.72	7642	1.22	8888	1.72	9573	2.22	9868	2.72	9967
0.23	5910	0.73	7673	1.23	8907	1.73	9582	2.23	9871	2.73	9968
0.24	5948	0.74	7704	1.24	8925	1.74	9591	2.24	9875	2.74	9969
0.25	5987	0.75	7734	1.25	8944	1.75	9599	2.25	9878	2.75	9970
0.26	6026	0.76	7764	1.26	8962	1.76	9608	2.26	9881	2.76	9971
0.27	6064	0.77	7794	1.27	8980	1.77	9616	2.27	9884	2.77	9972
0.28	6103	0.78	7823	1.28	8997	1.78	9625	2.28	9887	2.78	9973
0.29	6141	0.79	7852	1.29	9015	1.79	9633	2.29	9890	2.79	9974
0.30	6179	0.80	7881	1.30	9032	1.80	9641	2.30	9893	2.80	9974
0.31	6217	0.81	7910	1.31	9049	1.81	9649	2.31	9896	2.81	9975
0.32	6255	0.82	7939	1.32	9066	1.82	9656	2.32	9898	2.82	9976
0.33	6293	0.83	7967	1.33	9082	1.83	9664	2.33	9901	2.83	9977
0.34	6331	0.84	7995	1.34	9099	1.84	9671	2.34	9904	2.84	9977
0.35	6368	0.85	8023	1.35	9115	1.85	9678	2.35	9906	2.85	9978
0.36	6406	0.86	8051	1.36	9131	1.86	9686	2.36	9909	2.86	9979
0.37	6443	0.87	8078	1.37	9147	1.87	9693	2.37	9911	2.87	9979
0.38	6480	0.88	8106	1.38	9162	1.88	9699	2.38	9913	2.88	9980
0.39	6517	0.89	8133	1.39	9177	1.89	9706	2.39	9916	2.89	9981
0.40	6554	0.90	8159	1.40	9192	1.90	9713	2.40	9918	2.90	9981
0.41	6591	0.91	8186	1.41	9207	1.91	9719	2.41	9920	2.91	9982
0.42	6628	0.92	8212	1.42	9222	1.92	9726	2.42	9922	2.92	9982
0.43	6664	0.93	8238	1.43	9236	1.93	9732	2.43	9925	2.93	9983
0.44	6700	0.94	8264	1.44	9251	1.94	9738	2.44	9927	2.94	9984
0.45	6736	0.95	8289	1.45	9265	1.95	9744	2.45	9929	2.95	9984
0.46	6772	0.96	8315	1.46	9279	1.96	9750	2.46	9931	2.96	9985
0.47	6808	0.97	8340	1.47	9292	1.97	9756	2.47	9932	2.97	9985
0.48	6844	0.98	8365	1.48	9306	1.98	9761	2.48	9934	2.98	9986
0.49	6879	0.99	8389	1.49	9319	1.99	9767	2.49	9936	2.99	9986
0.50	6915	1.00	8413	1.50	9332	2.00	9772	2.50	9938	3.00	9987

Table A8 Normal Distribution

Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960
16	−0.994	0.202	56	0.151	0.772	96	1.751	2.054
17	−0.954	0.215	57	0.176	0.789	97	1.881	2.170
18	−0.915	0.228	58	0.202	0.806	97.5	1.960	2.241
19	−0.878	0.240	59	0.228	0.824	98	2.054	2.326
20	−0.842	0.253	60	0.253	0.842	99	2.326	2.576
21	−0.806	0.266	61	0.279	0.860	99.1	2.366	2.612
22	−0.772	0.279	62	0.305	0.878	99.2	2.409	2.652
23	−0.739	0.292	63	0.332	0.896	99.3	2.457	2.697
24	−0.706	0.305	64	0.358	0.915	99.4	2.512	2.748
25	−0.674	0.319	65	0.385	0.935	99.5	2.576	2.807
26	−0.643	0.332	66	0.412	0.954	99.6	2.652	2.878
27	−0.613	0.345	67	0.440	0.974	99.7	2.748	2.968
28	−0.583	0.358	68	0.468	0.994	99.8	2.878	3.090
29	−0.553	0.372	69	0.496	1.015	99.9	3.090	3.291
30	−0.524	0.385	70	0.524	1.036			
31	−0.496	0.399	71	0.553	1.058	99.91	3.121	3.320
32	−0.468	0.412	72	0.583	1.080	99.92	3.156	3.353
33	−0.440	0.426	73	0.613	1.103	99.93	3.195	3.390
34	−0.412	0.440	74	0.643	1.126	99.94	3.239	3.432
35	−0.385	0.454	75	0.674	1.150	99.95	3.291	3.481
36	−0.358	0.468	76	0.706	1.175	99.96	3.353	3.540
37	−0.332	0.482	77	0.739	1.200	99.97	3.432	3.615
38	−0.305	0.496	78	0.772	1.227	99.98	3.540	3.719
39	−0.279	0.510	79	0.806	1.254	99.99	3.719	3.891
40	−0.253	0.524	80	0.842	1.282			

Table A9 ***t*-Distribution**

Values of z for given values of the distribution function $F(z)$ (see (8) in Sec. 25.3). Example: For 9 degrees of freedom, $z = 1.83$ when $F(z) = 0.95$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.32	0.29	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26
0.7	0.73	0.62	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54
0.8	1.38	1.06	0.98	0.94	0.92	0.91	0.90	0.89	0.88	0.88
0.9	3.08	1.89	1.64	1.53	1.48	1.44	1.41	1.40	1.38	1.37
0.95	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
0.975	12.7	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23
0.99	31.8	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76
0.995	63.7	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17
0.999	318.3	22.3	10.2	7.17	5.89	5.21	4.79	4.50	4.30	4.14

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.7	0.54	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
0.8	0.88	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86
0.9	1.36	1.36	1.35	1.35	1.34	1.34	1.33	1.33	1.33	1.33
0.95	1.80	1.78	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.72
0.975	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09	2.09
0.99	2.72	2.68	2.65	2.62	2.60	2.58	2.57	2.55	2.54	2.53
0.995	3.11	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.85
0.999	4.02	3.93	3.85	3.79	3.73	3.69	3.65	3.61	3.58	3.55

$F(z)$	Number of Degrees of Freedom									
	22	24	26	28	30	40	50	100	200	∞
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.25
0.7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
0.8	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.84	0.84
0.9	1.32	1.32	1.31	1.31	1.31	1.30	1.30	1.29	1.29	1.28
0.95	1.72	1.71	1.71	1.70	1.70	1.68	1.68	1.66	1.65	1.65
0.975	2.07	2.06	2.06	2.05	2.04	2.02	2.01	1.98	1.97	1.96
0.99	2.51	2.49	2.48	2.47	2.46	2.42	2.40	2.36	2.35	2.33
0.995	2.82	2.80	2.78	2.76	2.75	2.70	2.68	2.63	2.60	2.58
0.999	3.50	3.47	3.43	3.41	3.39	3.31	3.26	3.17	3.13	3.09

Table A10 Chi-square Distribution

Values of x for given values of the distribution function $F(z)$ (see Sec. 25.3 before (17)). Example: For 3 degrees of freedom, $z = 11.34$ when $F(z) = 0.99$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.005	2.60	3.07	3.57	4.07	4.60	5.14	5.70	6.26	6.84	7.43
0.01	3.05	3.57	4.11	4.66	5.23	5.81	6.41	7.01	7.63	8.26
0.025	3.82	4.40	5.01	5.63	6.26	6.91	7.56	8.23	8.91	9.59
0.05	4.57	5.23	5.89	6.57	7.26	7.96	8.67	9.39	10.12	10.85
0.95	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41
0.975	21.92	23.34	24.74	26.12	27.49	28.85	30.19	31.53	32.85	34.17
0.99	24.72	26.22	27.69	29.14	30.58	32.00	33.41	34.81	36.19	37.57
0.995	26.76	28.30	29.82	31.32	32.80	34.27	35.72	37.16	38.58	40.00

$F(z)$	Number of Degrees of Freedom									
	21	22	23	24	25	26	27	28	29	30
0.005	8.0	8.6	9.3	9.9	10.5	11.2	11.8	12.5	13.1	13.8
0.01	8.9	9.5	10.2	10.9	11.5	12.2	12.9	13.6	14.3	15.0
0.025	10.3	11.0	11.7	12.4	13.1	13.8	14.6	15.3	16.0	16.8
0.05	11.6	12.3	13.1	13.8	14.6	15.4	16.2	16.9	17.7	18.5
0.95	32.7	33.9	35.2	36.4	37.7	38.9	40.1	41.3	42.6	43.8
0.975	35.5	36.8	38.1	39.4	40.6	41.9	43.2	44.5	45.7	47.0
0.99	38.9	40.3	41.6	43.0	44.3	45.6	47.0	48.3	49.6	50.9
0.995	41.4	42.8	44.2	45.6	46.9	48.3	49.6	51.0	52.3	53.7

$F(z)$	Number of Degrees of Freedom							
	40	50	60	70	80	90	100	> 100 (Approximation)
0.005	20.7	28.0	35.5	43.3	51.2	59.2	67.3	$\frac{1}{2}(h - 2.58)^2$
0.01	22.2	29.7	37.5	45.4	53.5	61.8	70.1	$\frac{1}{2}(h - 2.33)^2$
0.025	24.4	32.4	40.5	48.8	57.2	65.6	74.2	$\frac{1}{2}(h - 1.96)^2$
0.05	26.5	34.8	43.2	51.7	60.4	69.1	77.9	$\frac{1}{2}(h - 1.64)^2$
0.95	55.8	67.5	79.1	90.5	101.9	113.1	124.3	$\frac{1}{2}(h + 1.64)^2$
0.975	59.3	71.4	83.3	95.0	106.6	118.1	129.6	$\frac{1}{2}(h + 1.96)^2$
0.99	63.7	76.2	88.4	100.4	112.3	124.1	135.8	$\frac{1}{2}(h + 2.33)^2$
0.995	66.8	79.5	92.0	104.2	116.3	128.3	140.2	$\frac{1}{2}(h + 2.58)^2$

In the last column, $h = \sqrt{2m - 1}$, where m is the number of degrees of freedom.

Table A11 *F*-Distribution with (*m*, *n*) Degrees of Freedom

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

Example: For (7, 4) d.f., $z = 6.09$ if $F(z) = 0.95$.

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	242	246	248	250	251	252	253	254
2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
3	8.79	8.70	8.66	8.62	8.59	8.58	8.55	8.53
4	5.96	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	4.74	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	4.06	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	3.64	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	3.35	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	3.14	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	2.98	2.85	2.77	2.70	2.66	2.64	2.59	2.54
11	2.85	2.72	2.65	2.57	2.53	2.51	2.46	2.40
12	2.75	2.62	2.54	2.47	2.43	2.40	2.35	2.30
13	2.67	2.53	2.46	2.38	2.34	2.31	2.26	2.21
14	2.60	2.46	2.39	2.31	2.27	2.24	2.19	2.13
15	2.54	2.40	2.33	2.25	2.20	2.18	2.12	2.07
16	2.49	2.35	2.28	2.19	2.15	2.12	2.07	2.01
17	2.45	2.31	2.23	2.15	2.10	2.08	2.02	1.96
18	2.41	2.27	2.19	2.11	2.06	2.04	1.98	1.92
19	2.38	2.23	2.16	2.07	2.03	2.00	1.94	1.88
20	2.35	2.20	2.12	2.04	1.99	1.97	1.91	1.84
22	2.30	2.15	2.07	1.98	1.94	1.91	1.85	1.78
24	2.25	2.11	2.03	1.94	1.89	1.86	1.80	1.73
26	2.22	2.07	1.99	1.90	1.85	1.82	1.76	1.69
28	2.19	2.04	1.96	1.87	1.82	1.79	1.73	1.65
30	2.16	2.01	1.93	1.84	1.79	1.76	1.70	1.62
32	2.14	1.99	1.91	1.82	1.77	1.74	1.67	1.59
34	2.12	1.97	1.89	1.80	1.75	1.71	1.65	1.57
36	2.11	1.95	1.87	1.78	1.73	1.69	1.62	1.55
38	2.09	1.94	1.85	1.76	1.71	1.68	1.61	1.53
40	2.08	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	2.03	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	1.99	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	1.97	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	1.95	1.79	1.70	1.60	1.54	1.51	1.43	1.32
90	1.94	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	1.93	1.77	1.68	1.57	1.52	1.48	1.39	1.28
150	1.89	1.73	1.64	1.54	1.48	1.44	1.34	1.22
200	1.88	1.72	1.62	1.52	1.46	1.41	1.32	1.19
1000	1.84	1.68	1.58	1.47	1.41	1.36	1.26	1.08
∞	1.83	1.67	1.57	1.46	1.39	1.35	1.24	1.00

Table A11 ***F*-Distribution with (m, n) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	4052	4999	5403	5625	5764	5859	5928	5981	6022
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	6056	6157	6209	6261	6287	6303	6334	6366
2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	27.2	26.9	26.7	26.5	26.4	26.4	26.2	26.1
4	14.5	14.2	14.0	13.8	13.7	13.7	13.6	13.5
5	10.1	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	7.87	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	6.62	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	5.81	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	5.26	4.96	4.81	4.65	4.57	4.52	4.42	4.31
10	4.85	4.56	4.41	4.25	4.17	4.12	4.01	3.91
11	4.54	4.25	4.10	3.94	3.86	3.81	3.71	3.60
12	4.30	4.01	3.86	3.70	3.62	3.57	3.47	3.36
13	4.10	3.82	3.66	3.51	3.43	3.38	3.27	3.17
14	3.94	3.66	3.51	3.35	3.27	3.22	3.11	3.00
15	3.80	3.52	3.37	3.21	3.13	3.08	2.98	2.87
16	3.69	3.41	3.26	3.10	3.02	2.97	2.86	2.75
17	3.59	3.31	3.16	3.00	2.92	2.87	2.76	2.65
18	3.51	3.23	3.08	2.92	2.84	2.78	2.68	2.57
19	3.43	3.15	3.00	2.84	2.76	2.71	2.60	2.49
20	3.37	3.09	2.94	2.78	2.69	2.64	2.54	2.42
22	3.26	2.98	2.83	2.67	2.58	2.53	2.42	2.31
24	3.17	2.89	2.74	2.58	2.49	2.44	2.33	2.21
26	3.09	2.81	2.66	2.50	2.42	2.36	2.25	2.13
28	3.03	2.75	2.60	2.44	2.35	2.30	2.19	2.06
30	2.98	2.70	2.55	2.39	2.30	2.25	2.13	2.01
32	2.93	2.65	2.50	2.34	2.25	2.20	2.08	1.96
34	2.89	2.61	2.46	2.30	2.21	2.16	2.04	1.91
36	2.86	2.58	2.43	2.26	2.18	2.12	2.00	1.87
38	2.83	2.55	2.40	2.23	2.14	2.09	1.97	1.84
40	2.80	2.52	2.37	2.20	2.11	2.06	1.94	1.80
50	2.70	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	2.63	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	2.59	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	2.55	2.27	2.12	1.94	1.85	1.79	1.65	1.49
90	2.52	2.24	2.09	1.92	1.82	1.76	1.62	1.46
100	2.50	2.22	2.07	1.89	1.80	1.74	1.60	1.43
150	2.44	2.16	2.00	1.83	1.73	1.66	1.52	1.33
200	2.41	2.13	1.97	1.79	1.69	1.63	1.48	1.28
1000	2.34	2.06	1.90	1.72	1.61	1.54	1.38	1.11
∞	2.32	2.04	1.88	1.70	1.59	1.52	1.36	1.00

Table A12 Distribution Function $F(x) = P(T \leqq x)$ of the Random Variable T in Section 25.8

x	$n = 3$
	0.
0	167
1	500

x	$n = 4$
	0.
0	042
1	167
2	375

x	$n = 5$
	0.
0	008
1	042
2	117
3	242
4	408

x	$n = 6$
	0.
0	001
1	008
2	028
3	068
4	136
5	235
6	360
7	500

x	$n = 7$
	0.
1	001
2	005
3	015
4	035
5	068
6	119
7	191
8	281
9	386
10	500

x	$n = 8$
	0.
2	001
3	003
4	007
5	016
6	031
7	054
8	089
9	138
10	199
11	274
12	360
13	452

x	$n = 9$
	0.
4	001
5	003
6	006
7	012
8	022
9	038
10	060
11	090
12	130
13	179
14	238
15	306
16	381
17	460

x	$n = 10$
	0.
6	001
7	002
8	005
9	008
10	014
11	023
12	036
13	054
14	078
15	108
16	146
17	190
18	242
19	300
20	364
21	431
22	500

x	$n = 11$
	0.
8	001
9	002
10	003
11	005
12	008
13	013
14	020
15	030
16	043
17	060
18	082
19	109
20	141
21	179
22	223
23	271
24	324
25	381
26	440
27	500

x	$n = 20$
	0.
50	001
51	002
52	002
53	003
54	004
55	005
56	006
57	007
58	008
59	010
60	012
61	014
62	017
63	020
64	023
65	027
66	032
67	037
68	043
69	049
70	056
71	064
72	073
73	082
74	093
75	104
76	117
77	130
78	144
79	159
80	176
81	193
82	211
83	230
84	250
85	271
86	293
87	315
88	339
89	362
90	387
91	411
92	436
93	462
94	487

x	$n = 19$
	0.
43	001
44	002
45	002
46	003
47	003
48	004
49	005
50	006
51	008
52	010
53	012
54	014
55	017
56	021
57	025
58	029
59	034
60	040
61	047
62	054
63	062
64	072
65	082
66	093
67	105
68	119
69	133
70	149
71	166
72	184
73	203
74	223
75	245
76	267
77	290
78	314
79	339
80	365
81	391
82	418
83	445
84	473
85	500

x	$n = 18$
	0.
38	001
39	002
40	003
41	003
42	004
43	005
44	007
45	009
46	011
47	013
48	016
49	020
50	024
51	029
52	034
53	041
54	048
55	056
56	066
57	076
58	088
59	100
60	115
61	130
62	147
63	165
64	184
65	205
66	227
67	250
68	275
69	300
70	327
71	354
72	383
73	411
74	441
75	470
76	500

x	$n = 17$
	0.
32	001
33	002
34	002
35	003
36	004
37	005
38	007
39	009
40	011
41	014
42	017
43	021
44	026
45	032
46	038
47	046
48	054
49	064
50	076
51	088
52	102
53	118
54	135
55	154
56	174
57	196
58	220
59	245
60	271
61	299
62	328
63	358
64	388
65	420
66	452
67	484

x	$n = 16$
	0.
27	001
28	002
29	002
30	003
31	004
32	006
33	008
34	010
35	013
36	016
37	021
38	026
39	032
40	039
41	048
42	058
43	070
44	083
45	097
46	114
47	133
48	153
49	175
50	199
51	225
52	253
53	282
54	313
55	345
56	378
57	412
58	447
59	482

x	$n = 15$
	0.
23	001
24	002
25	003
26	004
27	006
28	008
29	010
30	014
31	018
32	023
33	029
34	037
35	046
36	057
37	070
38	084
39	101
40	120
41	141
42	164
43	190
44	218
45	248
46	279
47	313
48	349
49	385
50	423
51	461
52	500

x	$n = 14$
	0.
18	001
19	002
20	002
21	003
22	005
23	007
24	010
25	013
26	018
27	024
28	031
29	040
30	051
31	063
32	079
33	096
34	117
35	140
36	165
37	194
38	225
39	259
40	295
41	334
42	374
43	415
44	457
45	500

x	$n = 13$
	0.
14	001
15	001
16	002
17	003
18	005
19	007
20	011
21	015
22	021
23	029
24	038
25	050
26	064
27	082
28	102
29	126
30	153
31	184
32	218
33	255
34	295
35	338
36	383
37	429
38	476

x	$n = 12$
	0.
11	001
12	002
13	003
14	004
15	007
16	010
17	016
18	022
19	031
20	043
21	058
22	076
23	098
24	125
25	155
26	190
27	230
28	273
29	319
30	369
31	420
32	473

INDEX

PHOTO CREDITS

Part A Opener: © Denis Jr. Tangney/iStockphoto
Part B Opener: © Jill Fromer/iStockphoto
Part C Opener: © Science Photo Library/Photo Researchers, Inc
Part D Opener: © Rafa Irusta/iStockphoto
Part E Opener: © Alberto Pomares/iStockphoto
Chapter 19, Figure 437: © Eddie Gerald/Alamy
Part F Opener: © Rainer Plendl/iStockphoto
Part G Opener: © Sean Locke/iStockphoto
Appendix 1 Opener: © Ricardo De Mattos/iStockphoto
Appendix 2 Opener: © joel-t/iStockphoto
Appendix 3 Opener: © Luke Daniek/iStockphoto
Appendix 4 Opener: © Andrey Prokhorov/iStockphoto
Appendix 5 Opener: © Pedro Castellano/iStockphoto

Some Constants

$$
\begin{aligned}
e &= 2.71828\ 18284\ 59045\ 23536 \\
\sqrt{e} &= 1.64872\ 12707\ 00128\ 14685 \\
e^2 &= 7.38905\ 60989\ 30650\ 22723
\end{aligned}
$$

$$
\begin{aligned}
\pi &= 3.14159\ 26535\ 89793\ 23846 \\
\pi^2 &= 9.86960\ 44010\ 89358\ 61883 \\
\sqrt{\pi} &= 1.77245\ 38509\ 05516\ 02730
\end{aligned}
$$

$$
\begin{aligned}
\log_{10} \pi &= 0.49714\ 98726\ 94133\ 85435 \\
\ln \pi &= 1.14472\ 98858\ 49400\ 17414 \\
\log_{10} e &= 0.43429\ 44819\ 03251\ 82765 \\
\ln 10 &= 2.30258\ 50929\ 94045\ 68402
\end{aligned}
$$

$$
\begin{aligned}
\sqrt{2} &= 1.41421\ 35623\ 73095\ 04880 \\
\sqrt[3]{2} &= 1.25992\ 10498\ 94873\ 16477 \\
\sqrt{3} &= 1.73205\ 08075\ 68877\ 29353 \\
\sqrt[3]{3} &= 1.44224\ 95703\ 07408\ 38232 \\
\ln 2 &= 0.69314\ 71805\ 59945\ 30942 \\
\ln 3 &= 1.09861\ 22886\ 68109\ 69140
\end{aligned}
$$

$$
\begin{aligned}
\gamma &= 0.57721\ 56649\ 01532\ 86061 \\
\ln \gamma &= -0.54953\ 93129\ 81644\ 82234
\end{aligned}
$$

(see Sec. 5.6)

$$
\begin{aligned}
1^\circ &= 0.01745\ 32925\ 19943\ 29577 \text{ rad} \\
1 \text{ rad} &= 57.29577\ 95130\ 82320\ 87680^\circ \\
&= 57^\circ 17' 44.806''
\end{aligned}
$$

Polar Coordinates

$$x = r\cos\theta \qquad y = r\sin\theta$$

$$r = \sqrt{x^2 + y^2} \qquad \tan\theta = \frac{y}{x}$$

$$dx\,dy = r\,dr\,d\theta$$

Series

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m \qquad (|x| < 1)$$

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!}$$

$$\sin x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{(2m+1)!}$$

$$\cos x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{(2m)!}$$

$$\ln(1-x) = -\sum_{m=1}^{\infty} \frac{x^m}{m} \qquad (|x| < 1)$$

$$\arctan x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{2m+1} \qquad (|x| < 1)$$

Greek Alphabet

α	Alpha	ν	Nu
β	Beta	ξ	Xi
γ, Γ	Gamma	o	Omicron
δ, Δ	Delta	π	Pi
ϵ, ε	Epsilon	ρ	Rho
ζ	Zeta	σ, Σ	Sigma
η	Eta	τ	Tau
θ, ϑ, Θ	Theta	υ, Υ	Upsilon
ι	Iota	ϕ, φ, Φ	Phi
κ	Kappa	χ	Chi
λ, Λ	Lambda	ψ, Ψ	Psi
μ	Mu	ω, Ω	Omega

Vectors

$$\mathbf{a} \bullet \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$$

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

$$\operatorname{div} \mathbf{v} = \nabla \bullet \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$

$$\operatorname{curl} \mathbf{v} = \nabla \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix}$$