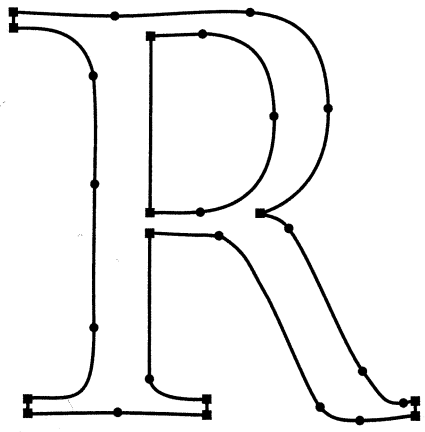
10.4 Cubic Spline Interpolation

In this section an artist's drafting aid is used as a physical model for the mathematical problem of finding a curve that passes through specified points in the plane. The parameters of the curve are determined by solving a linear system of equations.

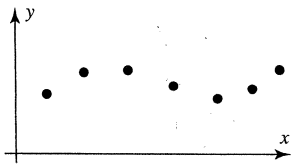
PREREQUISITES: Linear Systems
Matrix Algebra
Differential Calculus

Curve Fitting

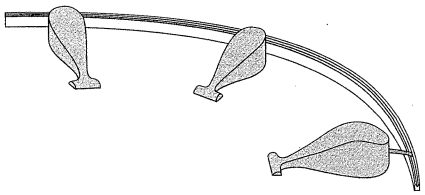
Fitting a curve through specified points in the plane is a common problem encountered in analyzing experimental data, in ascertaining the relations among variables, and in design work. A ubiquitous application is in the design and description of computer and printer fonts, such as PostScript™ and TrueType™ fonts (Figure 10.4.1). In Figure 10.4.2 seven points in the xy-plane are displayed, and in Figure 10.4.4 a smooth curve has been drawn that passes through them. A curve that passes through a set of points in the plane is said to ***interpolate*** those points, and the curve is called an ***interpolating curve*** for those points. The interpolating curve in Figure 10.4.4 was drawn with the aid of a *drafting spline* (Figure 10.4.3). This drafting aid consists of a thin, flexible strip of wood or other material that is bent to pass through the points to be interpolated. Attached sliding weights hold the spline in position while the artist draws the interpolating curve. The drafting spline will serve as the physical model for a mathematical theory of interpolation that we will discuss in this section.



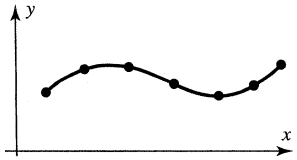
▶ Figure 10.4.1



▲ Figure 10.4.2



▲ Figure 10.4.3



▲ Figure 10.4.4

Statement of the Problem Suppose that we are given n points in the xy-plane,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

which we wish to interpolate with a "well-behaved" curve (Figure 10.4.5). For convenience, we take the points to be equally spaced in the x-direction, although our results can easily be extended to the case of unequally spaced points. If we let the common distance between the x-coordinates of the points be h, then we have

$$x_2 - x_1 = x_3 - x_2 = \cdots = x_n - x_{n-1} = h$$

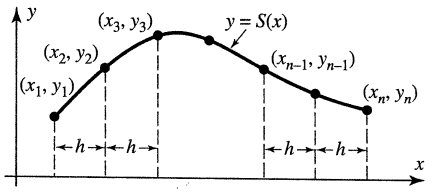
Let $y = S(x), x_1 \leq x \leq x_n$ denote the interpolating curve that we seek. We assume that this curve describes the displacement of a drafting spline that interpolates the n points when the weights holding down the spline are situated precisely at the n points. It is known from linear beam theory that for small displacements, the fourth derivative of the displacement of a beam is zero along any interval of the x-axis that contains no external forces acting on the beam. If we treat our drafting spline as a thin beam and realize that the only external forces acting on it arise from the weights at the n specified points, then it follows that

$$S^{(\text{iv})}(x) \equiv 0 \tag{1}$$

for values of x lying in the $n - 1$ open intervals

$$(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)$$

between the n points.



▶ **Figure 10.4.5**

We also need the result from linear beam theory that states that for a beam acted upon only by external forces, the displacement must have two continuous derivatives. In the case of the interpolating curve $y = S(x)$ constructed by the drafting spline, this means that $S(x)$, $S'(x)$, and $S''(x)$ must be continuous for $x_1 \leq x \leq x_n$.

The condition that $S''(x)$ be continuous is what causes a drafting spline to produce a pleasing curve, as it results in continuous *curvature*. The eye can perceive sudden changes in curvature—that is, discontinuities in $S''(x)$—but sudden changes in higher derivatives are not discernible. Thus, the condition that $S''(x)$ be continuous is the minimal prerequisite for the interpolating curve to be perceptible as a single smooth curve, rather than as a series of separate curves pieced together.

To determine the mathematical form of the function $S(x)$, we observe that because $S^{(\text{iv})}(x) \equiv 0$ in the intervals between the n specified points, it follows by integrating this equation four times that $S(x)$ must be a *cubic polynomial* in x in each such interval. In general, however, $S(x)$ will be a different cubic polynomial in each interval, so $S(x)$ must have the form

$$S(x) = \begin{cases} S_1(x), & x_1 \leq x \leq x_2 \\ S_2(x), & x_2 \leq x \leq x_3 \\ \quad\vdots & \\ S_{n-1}(x), & x_{n-1} \leq x \leq x_n \end{cases} \tag{2}$$

Section 10.10 Technology Exercises

The following exercises are designed to be solved using a technology utility. Typically, this will be MATLAB, *Mathematica*, Maple, Derive, or Mathcad, but it may also be some other type of linear algebra software or a scientific calculator with some linear algebra capabilities. For each exercise you will need to read the relevant documentation for the particular utility you are using. The goal of these exercises is to provide you with a basic proficiency with your technology utility. Once you have mastered the techniques in these exercises, you will be able to use your technology utility to solve many of the problems in the regular exercise sets.

T1. Let (a, b, c) be a unit vector normal to the plane $ax + by + cz = 0$, and let $\mathbf{r} = (x, y, z)$ be a vector. It can be shown that the mirror image of the vector $\mathbf{r}$ through the above plane has coordinates $\mathbf{r}_m = (x_m, y_m, z_m)$, where

$$\begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

with

$$M = I - 2\mathbf{n}\mathbf{n}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} a \\ b \\ c \end{bmatrix} [a \quad b \quad c]$$

a) Show that $M^2 = I$ and give a physical reason why this must be so. [*Hint:* Use the fact that (a, b, c) is a unit vector to show that $\mathbf{n}^T\mathbf{n} = 1$.]

b) Use a computer to show that $\det(M) = -1$.

c) The eigenvectors of M satisfy the equation

$$\begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

and therefore correspond to those vectors whose direction is not affected by a reflection through the plane. Use a computer to determine the eigenvectors and eigenvalues of M, and then give a physical argument to support your answer.

T2. A vector $\mathbf{v} = (x, y, z)$ is rotated by an angle θ about an axis having unit vector (a, b, c), thereby forming the rotated vector $\mathbf{v}_R = (x_R, y_R, z_R)$. It can be shown that

$$\begin{bmatrix} x_R \\ y_R \\ z_R \end{bmatrix} = R(\theta) \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

with

$$R(\theta) = \cos(\theta) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + (1 - \cos(\theta)) \begin{bmatrix} a \\ b \\ c \end{bmatrix} [a \quad b \quad c] + \sin(\theta) \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}$$

(a) Use a computer to show that $R(\theta)R(\varphi) = R(\theta + \varphi)$, and then give a physical reason why this must be so. Depending on the sophistication of the computer you are using, you may have to experiment using different values of a, b, and

$$c = \sqrt{1 - a^2 - b^2}$$

(b) Show also that $R^{-1}(\theta) = R(-\theta)$ and give a physical reason why this must be so.

(c) Use a computer to show that $\det(R(\theta)) = +1$.

10.11 Equilibrium Temperature Distributions

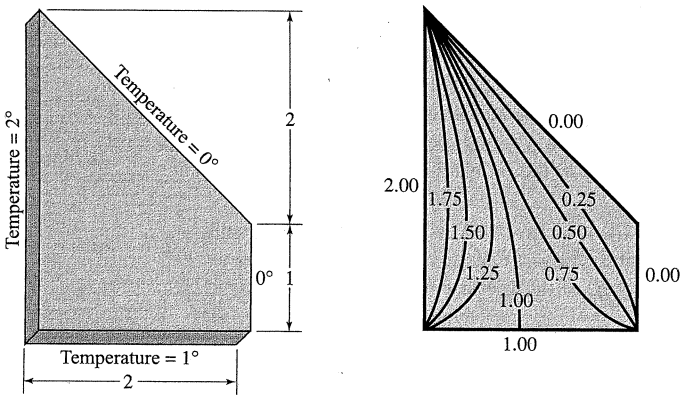
In this section we will see that the equilibrium temperature distribution within a trapezoidal plate can be found when the temperatures around the edges of the plate are specified. The problem is reduced to solving a system of linear equations. Also, an iterative technique for solving the problem and a "random walk" approach to the problem are described.

PREREQUISITES: Linear Systems
Matrices
Intuitive Understanding of Limits

Boundary Data

Suppose that the two faces of the thin trapezoidal plate shown in Figure 10.11.1*a* are insulated from heat. Suppose that we are also given the temperature along the four edges of the plate. For example, let the temperature be constant on each edge with values of

0°, 0°, 1°, and 2°, as in the figure. After a period of time, the temperature inside the plate will stabilize. Our objective in this section is to determine this equilibrium temperature distribution at the points inside the plate. As we will see, the interior equilibrium temperature is completely determined by the ***boundary data***—that is, the temperature along the edges of the plate.



▶ **Figure 10.11.1** (*a*) (*b*)

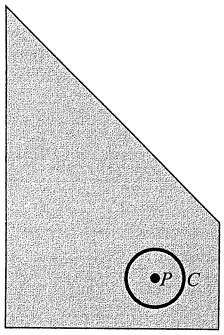
The equilibrium temperature distribution can be visualized by the use of curves that connect points of equal temperature. Such curves are called ***isotherms*** of the temperature distribution. In Figure 10.11.1*b* we have sketched a few isotherms, using information we derive later in the chapter.

Although all our calculations will be for the trapezoidal plate illustrated, our techniques generalize easily to a plate of any practical shape. They also generalize to the problem of finding the temperature within a three-dimensional body. In fact, our "plate" could be the cross section of some solid object if the flow of heat perpendicular to the cross section is negligible. For example, Figure 10.11.1 could represent the cross section of a long dam. The dam is exposed to three different temperatures: the temperature of the ground at its base, the temperature of the water on one side, and the temperature of the air on the other side. A knowledge of the temperature distribution inside the dam is necessary to determine the thermal stresses to which it is subjected.

Next we will consider a certain thermodynamic principle that characterizes the temperature distribution we are seeking.

The Mean-Value Property

There are many different ways to obtain a mathematical model for our problem. The approach we use is based on the following property of equilibrium temperature distributions.



▲ **Figure 10.11.2**

THEOREM 10.11.1 The Mean-Value Property

Let a plate be in thermal equilibrium and let P be a point inside the plate. Then if C is any circle with center at P that is completely contained in the plate, the temperature at P is the average value of the temperature on the circle (Figure 10.11.2).

This property is a consequence of certain basic laws of molecular motion, and we will not attempt to derive it. Basically, this property states that in equilibrium, thermal energy tends to distribute itself as evenly as possible consistent with the boundary conditions.

in these exercises, you will be able to use your technology utility to solve many of the problems in the regular exercise sets.

T1. Two integers that have no common factors (except 1) are said to be relatively prime. Given a positive integer n, let $S_n = \{a_1, a_2, a_3, \ldots, a_m\}$, where $a_1 < a_2 < a_3 < \cdots < a_m$, be the set of all positive integers less than n and relatively prime to n. For example, if $n = 9$, then

$$S_9 = \{a_1, a_2, a_3, \ldots, a_6\} = \{1, 2, 4, 5, 7, 8\}$$

(a) Construct a table consisting of n and S_n for $n = 2, 3, \ldots, 15$, and then compute

$$\sum_{k=1}^{m} a_k \quad \text{and} \quad \left(\sum_{k=1}^{m} a_k\right) \pmod{n}$$

in each case. Draw a conjecture for $n > 15$ and prove your conjecture to be true. [*Hint:* Use the fact that if a is relatively prime to n, then $n - a$ is also relatively prime to n.]

(b) Given a positive integer n and the set S_n, let P_n be the $m \times m$ matrix

$$P_n = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{m-1} & a_m \\ a_2 & a_3 & a_4 & \cdots & a_m & a_1 \\ a_3 & a_4 & a_5 & \cdots & a_1 & a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m-1} & a_m & a_1 & \cdots & a_{m-3} & a_{m-2} \\ a_m & a_1 & a_2 & \cdots & a_{m-2} & a_{m-1} \end{bmatrix}$$

so that, for example,

$$P_9 = \begin{bmatrix} 1 & 2 & 4 & 5 & 7 & 8 \\ 2 & 4 & 5 & 7 & 8 & 1 \\ 4 & 5 & 7 & 8 & 1 & 2 \\ 5 & 7 & 8 & 1 & 2 & 4 \\ 7 & 8 & 1 & 2 & 4 & 5 \\ 8 & 1 & 2 & 4 & 5 & 7 \end{bmatrix}$$

Use a computer to compute $\det(P_n)$ and $\det(P_n) \pmod{n}$ for $n = 2, 3, \ldots, 15$, and then use these results to construct a conjecture.

(c) Use the results of part (a) to prove your conjecture to be true. [*Hint:* Add the first $m - 1$ rows of P_n to its last row and then use Theorem 2.2.3.] What do these results imply about the inverse of $P_n \pmod{n}$?

T2. Given a positive integer n greater than 1, the number of positive integers less than n and relatively prime to n is called the ***Euler phi function*** of n and is denoted by $\varphi(n)$. For example, $\varphi(6) = 2$ since only two positive integers (1 and 5) are less than 6 and have no common factor with 6.

(a) Using a computer, for each value of $n = 2, 3, \ldots, 25$ compute and print out all positive integers that are less than n and relatively prime to n. Then use these integers to determine the values of $\varphi(n)$ for $n = 2, 3, \ldots, 25$. Can you discover a pattern in the results?

(b) It can be shown that if $\{p_1, p_2, p_3, \ldots, p_m\}$ are all the distinct prime factors of n, then

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\left(1 - \frac{1}{p_3}\right)\cdots\left(1 - \frac{1}{p_m}\right)$$

For example, since $\{2, 3\}$ are the distinct prime factors of 12, we have

$$\varphi(12) = 12\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right) = 4$$

which agrees with the fact that $\{1, 5, 7, 11\}$ are the only positive integers less than 12 and relatively prime to 12. Using a computer, print out all the prime factors of n for $n = 2, 3, \ldots, 25$. Then compute $\varphi(n)$ using the formula above and compare it to your results in part (a).

10.16 Genetics

In this section we investigate the propagation of an inherited trait in successive generations by computing powers of a matrix.

PREREQUISITES: Eigenvalues and Eigenvectors
Diagonalization of a Matrix
Intuitive Understanding of Limits

Inheritance Traits

In this section we examine the inheritance of traits in animals or plants. The inherited trait under consideration is assumed to be governed by a set of two genes, which we designate by A and a. Under ***autosomal inheritance*** each individual in the population of either gender possesses two of these genes, the possible pairings being designated AA, Aa, and aa. This pair of genes is called the individual's ***genotype***, and it determines how the trait controlled by the genes is manifested in the individual. For example,

in snapdragons a set of two genes determines the color of the flower. Genotype AA produces red flowers, genotype Aa produces pink flowers, and genotype aa produces white flowers. In humans, eye coloration is controlled through autosomal inheritance. Genotypes AA and Aa have brown eyes, and genotype aa has blue eyes. In this case we say that gene A ***dominates*** gene a, or that gene a is ***recessive*** to gene A, because genotype Aa has the same outward trait as genotype AA.

In addition to autosomal inheritance we will also discuss ***X-linked inheritance***. In this type of inheritance, the male of the species possesses only one of the two possible genes (A or a), and the female possesses a pair of the two genes (AA, Aa, or aa). In humans, color blindness, hereditary baldness, hemophilia, and muscular dystrophy, to name a few, are traits controlled by X-linked inheritance.

Below we explain the manner in which the genes of the parents are passed on to their offspring for the two types of inheritance. We construct matrix models that give the probable genotypes of the offspring in terms of the genotypes of the parents, and we use these matrix models to follow the genotype distribution of a population through successive generations.

Autosomal Inheritance

In autosomal inheritance an individual inherits one gene from each of its parents' pairs of genes to form its own particular pair. As far as we know, it is a matter of chance which of the two genes a parent passes on to the offspring. Thus, if one parent is of genotype Aa, it is equally likely that the offspring will inherit the A gene or the a gene from that parent. If one parent is of genotype aa and the other parent is of genotype Aa, the offspring will always receive an a gene from the aa parent and will receive either an A gene or an a gene, with equal probability, from the Aa parent. Consequently, each of the offspring has equal probability of being genotype aa or Aa. In Table 1 we list the probabilities of the possible genotypes of the offspring for all possible combinations of the genotypes of the parents.

Table 1

Genotype of Offspring	Genotypes of Parents					
	AA–AA	*AA–Aa*	*AA–aa*	*Aa–Aa*	*Aa–aa*	*aa–aa*
AA	1	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0
Aa	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0
aa	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$	1

▶ EXAMPLE 1 Distribution of Genotypes in a Population

Suppose that a farmer has a large population of plants consisting of some distribution of all three possible genotypes AA, Aa, and aa. The farmer desires to undertake a breeding program in which each plant in the population is always fertilized with a plant of genotype AA and is then replaced by one of its offspring. We want to derive an expression for the distribution of the three possible genotypes in the population after any number of generations.

For $n = 0, 1, 2, \ldots$, let us set

a_n = fraction of plants of genotype AA in nth generation

b_n = fraction of plants of genotype Aa in nth generation

c_n = fraction of plants of genotype aa in nth generation

10.17 Age-Specific Population Growth

In this section we investigate, using the Leslie matrix model, the growth over time of a female population that is divided into age classes. We then determine the limiting age distribution and growth rate of the population.

PREREQUISITES: Eigenvalues and Eigenvectors
Diagonalization of a Matrix
Intuitive Understanding of Limits

One of the most common models of population growth used by demographers is the so-called Leslie model developed in the 1940s. This model describes the growth of the female portion of a human or animal population. In this model the females are divided into age classes of equal duration. To be specific, suppose that the maximum age attained by any female in the population is L years (or some other time unit) and we divide the population into n age classes. Then each class is L/n years in duration. We label the age classes according to Table 1.

Table 1

Age Class	Age Interval
1	$[0, L/n)$
2	$[L/n, 2L/n)$
3	$[2L/n, 3L/n)$
$\vdots$	$\vdots$
$n-1$	$[(n-2)L/n, (n-1)L/n)$
n	$[(n-1)L/n, L]$

Suppose that we know the number of females in each of the n classes at time $t = 0$. In particular, let there be $x_1^{(0)}$ females in the first class, $x_2^{(0)}$ females in the second class, and so forth. With these n numbers we form a column vector:

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}$$

We call this vector the ***initial age distribution vector***.

As time progresses, the number of females within each of the n classes changes because of three biological processes: birth, death, and aging. By describing these three processes quantitatively, we will see how to project the initial age distribution vector into the future.

The easiest way to study the aging process is to observe the population at discrete times—say, $t_0, t_1, t_2, \ldots, t_k, \ldots$. The Leslie model requires that the duration between any two successive observation times be the same as the duration of the age intervals.

Therefore, we set

$$\begin{aligned} t_0 &= 0 \\ t_1 &= L/n \\ t_2 &= 2L/n \\ &\vdots \\ t_k &= kL/n \\ &\vdots \end{aligned}$$

With this assumption, all females in the $(i+1)$-st class at time t_{k+1} were in the ith class at time t_k.

The birth and death processes between two successive observation times can be described by means of the following demographic parameters:

a_i $(i = 1, 2, \ldots, n)$	The average number of daughters born to each female during the time she is in the ith age class
b_i $(i = 1, 2, \ldots, n-1)$	The fraction of females in the ith age class that can be expected to survive and pass into the $(i+1)$-st age class

By their definitions, we have that

$$\begin{aligned} &\text{(i) } a_i \geq 0 && \text{for } i = 1, 2, \ldots, n \\ &\text{(ii) } 0 < b_i \leq 1 && \text{for } i = 1, 2, \ldots, n-1 \end{aligned}$$

Note that we do not allow any b_i to equal zero, because then no females would survive beyond the ith age class. We also assume that at least one a_i is positive so that some births occur. Any age class for which the corresponding value of a_i is positive is called a ***fertile age class***.

We next define the age distribution vector $\mathbf{x}^{(k)}$ at time t_k by

$$\mathbf{x}^{(k)} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix}$$

where $x_i^{(k)}$ is the number of females in the ith age class at time t_k. Now, at time t_k, the females in the first age class are just those daughters born between times t_{k-1} and t_k. Thus, we can write

$$\left\{\begin{array}{c}\text{number of}\\ \text{females}\\ \text{in class 1}\\ \text{at time } t_k\end{array}\right\} = \left\{\begin{array}{c}\text{number of}\\ \text{daughters}\\ \text{born to}\\ \text{females in}\\ \text{class 1}\\ \text{between times}\\ t_{k-1} \text{ and } t_k\end{array}\right\} + \left\{\begin{array}{c}\text{number of}\\ \text{daughters}\\ \text{born to}\\ \text{females in}\\ \text{class 2}\\ \text{between times}\\ t_{k-1} \text{ and } t_k\end{array}\right\} + \cdots + \left\{\begin{array}{c}\text{number of}\\ \text{daughters}\\ \text{born to}\\ \text{females in}\\ \text{class } n\\ \text{between times}\\ t_{k-1} \text{ and } t_k\end{array}\right\}$$

T1. The methods of Exercise 4 show that for the cat map, $\Pi(p)$ is the smallest integer satisfying the equation

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{\Pi(p)} \bmod p = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This suggests that one way to determine $\Pi(p)$ is to compute

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{n} \bmod p$$

starting with $n = 1$ and stopping when this produces the identity matrix. Use this idea to compute $\Pi(p)$ for $p = 2, 3, \ldots, 10$. Compare your results to the formulas given in Exercise 1, if they apply. What can you conjecture about

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{\frac{1}{2}\Pi(p)} \bmod p$$

when $\Pi(p)$ is even?

T2. The eigenvalues and eigenvectors for the cat map matrix

$$C = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

are

$$\lambda_1 = \frac{3+\sqrt{5}}{2}, \qquad \lambda_2 = \frac{3-\sqrt{5}}{2},$$

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ \dfrac{1+\sqrt{5}}{2} \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 1 \\ \dfrac{1-\sqrt{5}}{2} \end{bmatrix}$$

Using these eigenvalues and eigenvectors, we can define

$$D = \begin{bmatrix} \dfrac{3+\sqrt{5}}{2} & 0 \\ 0 & \dfrac{3-\sqrt{5}}{2} \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} 1 & 1 \\ \dfrac{1+\sqrt{5}}{2} & \dfrac{1-\sqrt{5}}{2} \end{bmatrix}$$

and write $C = PDP^{-1}$; hence, $C^n = PD^nP^{-1}$. Use a computer to show that

$$C^n = \begin{bmatrix} c_{11}^{(n)} & c_{12}^{(n)} \\ c_{21}^{(n)} & c_{22}^{(n)} \end{bmatrix}$$

where

$$c_{11}^{(n)} = \left(\frac{1+\sqrt{5}}{2\sqrt{5}}\right)\left(\frac{3-\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2\sqrt{5}}\right)\left(\frac{3+\sqrt{5}}{2}\right)^n$$

$$c_{22}^{(n)} = \left(\frac{1+\sqrt{5}}{2\sqrt{5}}\right)\left(\frac{3+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2\sqrt{5}}\right)\left(\frac{3-\sqrt{5}}{2}\right)^n$$

and

$$c_{12}^{(n)} = c_{21}^{(n)} = \frac{1}{\sqrt{5}}\left\{\left(\frac{3+\sqrt{5}}{2}\right)^n - \left(\frac{3-\sqrt{5}}{2}\right)^n\right\}$$

How can you use these results and your conclusions in Exercise T1 to simplify the method for computing $\Pi(p)$?

10.15 Cryptography

In this section we present a method of encoding and decoding messages. We also examine modular arithmetic and show how Gaussian elimination can sometimes be used to break an opponent's code.

PREREQUISITES: Matrices
Gaussian Elimination
Matrix Operations
Linear Independence
Linear Transformations (Section 4.9)

Ciphers

The study of encoding and decoding secret messages is called ***cryptography***. Although secret codes date to the earliest days of written communication, there has been a recent surge of interest in the subject because of the need to maintain the privacy of information transmitted over public lines of communication. In the language of cryptography, codes are called ***ciphers***, uncoded messages are called ***plaintext***, and coded messages are called ***ciphertext***. The process of converting from plaintext to ciphertext is called ***enciphering***, and the reverse process of converting from ciphertext to plaintext is called ***deciphering***.

The simplest ciphers, called ***substitution ciphers***, are those that replace each letter of the alphabet by a different letter. For example, in the substitution cipher

Plain	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*
Cipher	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*	*A*	*B*	*C*

the plaintext letter A is replaced by D, the plaintext letter B by E, and so forth. With this cipher the plaintext message

ROME WAS NOT BUILT IN A DAY

becomes

URPH ZDV QRW EXLOW LQ D GDB

Hill Ciphers

A disadvantage of substitution ciphers is that they preserve the frequencies of individual letters, making it relatively easy to break the code by statistical methods. One way to overcome this problem is to divide the plaintext into groups of letters and encipher the plaintext group by group, rather than one letter at a time. A system of cryptography in which the plaintext is divided into sets of n letters, each of which is replaced by a set of n cipher letters, is called a ***polygraphic system***. In this section we will study a class of polygraphic systems based on matrix transformations. [The ciphers that we will discuss are called ***Hill ciphers*** after Lester S. Hill, who introduced them in two papers: "Cryptography in an Algebraic Alphabet," *American Mathematical Monthly, 36* (June–July 1929), pp. 306–312; and "Concerning Certain Linear Transformation Apparatus of Cryptography," *American Mathematical Monthly, 38* (March 1931), pp. 135–154.]

In the discussion to follow, we assume that each plaintext and ciphertext letter except Z is assigned the numerical value that specifies its position in the standard alphabet (Table 1). For reasons that will become clear later, Z is assigned a value of zero.

Table 1

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	0

In the simplest Hill ciphers, successive *pairs* of plaintext are transformed into ciphertext by the following procedure:

Step 1. Choose a 2×2 matrix with integer entries

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

to perform the encoding. Certain additional conditions on A will be imposed later.

Step 2. Group successive plaintext letters into pairs, adding an arbitrary "dummy" letter to fill out the last pair if the plaintext has an odd number of letters, and replace each plaintext letter by its numerical value.

Step 3. Successively convert each plaintext pair $p_1 p_2$ into a column vector

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

and form the product $A\mathbf{p}$. We will call $\mathbf{p}$ a ***plaintext vector*** and $A\mathbf{p}$ the corresponding ***ciphertext vector***.

Step 4. Convert each ciphertext vector into its alphabetic equivalent.

10.3 The Earliest Applications of Linear Algebra

Linear systems can be found in the earliest writings of many ancient civilizations. In this section we give some examples of the types of problems that they used to solve.

PREREQUISITES: Linear Systems

The practical problems of early civilizations included the measurement of land, the distribution of goods, the tracking of resources such as wheat and cattle, and taxation and inheritance calculations. In many cases, these problems led to linear systems of equations since linearity is one of the simplest relationships that can exist among variables. In this section we present examples from five diverse ancient cultures illustrating how they used and solved systems of linear equations. We restrict ourselves to examples before A.D. 500. These examples consequently predate the development of the field of algebra by Islamic/Arab mathematicians, a field that ultimately led in the nineteenth century to the branch of mathematics now called linear algebra.

▶ EXAMPLE 1 Egypt (about 1650 B.C.)

Problem 40 of the Ahmes Papyrus

The Ahmes (or Rhind) Papyrus is the source of most of our information about ancient Egyptian mathematics. This 5-meter-long papyrus contains 84 short mathematical problems, together with their solutions, and dates from about 1650 B.C. Problem 40 in this papyrus is the following:

Divide 100 hekats of barley among five men in arithmetic progression so that the sum of the two smallest is one-seventh the sum of the three largest.

Let a be the least amount that any man obtains, and let d be the common difference of the terms in the arithmetic progression. Then the other four men receive $a + d$, $a + 2d$, $a + 3d$, and $a + 4d$ hekats. The two conditions of the problem require that

$$a + (a + d) + (a + 2d) + (a + 3d) + (a + 4d) = 100$$
$$\tfrac{1}{7}[(a + 2d) + (a + 3d) + (a + 4d)] = a + (a + d)$$

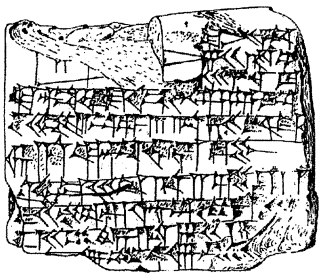
These equations reduce to the following system of two equations in two unknowns:

$$\begin{aligned} 5a + 10d &= 100 \\ 11a - 2d &= 0 \end{aligned} \tag{1}$$

The solution technique described in the papyrus is known as the method of false position or false assumption. It begins by assuming some convenient value of a (in our case $a = 1$), substituting that value into the second equation, and obtaining $d = 11/2$. Substituting $a = 1$ and $d = 11/2$ into the left-hand side of the first equation gives 60, whereas

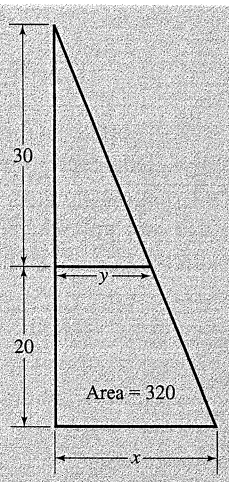
the right-hand side is 100. Adjusting the initial guess for a by multiplying it by 100/60 leads to the correct value $a = 5/3$. Substituting $a = 5/3$ into the second equation then gives $d = 55/6$, so the quantities of barley received by the five men are 10/6, 65/6, 120/6, 175/6, and 230/6 hekats. This technique of guessing a value of an unknown and later adjusting it has been used by many cultures throughout the ages.

▶ EXAMPLE 2 Babylonia (1900–1600 B.C.)



Babylonian clay tablet Ca MLA 1950

The Old Babylonian Empire flourished in Mesopotamia between 1900 and 1600 B.C. Many clay tablets containing mathematical tables and problems survive from that period, one of which (designated Ca MLA 1950) contains the next problem. The statement of the problem is a bit muddled because of the condition of the tablet, but the diagram and the solution on the tablet indicate that the problem is as follows:



A trapezoid with an area of 320 square units is cut off from a right triangle by a line parallel to one of its sides. The other side has length 50 units, and the height of the trapezoid is 20 units. What are the upper and the lower widths of the trapezoid?

Let x be the lower width of the trapezoid and y its upper width. The area of the trapezoid is its height times its average width, so $20\left(\frac{x+y}{2}\right) = 320$. Using similar triangles, we also have $\frac{x}{50} = \frac{y}{30}$. The solution on the tablet uses these relations to generate the linear system

$$\begin{aligned} \tfrac{1}{2}(x + y) &= 16 \\ \tfrac{1}{2}(x - y) &= 4 \end{aligned} \tag{2}$$

Adding and subtracting these two equations then gives the solution $x = 20$ and $y = 12$.

▶ EXAMPLE 3 China (A.D. 263)

九章算術

Chiu Chang Suan Shu in Chinese characters

The most important treatise in the history of Chinese mathematics is the Chiu Chang Suan Shu, or "The Nine Chapters of the Mathematical Art." This treatise, which is a collection of 246 problems and their solutions, was assembled in its final form by Liu Hui in A.D. 263. Its contents, however, go back to at least the beginning of the Han dynasty in the second century B.C. The eighth of its nine chapters, entitled "The Way of Calculating by Arrays," contains 18 word problems that lead to linear systems in three to six unknowns. The general solution procedure described is almost identical to the