Section 2
Mathematics

BY

C. EDWARD SANDIFER *Professor, Western Connecticut State University, Danbury, CT.*
GEORGE J. MOSHOS *Professor Emeritus of Computer and Information Science, New Jersey Institute of Technology*

2.1 MATHEMATICS
by C. Edward Sandifer

2.2 COMPUTERS
by George J. Moshos

2.1 MATHEMATICS

by C. Edward Sandifer

REFERENCES: Conte and DeBoor, ‘‘Elementary Numerical Analysis: An Algorithmic Approach,’’ McGraw-Hill. Boyce and DiPrima, ‘‘Elementary Differential Equations and Boundary Value Problems,’’ Wiley. Hamming, ‘‘Numerical Methods for Scientists and Engineers,’’ McGraw-Hill. Kreyszig, ‘‘Advanced Engineering Mathematics,’’ Wiley.

SETS, NUMBERS, AND ARITHMETIC

Sets and Elements

The concept of a set appears throughout modern mathematics. A **set** is a well-defined list or collection of objects and is generally denoted by capital letters, $A, B, C, \ldots$. The objects composing the set are called **elements** and are denoted by lowercase letters, $a, b, x, y, \ldots$. The notation

$$x \in A$$

is read ‘‘x is an element of A’’ and means that x is one of the objects composing the set A.

There are two basic ways to describe a set. The first way is to list the elements of the set.

$$A = \{2, 4, 6, 8, 10\}$$

This often is not practical for very large sets.

The second way is to describe properties which determine the elements of the set.

$$A = \{\text{even numbers from 2 to 10}\}$$

This method is sometimes awkward since a single set may sometimes be described in several different ways.

In describing sets, the symbol : is read ‘‘such that.’’ The expression

$$B = \{x : x \text{ is an even integer}, x > 1, x < 11\}$$

is read ‘‘B equals the set of all x such that x is an even integer, x is greater than 1, and x is less than 11.’’

Two sets, A and B, are equal, written $A = B$, if they contain exactly the same elements. The sets A and B above are equal. If two sets, X and Y, are not equal, it is written $X \neq Y$.

Subsets A set C is a subset of a set A, written $C \subseteq A$, if each element in C is also an element in A. It is also said that C is contained in A. Any set is a subset of itself. That is, $A \subseteq A$ always. A is said to be an ‘‘improper subset of itself.’’ Otherwise, if $C \subseteq A$ and $C \neq A$, then C is a proper subset of A.

Two theorems are important about subsets:
(Fundamental theorem of set equality)

$$\text{If } X \subseteq Y \quad \text{and} \quad Y \subseteq X, \quad \text{then } X = Y \tag{2.1.1}$$

(Transitivity)

$$\text{If } X \subseteq Y \quad \text{and} \quad Y \subseteq Z, \quad \text{then } X \subseteq Z \tag{2.1.2}$$

Universe and Empty Set In an application of set theory, it often happens that all sets being considered are subsets of some fixed set, say integers or vectors. This fixed set is called the **universe** and is sometimes denoted U.

It is possible that a set contains no elements at all. The set with no elements is called the **empty set** or the **null set** and is denoted $\varnothing$.

Set Operations New sets may be built from given sets in several ways. The **union** of two sets, denoted $A \cup B$, is the set of all elements belonging to A or to B, or to both.

$$A \cup B = \{x : x \in A \quad \text{or} \quad x \in B\}$$

The union has the properties:

$$A \subseteq A \cup B \qquad \text{and} \qquad B \subseteq A \cup B \tag{2.1.3}$$

The **intersection** is denoted $A \cap B$ and consists of all elements, each of which belongs to both A and B.

$$A \cap B = \{x : x \in A \quad \text{and} \quad x \in B\}$$

The intersection has the properties

$$A \cap B \subseteq A \qquad \text{and} \qquad A \cap B \subseteq B \tag{2.1.4}$$

If $A \cap B = \varnothing$, then A and B are called **disjoint.**

In general, a union makes a larger set and an intersection makes a smaller set.

The **complement** of a set A is the set of all elements in the universe set which are not in A. This is written

$$\sim\!A = \{x : x \in U, \quad x \notin A\}$$

The difference of two sets, denoted $A - B$, is the set of all elements which belong to A but do not belong to B.

Algebra on Sets The operations of union, intersection, and complement obey certain laws known as **Boolean algebra.** Using these laws, it is possible to convert an expression involving sets into other equivalent expressions. The laws of Boolean algebra are given in Table 2.1.1.

Venn Diagrams To give a pictorial representation of a set, **Venn diagrams** are often used. Regions in the plane are used to correspond to sets, and areas are shaded to indicate unions, intersections, and complements. Examples of Venn diagrams are given in Fig. 2.1.1.

Numbers

Numbers are the basic instruments of computation. It is by operations on numbers that calculations are made. There are several different kinds of numbers.

Natural numbers, or counting numbers, denoted **N,** are the whole numbers greater than zero. Sometimes zero is included as a natural number. Any two natural numbers may be added or multiplied to give

Table 2.1.1 Laws of Boolean Algebra

1. Idempotency	
$A \cup A = A$	$A \cap A = A$
2. Associativity	
$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
3. Commutativity	
$A \cup B = B \cup A$	$A \cap B = B \cap A$
4. Distributivity	
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
5. Identity	
$A \cup \varnothing = A$	$A \cap U = A$
$A \cup U = U$	$A \cap \varnothing = \varnothing$
6. Complement	
$A \cup \sim\!A = U$	$A \cap \sim\!A = \varnothing$
$\sim(\sim\!A) = A$	
$\sim\!U = \varnothing$	
$\sim\!\varnothing = U$	
7. DeMorgan's laws	
$\sim(A \cup B) = \sim\!A \cap \sim\!B$	$\sim(A \cap B) = \sim\!A \cup \sim\!B$

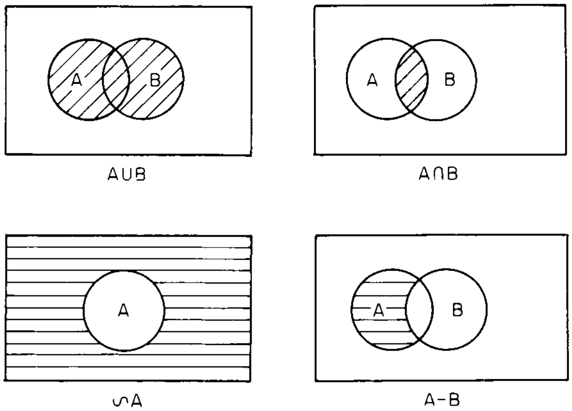


Fig. 2.1.1 Venn diagrams.

another natural number, but subtracting them may produce a negative number, which is not a natural number, and dividing them may produce a fraction, which is not a natural number.

Integers, or whole numbers, are denoted by **Z.** They include both positive and negative numbers and zero. Integers may be added, subtracted, and multiplied, but division might not produce an integer.

Real numbers, denoted **R,** are essentially all values which it is possible for a measurement to take, or all possible lengths for line segments. **Rational numbers** are real numbers that are the quotient of two integers, for example, $^{11}/_{78}$. **Irrational numbers** are not the quotient of two integers, for example, π and $\sqrt{2}$. Within the real numbers, it is always possible to add, subtract, multiply, and divide (except division by zero).

Complex numbers, or **imaginary numbers,** denoted **C,** are an extension of the real numbers that include the square root of -1, denoted i. Within the real numbers, only positive numbers have square roots. Within the complex numbers, all numbers have square roots.

Any complex number z can be written uniquely as $z = x + iy$, where x and y are real. Then x is the real part of z, denoted Re(z), and y is the imaginary part, denoted Im(z).

The **complex conjugate,** or simply conjugate of a complex number, z is $\bar{z} = x - iy$.

If $z = x + iy$ and $w = u + iv$, then z and w may be manipulated as follows:

$$\begin{aligned} z + w &= (x + u) + i(y + v) \\ z - w &= (x - u) + i(y - v) \\ zw &= xu - yv + i(xv + yu) \\ \frac{z}{w} &= \frac{xu + yv + i(yu - xv)}{u^2 + v^2} \end{aligned}$$

As sets, the following relation exists among these different kinds of numbers:

$$\mathbf{N} \subseteq \mathbf{Z} \subseteq \mathbf{R} \subseteq \mathbf{C}$$

Functions

A **function** f is a rule that relates two sets A and B. Given an element x of the set A, the function assigns a unique element y from the set B. This is written

$$y = f(x)$$

The set A is called the **domain** of the function, and the set B is called the **range.** It is possible for A and B to be the same set.

Functions are usually described by giving the rule. For example,

$$f(x) = 3x + 4$$

is a rule for a function with range and domain both equal to **R.** Given a value, say, 2, from the domain, $f(2) = 3(2) + 4 = 10$.

If two functions f and g have the same range and domain and if the ranges are numbers, then f and g may be added, subtracted, multiplied, or divided according to the rules of the range. If $f(x) = 3x + 4$ and $g(x) = \sin(x)$ and both have range and domain equal to **R,** then

$$f + g(x) = 3x + 4 + \sin(x)$$

and

$$\frac{f}{g}(x) = \frac{3x + 4}{\sin x}$$

Dividing functions occasionally leads to complications when one of the functions assumes a value of zero. In the example f/g above, this occurs when $x = 0$. The quotient cannot be evaluated for $x = 0$ although the quotient function is still meaningful. In this case, the function f/g is said to have a **pole** at $x = 0$.

Polynomial functions are functions of the form

$$f(x) = \sum_{i=0}^{n} a_i x^i$$

where $a_n \neq 0$. The domain and range of polynomial functions are always either **R** or **C.** The number n is the degree of the polynomial.

Polynomials of degree 0 or 1 are called **linear;** of degree 2 they are called **parabolic** or **quadratic;** and of degree 3 they are called **cubic.**

The values of f for which $f(x) = 0$ are called the **roots of *f*.** A polynomial of degree n has at most n roots. There is exactly one exception to this rule: If $f(x) = 0$ is the constant zero function, the degree of f is zero, but f has infinitely many roots.

Roots of polynomials of degree 1 are found as follows: Suppose the polynomial is $f(x) = ax + b$. Set $f(x) = 0$ and solve for x. Then $x = -b/a$.

Roots of polynomials of degree 2 are often found using the quadratic formula. If $f(x) = ax^2 + bx + c$, then the two roots of f are given by the **quadratic formula:**

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad \text{and} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Roots of a polynomial of degree 3 fall into two types.

Equations of the Third Degree with Term in x^2 Absent

Solution: After dividing through by the coefficient of x^3, any equation of this type can be written $x^3 = Ax + B$. Let $p = A/3$ and $q = B/2$. The general solution is as follows:

CASE 1. $q^2 - p^3$ positive. One root is real, viz.,

$$x_1 = \sqrt[3]{q + \sqrt{q^2 - p^3}} + \sqrt[3]{q - \sqrt{q^2 - p^3}}$$

The other two roots are imaginary.

CASE 2. $q^2 - p^3 =$ zero. Three roots real, but two of them equal.

$$x_1 = 2\sqrt[3]{q} \qquad x_2 = -\sqrt[3]{q} \quad x_3 = -\sqrt[3]{q}$$

CASE 3. $q^2 - p^3$ negative. All three roots are real and distinct. Determine an angle u between 0 and 180°, such that $\cos u = q/(p\sqrt{p})$. Then

$$\begin{aligned} x_1 &= 2\sqrt{p}\cos(u/3) \\ x_2 &= 2\sqrt{p}\cos(u/3 + 120°) \\ x_3 &= 2\sqrt{p}\cos(u/3 + 240°) \end{aligned}$$

Graphical Solution: Plot the curve $y_1 = x^3$, and the straight line $y_2 = Ax + B$. The abscissas of the points of intersection will be the roots of the equation.

Equations of the Third Degree (General Case)

Solution: The general cubic equation, after dividing through by the coefficient of the highest power, may be written $x^3 + ax^2 + bx + c = 0$. To get rid of the term in x^2, let $x = x_1 - a/3$. The equation then becomes $x_1^3 = Ax_1 + B$, where $A = 3(a/3)^2 - b$, and $B = -2(a/3)^3 + b(a/3) - c$. Solve this equation for x_1, by the method above, and then find x itself from $x = x_1 - (a/3)$.

Graphical Solution: Without getting rid of the term in x^2, write the equation in the form $x^3 = -a[x + (b/2a)]^2 + [a(b/2a)^2 - c]$, and solve by the graphical method.

Arithmetic

When numbers, functions, or vectors are manipulated, they always obey certain properties, regardless of the types of the objects involved. Elements may be added or subtracted only if they are in the same universe set. Elements in different universes may sometimes be multiplied or divided, but the result may be in a different universe. Regardless of the universe sets involved, the following properties hold true:

1. Associative laws. $a + (b + c) = (a + b) + c$, $a(bc) = (ab)c$
2. Identity laws. $0 + a = a$, $1a = a$
3. Inverse laws. $a - a = 0$, $a/a = 1$
4. Distributive law. $a(b + c) = ab + ac$
5. Commutative laws. $a + b = b + a$, $ab = ba$

Certain universes, for example, matrices, do not obey the commutative law for multiplication.

SIGNIFICANT FIGURES AND PRECISION

Number of Significant Figures In engineering computations, the data are ordinarily the result of measurement and are correct only to a limited number of significant figures. Each of the numbers 3.840 and 0.003840 is said to be given ‘‘correct to four figures’’; the true value lies in the first case between 0.0038395 and 0.0038405. The **absolute error** is less than 0.001 in the first case, and less than 0.000001 in the second; but the **relative error** is the same in both cases, namely, an error of less than ‘‘one part in 3,840.’’

If a number is written as 384,000, the reader is left in doubt whether the number of correct significant figures is 3, 4, 5, or 6. This doubt can be removed by writing the number as 3.84×10^5, or 3.840×10^5, or 3.8400×10^5, or 3.84000×10^5.

In any numerical computation, the possible or desirable degree of accuracy should be decided on and the computation should then be so arranged that the required number of significant figures, and no more, is secured. Carrying out the work to a larger number of places than is justified by the data is to be avoided, (1) because the form of the results leads to an erroneous impression of their accuracy and (2) because time and labor are wasted in superfluous computation.

The unit value of the least significant figure in a number is its **precision.** The number 123.456 has six significant figures and has precision 0.001.

Two ways to represent a real number are as **fixed-point** or as **floating-point,** also known as ‘‘scientific notation.’’

In scientific notation, a number is represented as a product of a **mantissa** and a power of 10. The mantissa has its first significant figure either immediately before or immediately after the decimal point, depending on which convention is being used. The power of 10 used is called the **exponent.** The number 123.456 may be represented as either

$$0.123456 \times 10^3 \qquad \text{or} \qquad 1.23456 \times 10^2$$

Fixed-point representations tend to be more convenient when the quantities involved will be added or subtracted or when all measurements are taken to the same precision. Floating-point representations are more convenient for very large or very small numbers or when the quantities involved will be multiplied or divided.

Many different numbers may share the same representation. For example, 0.05 may be used to represent, with precision 0.01, any value between 0.045000 and 0.054999. The largest value a number represents, in this case 0.0549999, is sometimes denoted x^*, and the smallest is denoted x_*.

An awareness of precision and significant figures is necessary so that answers correctly represent their accuracy.

Multiplication and Division A product or quotient should be written with the smallest number of significant figures of any of the factors involved. The product often has a different precision than the factors, but the significant figures must not increase.

EXAMPLES. (6.)(8.) = 48 should be written as 50 since the factors have one significant figure. There is a loss of precision from 1 to 10.

$0.6 \times 0.8 = .048$ should be written as 0.5 since the factors have one significant figure. There is a gain of precision from 0.1 to 0.01.

Addition and Subtraction A sum or difference should be represented with the same precision as the least precise term involved. The number of significant figures may change.

EXAMPLES. $3.14 + 0.001 = 3.141$ should be represented as 3.14 since the least precise term has precision 0.01.

$3.14 + 0.1 = 3.24$ should be represented as 3.2 since the least precise term has precision 0.1.

Loss of Significant Figures Addition and subtraction may result in serious loss of significant figures and resultant large relative errors if the sums are near zero. For example,

$$3.15 - 3.14 = 0.01$$

shows a loss from three significant figures to just one. Where it is possible, calculations and measurements should be planned so that loss of significant figures can be avoided.

Mixed Calculations When an expression involves both products and sums, significant figures and precision should be noted for each term or factor as it is calculated, so that correct significant figures and precision for the result are known. The calculation should be performed to as much precision as is available and should be rounded to the correct precision when the calculation is finished. This process is frequently done incorrectly, particularly when calculators or computers provide many decimal places in their result but provide no clue as to how many of those figures are significant.

Significant Figures in Evaluating Functions If $y = f(x)$, then the correct number of significant figures in y depends on the number of significant figures in x and on the behavior of the function f in the neighborhood of x. In general, y should be represented so that all of $f(x)$, $f(x^*)$, and $f(x_*)$ are between y^* and y_*.

EXAMPLES.

sqr (2.0)	sqr (1.95) = 1.39642
	sqr (2.00) = 1.41421
	sqr (2.05) = 1.43178
	so y = 1.4
sin (1°)	sin (0.5) = 0.00872
	sin (1.0) = 0.01745
	sin (1.5) = 0.02617
	so sin (1°) = 0.0
sin (90°)	sin (89.5) = 0.99996
	sin (90.0) = 1.00000
	sin (90.5) = 0.99996
	so sin (90°) = 1.0000

Note that in finding sin (90°), there was a gain in significant figures from two to five and also a gain in precision. This tends to happen when $f'(x)$ is close to zero. On the other hand, precision and significant figures are often lost when $f'(x)$ or $f''(x)$ are large.

Rearrangement of Formulas Often a formula may be rewritten in order to avoid a loss of significant figures.

In using the quadratic formula to find the roots of a polynomial, significant figures may be lost if the $ax^2 + bx + c$ has a root near zero. The quadratic formula may be rearranged as follows:

1. Use the quadratic formula to find the root that is not close to 0. Call this root x_1.
2. Then $x_2 = c/ax_1$.

If $f(x) = \sqrt{x+1} - \sqrt{x}$, then loss of significant figures occurs if x is large. This can be eliminated by ‘‘rationalizing the numerator’’ as follows:

$$\frac{(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{\sqrt{x+1} + \sqrt{x}} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

and this has no loss of significant figures.

There is an almost unlimited number of ‘‘tricks’’ for rearranging formulas to avoid loss of significant figures, but many of these are very similar to the tricks used in calculus to evaluate limits.

GEOMETRY, AREAS, AND VOLUMES

Geometrical Theorems

Right Triangles $a^2 + b^2 = c^2$. (See Fig. 2.1.2.) $\angle A + \angle B = 90°$. $p^2 = mn$. $a^2 = mc$. $b^2 = nc$.

Oblique Triangles Sum of angles = 180°. An exterior angle = sum of the two opposite interior angles (Fig. 2.1.2).

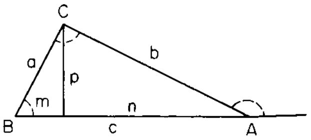


Fig. 2.1.2 Right triangle.

The medians, joining each vertex with the middle point of the opposite side, meet in the center of gravity G (Fig. 2.1.3), which trisects each median.

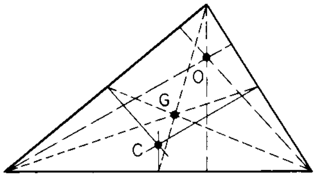


Fig. 2.1.3 Triangle showing medians and center of gravity.

The altitudes meet in a point called the **orthocenter,** O.

The perpendiculars erected at the midpoints of the sides meet in a point C, the center of the circumscribed circle. (In any triangle G, O, and C lie in line, and G is two-thirds of the way from O to C.)

The bisectors of the angles meet in the center of the inscribed circle (Fig. 2.1.4).

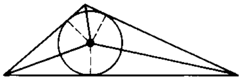


Fig. 2.1.4 Triangle showing bisectors of angles.

The largest side of a triangle is opposite the largest angle; it is less than the sum of the other two sides.

Similar Figures Any two similar figures, in a plane or in space, can be placed in ‘‘perspective,’’ i.e., so that straight lines joining corresponding points of the two figures will pass through a common point (Fig. 2.1.5). That is, of two similar figures, one is merely an enlargement of the other. Assume that each length in one figure is k times the corresponding length in the other; then each area in the first figure is k^2 times the corresponding area in the second, and each volume in the first figure is k^3 times the corresponding volume in the second. If two lines are cut by a set of parallel lines (or parallel planes), the corresponding segments are proportional.

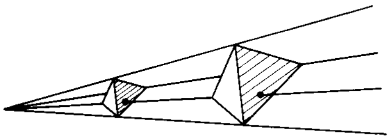


Fig. 2.1.5 Similar figures.

The Circle An angle that is inscribed in a semicircle is a right angle (Fig. 2.1.6).

A tangent is perpendicular to the radius drawn to the point of contact.

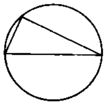


Fig. 2.1.6 Angle inscribed in a semicircle.

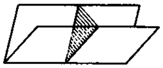


Fig. 2.1.7 Dihedral angle.

Dihedral Angles The dihedral angle between two planes is measured by a plane angle formed by two lines, one in each plane, perpendicular to the edge (Fig. 2.1.7). (For solid angles, see Surfaces and Volumes of Solids.)

In a **tetrahedron,** or triangular pyramid, the four medians, joining each vertex with the center of gravity of the opposite face, meet in a point, the center of gravity of the tetrahedron; this point is ¾ of the way from any vertex to the center of gravity of the opposite face.

The Sphere (See also Surfaces and Volumes of Solids.) If AB is a diameter, any plane perpendicular to AB cuts the sphere in a circle, of which A and B are called the poles. A great circle on the sphere is formed by a plane passing through the center.

Geometrical Constructions

To Bisect a Line AB (Fig. 2.1.8) (1) From A and B as centers, and with equal radii, describe arcs intersecting at P and Q, and draw PQ, which will bisect AB in M. (2) Lay off $AC = BD$ = approximately half of AB, and then bisect CD.

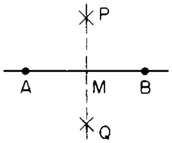


Fig. 2.1.8 Bisectors of a line.

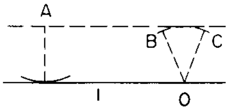


Fig. 2.1.9 Construction of a line parallel to a given line.

To Draw a Parallel to a Given Line l **through a Given Point** A (Fig. 2.1.9) With point A as center draw an arc just touching the line l; with any point O of the line as center, draw an arc BC with the same radius. Then a line through A touching this arc will be the required parallel. Or, use a straightedge and triangle. Or, use a sheet of celluloid with a set of lines parallel to one edge and about ¼ in apart ruled upon it.

To Draw a Perpendicular to a Given Line from a Given Point A **Outside the Line** (Fig. 2.1.10) (1) With A as center, describe an arc cutting the line at R and S, and bisect RS at M. Then M is the foot of the perpendicular. (2) If A is nearly opposite one end of the line, take any point B of the line and bisect AB in O; then with O as center, and OA or OB as radius, draw an arc cutting the line in M. Or, (3) use a straightedge and triangle.

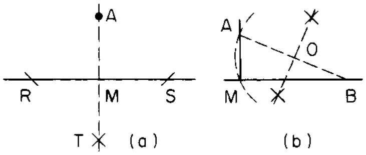


Fig. 2.1.10 Construction of a line perpendicular to a given line from a point not on the line.

To Erect a Perpendicular to a Given Line at a Given Point P (1) Lay off $PR = PS$ (Fig. 2.1.11), and with R and S as centers draw arcs intersecting at A. Then PA is the required perpendicular. (2) If P is near the end of the line, take any convenient point O (Fig. 2.1.12) above the line as center, and with radius OP draw an arc cutting the line at Q. Produce QO to meet the arc at A; then PA is the required perpendicular. (3) Lay off $PB = 4$ units of any scale (Fig. 2.1.13); from P and B as centers lay off $PA = 3$ and $BA = 5$; then APB is a right angle.

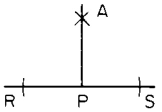


Fig. 2.1.11 Construction of a line perpendicular to a given line from a point on the line.

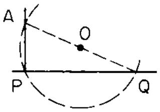


Fig. 2.1.12 Construction of a line perpendicular to a given line from a point on the line.

To Divide a Line *AB* into *n* Equal Parts (Fig. 2.1.14) Through A draw a line AX at any angle, and lay off n equal steps along this line. Connect the last of these divisions with B, and draw parallels through the other divisions. These parallels will divide the given line into n equal parts. A similar method may be used to divide a line into parts which shall be proportional to any given numbers.

To Bisect an Angle *AOB* (Fig. 2.1.15) Lay off $OA = OB$. From A and B as centers, with any convenient radius, draw arcs meeting at M; then OM is the required bisector.

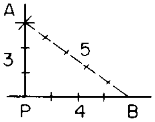


Fig. 2.1.13 Construction of a line perpendicular to a given line from a point on the line.

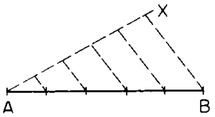


Fig. 2.1.14 Division of a line into equal parts.

To draw the bisector of an angle when the vertex of the angle is not accessible. Parallel to the given lines a, b, and equidistant from them, draw two lines a', b' which intersect; then bisect the angle between a' and b'.

To Inscribe a Hexagon in a Circle (Fig. 2.1.16) Step around the circumference with a chord equal to the radius. Or, use a 60° triangle.

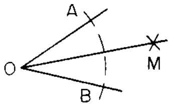


Fig. 2.1.15 Bisection of an angle.

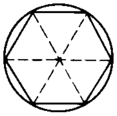


Fig. 2.1.16 Hexagon inscribed in a circle.

To Circumscribe a Hexagon about a Circle (Fig. 2.1.17) Draw a chord AB equal to the radius. Bisect the arc AB at T. Draw the tangent at T (parallel to AB), meeting OA and OB at P and Q. Then draw a circle with radius OP or OQ and inscribe in it a hexagon, one side being PQ.

To Construct a Polygon of *n* Sides, One Side *AB* Being Given (Fig. 2.1.18) With A as center and AB as radius, draw a semicircle, and divide it into n parts, of which $n - 2$ parts (counting from B) are to be used. Draw rays from A through these points of division, and complete the construction as in the figure (in which $n = 7$). Note that the center of the polygon must lie in the perpendicular bisector of each side.

To Draw a Tangent to a Circle from an external point A (Fig. 2.1.19) Bisect AC in M; with M as center and radius MC, draw arc cutting circle in P; then P is the required point of tangency.

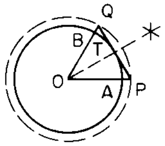


Fig. 2.1.17 Hexagon circumscribed about a circle.

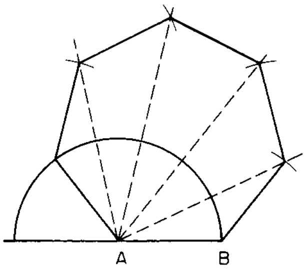


Fig. 2.1.18 Construction of a polygon with a given side.

To Draw a Common Tangent to Two Given Circles (Fig. 2.1.20) Let C and c be centers and R and r the radii ($R > r$). From C as center, draw two concentric circles with radii $R + r$ and $R - r$; draw tangents to these circles from c; then draw parallels to these lines at distance r. These parallels will be the required common tangents.

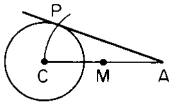


Fig. 2.1.19 Construction of a tangent to a circle.

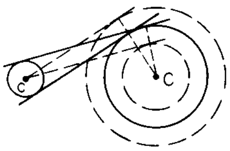


Fig. 2.1.20 Construction of a tangent common to two circles.

To Draw a Circle through Three Given Points A, B, C, or to find the center of a given circular arc (Fig. 2.1.21) Draw the perpendicular bisectors of AB and BC; these will meet at the center, O.

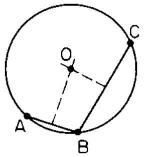


Fig. 2.1.21 Construction of a circle passing through three given points.

To Draw a Circle through Two Given Points *A*, *B*, and Touching a Given Circle (Fig. 2.1.22) Draw any circle through A and B, cutting the given circle at C and D. Let AB and CD meet at E, and let ET be tangent from E to the circle just drawn. With E as center, and radius ET, draw an arc cutting the given circle at P and Q. Either P or Q is the required point of contact. (Two solutions.)

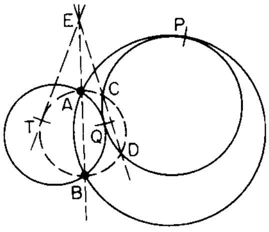


Fig. 2.1.22 Construction of a circle through two given points and touching a given circle.

To Draw a Circle through One Given Point, *A*, and Touching Two Given Circles (Fig. 2.1.23) Let S be a center of similitude for the two given circles, i.e., the point of intersection of two external (or internal) common tangents. Through S draw any line cutting one circle at two points, the nearer of which shall be called P, and the other at two points, the more remote of which shall be called Q. Through A, P, Q draw a circle cutting SA at B. Then draw a circle through A and B and touching one of the given circles (see preceding construction). This circle will touch the other given circle also. (Four solutions.)

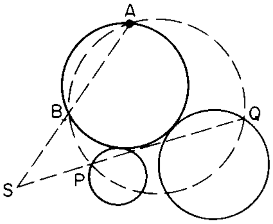


Fig. 2.1.23 Construction of a circle through a given point and touching two given circles.

To Draw an Annulus Which Shall Contain a Given Number of Equal Contiguous Circles (Fig. 2.1.24) (An annulus is a ring-shaped area enclosed between two concentric circles.) Let $R + r$ and $R - r$ be the inner and outer radii of the annulus, r being the radius of each of the n circles. Then the required relation between these quantities is given by $r = R \sin (180°/n)$, or $r = (R + r) [\sin (180°/n)]/[1 + \sin (180°/n)]$.

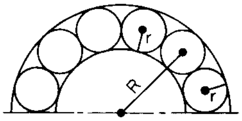


Fig. 2.1.24 Construction of an annulus containing a given number of contiguous circles.

Lengths and Areas of Plane Figures

Right Triangle (Fig. 2.1.25) $a^2 + b^2 = c^2$. Area $= \frac{1}{2}ab = \frac{1}{2}a^2 \cot A = \frac{1}{2}b^2 \tan A = \frac{1}{4}c^2 \sin 2A$.

Equilateral Triangle (Fig. 2.1.26) Area $= \frac{1}{4}a^2\sqrt{3} = 0.43301a^2$.

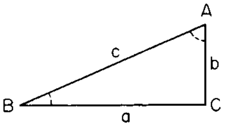


Fig. 2.1.25 Right triangle.

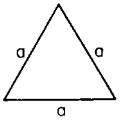


Fig. 2.1.26 Equilateral triangle.

Any Triangle (Fig. 2.1.27)

$$s = \tfrac{1}{2}(a + b + c),\ t = \tfrac{1}{2}(m1 + m2 + m3)$$

$$r = \sqrt{(s - a)(s - b)(s - c)/s} = \text{radius inscribed circle}$$

$$R = \tfrac{1}{2}a/\sin A = \tfrac{1}{2}b/\sin B = \tfrac{1}{2}c/\sin C = \text{radius circumscribed circle}$$

$$\text{Area} = \tfrac{1}{2}\text{ base} \times \text{altitude} = \tfrac{1}{2}ah = \tfrac{1}{2}ab \sin C = rs = abc/4R = \pm\tfrac{1}{2}\{(x_1 y_2 - x_2 y_1) + (x_2 y_3 - x_3 y_2) + (x_3 y_1 - x_1 y_3)\},$$
where (x_1, y_1), (x_2, y_2), (x_3, y_3) are coordinates of vertices.

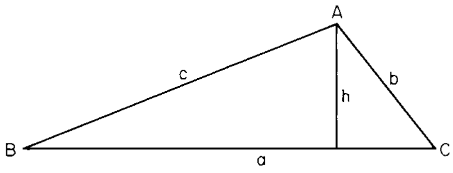


Fig. 2.1.27 Triangle.

Rectangle (Fig. 2.1.28) Area $= ab = \frac{1}{2}D^2 \sin u$, where u = angle between diagonals D, D.

Rhombus (Fig. 2.1.29) Area $= a^2 \sin C = \frac{1}{2}D_1D_2$, where C = angle between two adjacent sides; D_1, D_2 = diagonals.

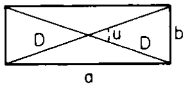


Fig. 2.1.28 Rectangle.

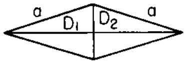


Fig. 2.1.29 Rhombus.

Parallelogram (Fig. 2.1.30) Area $= bh = ab \sin C = \frac{1}{2}D_1D_2 \sin u$, where u = angle between diagonals D_1 and D_2.

Trapezoid (Fig. 2.1.31) Area $= \frac{1}{2}(a + b)\,h$ where bases a and b are parallel.

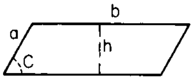


Fig. 2.1.30 Parallelogram.

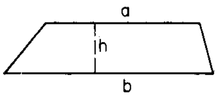


Fig. 2.1.31 Trapezoid.

Any Quadrilateral (Fig. 2.1.32) Area $= \frac{1}{2}D_1D_2 \sin u$.

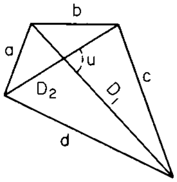


Fig. 2.1.32 Quadrilateral.

Regular Polygons n = number of sides; $v = 360°/n$ = angle subtended at center by one side; a = length of one side $= 2R \sin (v/2) = 2r \tan (v/2)$; R = radius of circumscribed circle $= 0.5\, a \csc (v/2) = r \sec (v/2)$; r = radius of inscribed circle $= R \cos (v/2) = 0.5 \cot (v/2)$; area $= 0.25\, a^2n \cot (v/2) = 0.5\, R^2n \sin (v) = r^2n \tan (v/2)$. Areas of regular polygons are tabulated in Table 1.1.3.

Circle Area $= \pi r^2 = \frac{1}{2}Cr = \frac{1}{4}Cd = \frac{1}{4}\pi d^2 = 0.785398d^2$, where r = radius, d = diameter, C = circumference $= 2\,\pi r = \pi d$.

Annulus (Fig. 2.1.33) Area $= \pi(R^2 - r^2) = \pi(D^2 - d^2)/4 = 2\pi R'b$, where R' = mean radius $= \frac{1}{2}(R + r)$, and $b = R - r$.

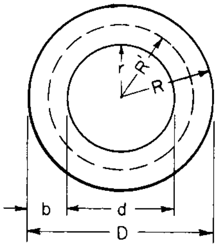


Fig. 2.1.33 Annulus.

Sector (Fig. 2.1.34) Area $= \frac{1}{2}rs = \pi r^2A/360° = \frac{1}{2}\, r^2 \text{ rad} A$, where rad A = radian measure of angle A, and s = length of arc $= r$ rad A.

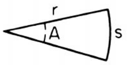


Fig. 2.1.34 Sector.

Segment (Fig. 2.1.35) Area $= \frac{1}{2}r^2(\text{rad } A - \sin A) = \frac{1}{2}[r(s - c) + ch]$, where rad A radian measure of angle A. For small arcs, $s = \frac{1}{3}(8c' - c)$, where c' = chord of half of the arc (Huygens' approximation). Areas of segments are tabulated in Tables 1.1.1 and 1.1.2.

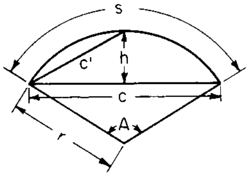


Fig. 2.1.35 Segment.

Ribbon bounded by two parallel curves (Fig. 2.1.36). If a straight line AB moves so that it is always perpendicular to the path traced by its middle point G, then the area of the ribbon or strip thus generated is equal to the length of AB times the length of the path traced by G. (It is assumed that the radius of curvature of G's path is never less than $\frac{1}{2}AB$, so that successive positions of generating line will not intersect.)

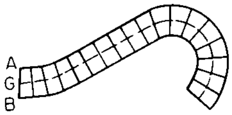


Fig. 2.1.36 Ribbon.

Ellipse (Fig. 2.1.37) Area of ellipse $= \pi ab$. Area of shaded segment $= xy + ab \sin^{-1}(x/a)$. Length of perimeter of ellipse $= \pi(a + b)K$, where $K = (1 + \frac{1}{4}m^2 + \frac{1}{64}m^4 + \frac{1}{256}m^6 + \ldots)$, $m = (a - b)/(a + b)$.

For m =	0.1	0.2	0.3	0.4	0.5
K =	1.002	1.010	1.023	1.040	1.064
For m =	0.6	0.7	0.8	0.9	1.0
K =	1.092	1.127	1.168	1.216	1.273

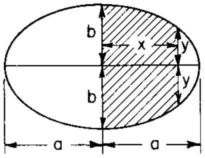


Fig. 2.1.37 Ellipse.

Hyperbola (Fig. 2.1.38) In any hyperbola, shaded area $A = ab \ln[(x/a) + (y/b)]$. In an equilateral hyperbola ($a = b$), area $A = a^2 \sinh^{-1}(y/a) = a^2 \cosh^{-1}(x/a)$. Here x and y are coordinates of point P.

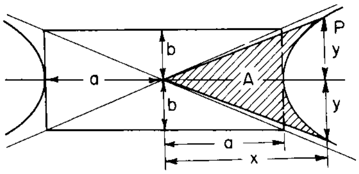


Fig. 2.1.38 Hyperbola.

For lengths and areas of **other curves** see Analytical Geometry.

Surfaces and Volumes of Solids

Regular Prism (Fig. 2.1.39) Volume $= \frac{1}{2}nrah = Bh$. Lateral area $= nah = Ph$. Here n = number of sides; B = area of base; P = perimeter of base.

Right Circular Cylinder (Fig. 2.1.40) Volume $= \pi r^2 h = Bh$. Lateral area $= 2\pi rh = Ph$. Here B = area of base; P = perimeter of base.

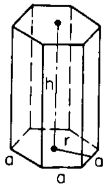


Fig. 2.1.39 Regular prism.

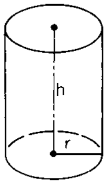


Fig. 2.1.40 Right circular cylinder.

Truncated Right Circular Cylinder (Fig. 2.1.41) Volume $= \pi r^2 h = Bh$. Lateral area $= 2\pi rh = Ph$. Here h = mean height $= \frac{1}{2}(h_1 + h_2)$; B = area of base; P = perimeter of base.

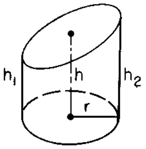


Fig. 2.1.41 Truncated right circular cylinder.

Any Prism or Cylinder (Fig. 2.1.42) Volume $= Bh = Nl$. Lateral area $= Ql$. Here l = length of an element or lateral edge; B = area of base; N = area of normal section; Q = perimeter of normal section.

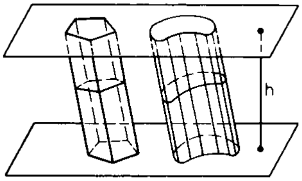


Fig. 2.1.42 Any prism or cylinder.

Special Ungula of a Right Cylinder (Fig. 2.1.43) Volume $= \frac{2}{3}r^2 H$. Lateral area $= 2rH$. r = radius. (Upper surface is a semiellipse.)

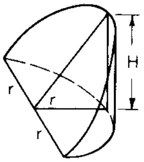


Fig. 2.1.43 Special ungula of a right circular cylinder.

Any Ungula of a right circular cylinder (Figs. 2.1.44 and 2.1.45) Volume $= H(\frac{2}{3}a^3 \pm cB)/(r \pm c) = H[a(r^2 - \frac{1}{3}a^2) \pm r^2 c \text{ rad } u]/(r \pm c)$. Lateral area $= H(2ra \pm cs)/(r \pm c) = 2rH(a \pm c \text{ rad } u)/$

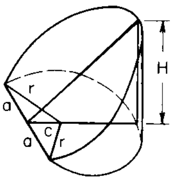


Fig. 2.1.44 Ungula of a right circular cylinder.

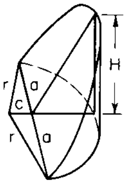


Fig. 2.1.45 Ungula of a right circular cylinder.

($r \pm c$). If base is greater (less) than a semicircle, use + (−) sign. r = radius of base; B = area of base; s = arc of base; u = half the angle subtended by arc s at center; rad u = radian measure of angle u.

Regular Pyramid (Fig. 2.1.46) Volume = ⅓ altitude × area of base = $\frac{1}{6}hran$. Lateral area = ½ slant height × perimeter of base = $\frac{1}{2}san$. Here r = radius of inscribed circle; a = side (of regular polygon); n = number of sides; $s = \sqrt{r^2 + h^2}$. Vertex of pyramid directly above center of base.

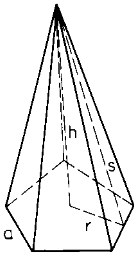


Fig. 2.1.46 Regular pyramid.

Right Circular Cone Volume = $\frac{1}{3}\pi r^2 h$. Lateral area = πrs. Here r = radius of base; h = altitude; s = slant height = $\sqrt{r^2 + h^2}$.

Frustum of Regular Pyramid (Fig. 2.1.47) Volume = $\frac{1}{6}hran[1 + (a'/a) + (a'/a)^2]$. Lateral area = slant height × half sum of perimeters of bases = slant height × perimeter of midsection = $\frac{1}{2}sn(r + r')$. Here r,r' = radii of inscribed circles; $s = \sqrt{(r - r')^2 + h^2}$; a,a' = sides of lower and upper bases; n = number of sides.

Frustum of Right Circular Cone (Fig. 2.1.48) Volume = $\frac{1}{3}\pi r^2 h[1 + (r'/r) + (r'/r)^2] = \frac{1}{3}\pi h(r^2 + rr' + r'^2) = \frac{1}{4}\pi h[(r + r')^2 + \frac{1}{3}(r - r')^2]$. Lateral area = $\pi s(r + r')$; $s = \sqrt{(r - r')^2 + h^2}$.

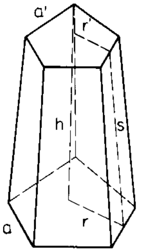


Fig. 2.1.47 Frustum of a regular pyramid.

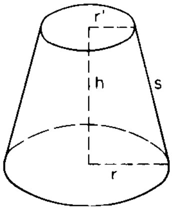


Fig. 2.1.48 Frustum of a right circular cone.

Any Pyramid or Cone Volume = $\frac{1}{3}Bh$. B = area of base; h = perpendicular distance from vertex to plane in which base lies.

Any Pyramidal or Conical Frustum (Fig. 2.1.49) Volume = $\frac{1}{3}h(B + \sqrt{BB'} + B') = \frac{1}{3}hB[1 + (P'/P) + (P'/P)^2]$. Here B, B' = areas of lower and upper bases; P, P' = perimeters of lower and upper bases.

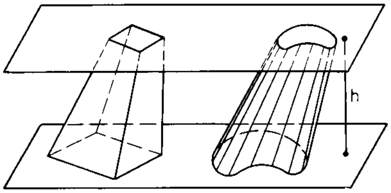


Fig. 2.1.49 Pyramidal frustum and conical frustum.

Sphere Volume = $V = \frac{4}{3}\pi r^3 = 4.188790r^3 = \frac{1}{6}\pi d^3$ = ⅔ volume of circumscribed cylinder. Area = $A = 4\pi r^2$ = four great circles = πd^2 = lateral area of circumscribed cylinder. Here r = radius; $d = 2r$ = diameter = $\sqrt[3]{6V/\pi} = \sqrt{A/\pi}$.

Hollow Sphere or spherical shell. Volume = $\frac{4}{3}\pi(R^3 - r^3) = \frac{1}{6}\pi(D^3 - d^3) = 4\pi R_1^2 t + \frac{1}{3}\pi t^3$. Here R,r = outer and inner radii; D,d = outer and inner diameters; t = thickness = $R - r$; R_1 = mean radius = $\frac{1}{2}(R + r)$.

Any Spherical Segment. Zone (Fig. 2.1.50) Volume = $\frac{1}{6}\pi h(3a^2 + 3a_1^2 + h^2)$. Lateral area (zone) = $2\pi rh$. Here r = radius of sphere. If the inscribed frustum of a cone is removed from the spherical segment, the volume remaining is $\frac{1}{6}\pi hc^2$, where c = slant height of frustum = $\sqrt{h^2 + (a - a_1)^2}$.

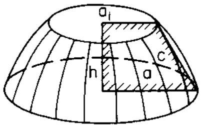


Fig. 2.1.50 Any spherical segment.

Spherical Segment of One Base. Zone (spherical ‘‘cap’’ of Fig. 2.1.51) Volume = $\frac{1}{6}\pi h(3a^2 + h^2) = \frac{1}{3}\pi h^2(3r - h)$. Lateral area (of zone) = $2\pi rh = \pi(a^2 + h^2)$.

NOTE. $a^2 = h(2r - h)$, where r = radius of sphere.

Spherical Sector (Fig. 2.1.51) Volume = $\frac{1}{3}r$ × area of cap = $\frac{2}{3}\pi r^2 h$. Total area = area of cap + area of cone = $2\pi rh + \pi ra$.

NOTE. $a^2 = h(2r - h)$.

Spherical Wedge bounded by two plane semicircles and a **lune** (Fig. 2.1.52). Volume of wedge ÷ volume of sphere = $u/360°$. Area of lune ÷ area of sphere = $u/360°$. u = dihedral angle of the wedge.

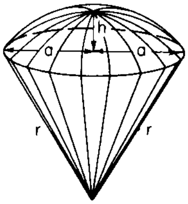


Fig. 2.1.51 Spherical sector.

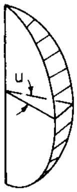


Fig. 2.1.52 Spherical wedge.

Solid Angles Any portion of a spherical surface subtends what is called a **solid angle** at the center of the sphere. If the area of the given portion of spherical surface is equal to the square of the radius, the subtended solid angle is called a **steradian,** and this is commonly taken as the unit. The entire solid angle about the center is called a **steregon,** so that 4π steradians = 1 steregon. A so-called ‘‘solid right angle’’ is the solid angle subtended by a quadrantal (or trirectangular) spherical triangle, and a ‘‘spherical degree’’ (now little used) is a solid angle equal to $\frac{1}{90}$ of a solid right angle. Hence 720 spherical degrees = 1 steregon, or π steradians = 180 spherical degrees. If u = the angle which an element of a cone makes with its axis, then the solid angle of the cone contains $2\pi(1 - \cos u)$ steradians.

Regular Polyhedra A = area of surface; V = volume; a = edge.

Name of solid	Bounded by	A/a^2	V/a^3
Tetrahedron	4 triangles	1.7321	0.1179
Cube	6 squares	6.0000	1.0000
Octahedron	8 triangles	3.4641	0.4714
Dodecahedron	12 pentagons	20.6457	7.6631
Icosahedron	20 triangles	8.6603	2.1917

Ellipsoid (Fig. 2.1.53) Volume $= \tfrac{4}{3}\pi abc$, where a, b, c = semi-axes.

Torus, or Anchor Ring (Fig. 2.1.54) Volume $= 2\pi^2cr^2$. Area $= 4\pi^2cr$.

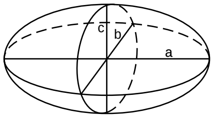


Fig. 2.1.53 Ellipsoid.

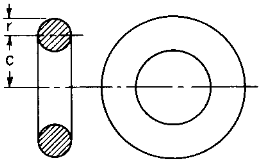


Fig. 2.1.54 Torus.

Volume of a Solid of Revolution (solid generated by rotating an area bounded above by $f(x)$ around the x axis)

$$V = \pi\int_a^b |\, f(x)|^2\, dx$$

Area of a Surface of Revolution

$$A = 2\pi\int_a^b y\sqrt{1 + (dy/dx)^2}\, dx$$

Length of Arc of a Plane Curve $y = f(x)$ between values $x = a$ and $x = b$. $s = \int_a^b \sqrt{1 + (dy/dx)^2}\, dx$. If $x = f(t)$ and $y = g(t)$, for $a < t < b$, then

$$s = \int_a^b \sqrt{(dx/dt)^2 + (dy/dt)^2}\, dt$$

PERMUTATIONS AND COMBINATIONS

The product (1)(2)(3) . . . (n) is written $n!$ and is read ''n factorial.'' By convention, $0! = 1$, and $n!$ is not defined for negative integers.

For large values of n, $n!$ may be approximated by Stirling's formula:

$$n! \approx 2.50663 n^{n+.5}e^{-n}$$

The **binomial coefficient** $C(n, k)$, also written $\binom{n}{k}$, is defined as:

$$C(n, k) = \frac{n!}{k!(n-k)!}$$

$C(n, k)$ is read ''n choose k'' or as ''binomial coefficient n-k.''

Binomial coefficients have the following properties:

1. $C(n, 0) = C(n, n) = 1$
2. $C(n, 1) = C(n, n-1) = n$
3. $C(n+1, k) = C(n, k) + C(n, k-1)$
4. $C(n, \mathrm{k}) = C(n, n-k)$

Binomial coefficients are tabulated in Sec. 1.

Binomial Theorem

If n is a positive integer, then

$$(a+b)^n = \sum_{k=0}^{n} C(n, k)a^kb^{n-k}$$

EXAMPLE. The third term of $(2x+3)^7$ is $C(7, 4)(2x)^{7-4}3^4 = [7!/(4!3!)](2x)^33^4 = (35)(8\,x^3)(81) = 22680x^3$.

Combinations $C(n, k)$ gives the number of ways k distinct objects can be chosen from a set of n elements. This is the number of k-element subsets of a set of n elements.

EXAMPLE. The set of four elements $\{a, b, c, d\}$ has $C\{4, 2\} = 6$ two-element subsets, $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{b, c\}$, $\{b, d\}$, and $\{c, d\}$. (Note that $\{a, c\}$ is the same set as $\{c, a\}$.)

Permutations The number of ways k objects may be arranged from a set of n elements is given by

$$P(n, k) = \frac{n!}{(n-k)!}$$

EXAMPLE. Two elements from the set $\{a, b, c, d\}$ may be arranged in $C(4, 2) = 12$ ways: ab, ac, ad, ba, bc, bd, ca, cb, cd, da, db, and dc. Note that ac is a different arrangement than ca.

Permutations and combinations are examined in detail in most texts on probability and statistics and on discrete mathematics.

If an event can occur in s ways and can fail to occur in f ways, and if all ways are equally likely, then the probability of the event's occurring is $p = s/(s+f)$, and the probability of failure is $q = f/(s+f) = 1 - p$.

The set of all possible outcomes of an experiment is called the **sample space,** denoted S. Let n be the number of outcomes in the sample set. A subset A of the sample space is called an **event**. The number of outcomes in A is s. Therefore $P(A) = s/n$. The probability that A does not occur is $P({\sim}A) = q = 1 - p$.

Always $0 \le p \le 1$ and $P(S) = 1$.

If two events cannot occur simultaneously, then $A \cap B = \varnothing$, and A and B are said to be **mutually exclusive**. Then $P(A \cup B) = P(A) + P(B)$. Otherwise, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Events A and B are **independent** if $P(A \cap B) = P(A)P(B)$.

If E is an event and if $P(E) > 0$, then the probability that A occurs once E has already occurred is called the ''conditional probability of A given E,'' written $P(A|E)$ and defined as

$$P(A|E) = P(A \cap E)/P(E)$$

A and E are independent if $P(A|E) = P(A)$.

If the outcomes in a sample space X are all numbers, then X, together with the probabilities of the outcomes, is called a **random variable**. If x_i is an outcome, then $p_i = P(x_i)$.

The **expected value** of a random variable is

$$E(X) = \Sigma\, e_ip_i$$

The **variance** of X is

$$V(X) = \Sigma[x_i - E(X)]^2p_i$$

The **standard deviation** is

$$S(X) = \sqrt{[V(X)]}$$

The Binomial, or Bernoulli, Distribution If an experiment is repeated n times and the probability of a success on any trial is p, then the probability of k successes among those n trials is

$$f(n, k, p) = C(n, k)p^kq^{n-k}$$

Geometric Distribution If an experiment is repeated until it finally succeeds, let x be the number of failures observed before the first success. Let p be the probability of success on any trial and let $q = 1 - p$. Then

$$P(x = k) = q^k \cdot p$$

Uniform Distribution If the random variable x assumes the values 1, 2, . . . , n, with equal probabilities, then the distribution is uniform, and

$$P(x = k) = \frac{1}{n}$$

Hypergeometric Distribution—Sampling without Replacement If a finite population of N elements contains x successes and if n items are selected randomly without replacement, then the probability that k successes will occur among those n samples is

$$h(x; N, n, k) = \frac{C(k, x)C(N - k, n - x)}{C(N, n)}$$

For large values of N, the hypergeometric distribution approaches the binomial distribution, so

$$h(x; N, n, k) \approx f\left(n, k, \frac{x}{N}\right)$$

Poisson Distribution If the average number of successes which occur in a given fixed time interval is m, then let x be the number of successes observed in that time interval. The probability that $x = k$ is

$$p(k, m) = \frac{e^{-m}m^x}{x!} \qquad \text{where } e = 2.71828\ .\ .\ .$$

Negative Binomial Distribution If repeated independent trials have probability of success p, then let x be the trial number upon which success number n occurs. Then the probability that $x = k$ is

$$b^*(k; n, p) = C(k - 1, n - 1)p^n q^{k-n}$$

The expected values and variances of these distributions are summarized in the following table:

Distribution	$E(X)$	$V(X)$
Uniform	$(n + 1)/2$	$(n^2 - 1)/12$
Binomial	np	npq
Hypergeometric	nk/N	$[nk(N - n)(1 - k/N)]/[N(N - 1)]$
Poisson	m	m
Geometric	q/p	q/p^2
Negative binomial	nq/p	nq/p^2

LINEAR ALGEBRA

Using linear algebra, it is often possible to express in a single equation a set of relations that would otherwise require several equations. Similarly, it is possible to replace many calculations involving several variables with a few calculations involving vectors and matrices. In general, the equations to which the techniques of linear algebra apply must be linear equations; they can involve no polynomial, exponential, or trigonometric terms.

Vectors

A **row vector v** is a list of numbers written in a row, usually enclosed by parentheses.

$$\mathbf{v} = (v_1, v_2, \ .\ .\ .\ .\ , v_n)$$

A **column vector u** is a list of numbers written in a column:

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix}$$

The numbers u_i and v_i may be real or complex, or they may even be variables or functions.

A vector is sometimes called an **ordered *n*-tuple**. In the case where $n = 2$, it may be called an **ordered pair**.

The numbers v_i are called **components** or **coordinates** of the vector **v**. The number n is called the **dimension** of **v**.

Two-dimensional vectors correspond with points in the plane, where v_1 is the x coordinate and v_2 is the y coordinate of the point **v**. Two-dimensional vectors also correspond with complex numbers, where $z = v_1 + iv_2$.

Three-dimensional vectors correspond to points in space, where v_1, v_2, and v_3 are the x, y, and z coordinates of the point, respectively.

Two- and three-dimensional vectors may be thought of as having a direction and a magnitude. See the section ‘‘Analytical Geometry.’’

Two vectors **u** and **v** are equal if:

1. **u** and **v** are the same type (either row or column).
2. **u** and **v** have the same dimension.
3. Corresponding components are equal; that is, $u_i = v_i$ for $i = 1, 2, \ .\ .\ .\ .\ , n$.

Note that the row vectors

$$\mathbf{u} = (1, 2, 3) \qquad \text{and} \qquad \mathbf{v} = (3, 2, 1)$$

are not equal since the components are not in the same order. Also,

$$\mathbf{u} = (1, 2, 3) \qquad \text{and} \qquad \mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

are not equal since **u** is a row vector and **v** is a column vector.

Vector Transpose If **u** is a row vector, then the **transpose** of **u**, written $\mathbf{u}^T$, is the column vector with the same components in the same order as **u**. Similarly, the transpose of a column vector is the row vector with the same components in the same order. Note that $(\mathbf{u}^T)^T = \mathbf{u}$.

Vector Addition If **u** and **v** are vectors of the same type and the same dimension, then the sum of **u** and **v**, written $\mathbf{u} + \mathbf{v}$, is the vector obtained by adding corresponding components. In the case of row vectors,

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \ .\ .\ .\ .\ , u_n + v_n)$$

Scalar Multiplication If a is a number and **u** is a vector, then the **scalar product** $a\mathbf{u}$ is the vector obtained by multiplying each component of **u** by a.

$$a\mathbf{u} = (au_1, au_2, \ .\ .\ .\ .\ , au_n)$$

A number by which a vector is multiplied is called a **scalar**.

The **negative** of vector **u** is written $-\mathbf{u}$, and

$$-\mathbf{u} = -1\mathbf{u}$$

The **zero vector** is the vector with all its components equal to zero.

Arithmetic Properties of Vectors If **u, v,** and **w** are vectors of the same type and dimensions, and if a and b are scalars, then vector addition and scalar multiplication obey the following seven rules, known as the *properties of a vector space:*

1. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$	associative law
2. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$	commutative law
3. $\mathbf{u} + \mathbf{0} = \mathbf{u}$	additive identity
4. $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$	additive inverse
5. $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$	distributive law
6. $(ab)\mathbf{u} = a(b\mathbf{u})$	associative law of multiplication
7. $1\mathbf{u} = \mathbf{u}$	multiplicative identity

Inner Product or Dot Product If **u** and **v** are vectors of the same type and dimension, then their **inner product** or **dot product,** written **uv** or **u** · **v,** is the scalar

$$\mathbf{uv} = u_1v_1 + u_2v_2 + \cdots + u_nv_n$$

Vectors **u** and **v** are **perpendicular** or **orthogonal** if $\mathbf{uv} = 0$.

Magnitude There are two equivalent ways to define the magnitude of a vector **u**, written $|\mathbf{u}|$ or $\|\mathbf{u}\|$.

$$|\mathbf{u}| = \sqrt{(\mathbf{u} \cdot \mathbf{u})}$$

or

$$|\mathbf{u}| = \sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)}$$

Cross Product or Outer Product If **u** and **v** are three-dimensional vectors, then they have a **cross product,** also called **outer product** or **vector product**.

$$\mathbf{u} \times \mathbf{v} = (u_2v_3 - u_3v_2, v_1u_3 - v_3u_1, u_1v_2 - u_2v_1)$$

The cross product $\mathbf{u} \times \mathbf{v}$ is a three-dimensional vector that is perpendicular to both $\mathbf{u}$ and $\mathbf{v}$. The cross product is not commutative. In fact,

$$\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$$

Cross product and inner product have two properties involving trigonometric functions. If θ is the angle between vectors $\mathbf{u}$ and $\mathbf{v}$, then

$$\mathbf{uv} = |\mathbf{u}|\,|\mathbf{v}| \cos\theta \qquad \text{and} \qquad |\mathbf{u} \times \mathbf{v}| = |\mathbf{u}|\,|\mathbf{v}| \sin\theta$$

Matrices

A **matrix** is a rectangular array of numbers. A matrix A with m rows and n columns may be written

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix}$$

The numbers a_{ij} are called the **entries** of the matrix. The first subscript i identifies the row of the entry, and the second subscript j identifies the column.

Matrices are denoted either by capital letters, A, B, etc., or by writing the general entry in parentheses, (a_{ij}).

The number of rows and the number of columns together define the dimensions of the matrix. The matrix A is an $m \times n$ matrix, read ''m by n.''

A row vector may be considered to be a $1 \times n$ matrix, and a column vector may be considered as a $n \times 1$ matrix.

The rows of a matrix are sometimes considered as row vectors, and the columns may be considered as column vectors.

If a matrix has the same number of rows as columns, the matrix is called a **square matrix**.

In a square matrix, the entries a_{ii}, where the row index is the same as the column index, are called the diagonal entries.

If a matrix has all its entries equal to zero, it is called a **zero matrix.**

If a square matrix has all its entries equal to zero except its diagonal entries, it is called a **diagonal matrix.**

The diagonal matrix with all its diagonal entries equal to 1 is called the **identity matrix,** and is denoted I, or $I_{n \times n}$ if it is important to emphasize the dimensions of the matrix. The 2×2 and 3×3 identity matrices are:

$$I_{2 \times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad I_{3 \times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The entries of a square matrix a_{ij} where $i > j$ are said to be below the diagonal. Similarly, those where $i < j$ are said to be above the diagonal.

A square matrix with all entries below (resp. above) the diagonal equal to zero is called *upper-triangular* (resp. *lower-triangular*).

Matrix Addition Matrices A and B may be added only if they have the same dimensions. Then the sum $C = A + B$ is defined by

$$c_{ij} = a_{ij} + b_{ij}$$

That is, corresponding entries of the matrices are added together, just as with vectors. Similarly, matrices may be multiplied by scalars.

Matrix Multiplication Matrices A and B may be multiplied only if the number of columns in A equals the number of rows of B. If A is an $m \times n$ matrix and B is an $n \times p$ matrix, then the product $C = AB$ is an $m \times p$ matrix, defined as follows:

$$\begin{aligned} c_{ij} &= a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \\ &= \sum_{k=1}^{n} a_{ik}b_{kj} \end{aligned}$$

The entry c_{ij} may also be defined as the dot product of row i of A with the transpose of column j of B.

EXAMPLE. $\begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ 7 & 8 \end{pmatrix} =$

$$\begin{pmatrix} 1 \times 3 + 2 \times 7 & 1 \times 4 + 2 \times 8 \\ 5 \times 3 + 6 \times 7 & 5 \times 4 + 6 \times 8 \end{pmatrix} = \begin{pmatrix} 17 & 20 \\ 57 & 68 \end{pmatrix}$$

Matrix multiplication is not commutative. Even if A and B are both square, it is hardly ever true that $AB = BA$. Matrix multiplication does have the following properties:

1. $(AB)C = A(BC)$ associative law
2. $A(B + C) = AB + AC$ } distributive laws
3. $(B + C)A = BA + CA$ }

If A is square, then also

4. $AI = IA = A$ multiplicative identity

If A is square, then powers of A, AA, and AAA are denoted A^2 and A^3, respectively.

The transpose of a matrix A, written A^T, is obtained by writing the rows of A as columns. If A is $m \times n$, then A^T is $n \times m$.

EXAMPLE. $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$

The transpose has the following properties:

1. $(A^T)^T = A$
2. $(A + B)^T = A^T + B^T$
3. $(AB)^T = B^TA^T$

Note that in property 3, the order of multiplication is reversed.

If $A^T = A$, then A is called *symmetric.*

Linear Equations

A linear equation in two variables is of the form

$$a_1x_1 + a_2x_2 = b \qquad \text{or} \qquad a_1x + a_2y = b$$

depending on whether the variables are named x_1 and x_2 or x and y.

In n variables, such an equation has the form

$$a_1x_1 + a_2x_2 + \cdots\ a_nx_n = b$$

Such equations describe lines and planes. Often it is necessary to solve several such equations simultaneously. A set of m linear equations in n variables is called an $m \times n$ system of simultaneous linear equations.

Systems with Two Variables

1 × 2 Systems An equation of the form

$$a_1x + a_2y = b$$

has infinitely many solutions which form a straight line in the xy plane. That line has slope $-a_1/a_2$ and y intercept b/a_2.

2 × 2 Systems A 2×2 system has the form

$$a_{11}x + a_{12}y = b_1 \qquad a_{21}x + a_{22}y = b_2$$

Solutions to such systems do not always exist.

CASE 1. The system has exactly one solution (Fig. 2.1.55*a*). The lines corresponding to the equations intersect at a single point. This occurs whenever the two lines have different slopes, so they are not

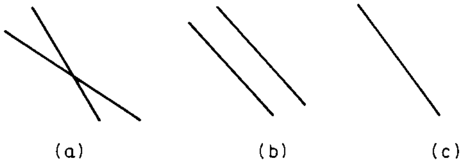


Fig. 2.1.55 Line corresponding to linear equations. (*a*) One solution; (*b*) no solutions; (*c*) infinitely many solutions.

parallel. In this case,

$$\frac{a_{11}}{a_{21}} \neq \frac{a_{12}}{a_{22}} \qquad \text{so} \qquad a_{11}a_{22} - a_{21}a_{12} \neq 0$$

CASE 2. The system has no solutions (Fig. 2.1.55*b*). This occurs whenever the two lines have the same slope and different *y* intercepts, so they are parallel. In this case,

$$\frac{a_{11}}{a_{21}} = \frac{a_{12}}{a_{22}}$$

CASE 3. The system has infinitely many solutions (Fig. 2.1.55*c*). This occurs whenever the two lines coincide. They have the same slope and *y* intercept. In this case,

$$\frac{a_{11}}{a_{21}} = \frac{a_{12}}{a_{22}} = \frac{b_1}{b_2}$$

The value $a_{11}a_{22} - a_{21}a_{12}$ is called the **determinant** of the system. A larger $n \times n$ system also has a determinant (see below). A system has exactly one solution when its determinant is not zero.

3 × 2 Systems Any system with more equations than variables is called **overdetermined**. The only case in which a 3×2 system has exactly one solution is when one of the equations can be derived from the other two.

One basic way to solve such a system is to treat any two equations as a 2×2 system and see if the solution to that subsystem of equations is also a solution to the third equation.

Matrix Form for Systems of Equations The 2×2 system of linear equations

$$a_{11}x_1 + a_{12}x_2 = b_1 \qquad a_{21}x_1 + a_{22}x_2 = b_2$$

may be written as a matrix equation as follows:

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

or as

$$A\mathbf{x} = \mathbf{b}$$

where A is the 2×2 matrix and $\mathbf{x}$ and $\mathbf{b}$ are two-dimensional column vectors. Then, the determinant of A, written $\det A$ or $|A|$, is the same as the determinant of the 2×2 system:

$$\det A = a_{11}a_{22} - a_{21}a_{12}$$

In general, any $m \times n$ system of simultaneous linear equations may be written as

$$A\mathbf{x} = \mathbf{b}$$

where A is an $m \times n$ matrix, $\mathbf{x}$ is an n-dimensional column vector, and $\mathbf{b}$ is an m-dimensional column vector.

An $n \times n$ (square) system of simultaneous linear equations has exactly one solution whenever its determinant is not zero. Then the system and the matrix A are called **nonsingular**. If the determinant is zero, the system is called **singular**.

Elementary Row Operations on a Matrix There are three operations on a matrix which change the matrix:

1. Multiply each entry in row i by a scalar k (not zero).
2. Interchange row i with row j.
3. Add row i to row j.

Similarly, there are three elementary column operations.

The elementary row operations have the following effects on $|A|$:

1. Multiplying a row (or column) by k multiplies $|A|$ by k.
2. Interchanging two rows (or columns) multiplies $|A|$ by -1.
3. Adding one row (or column) to another does not change $|A|$.

Pivoting, or Reducing, a Column The process of changing the ij entry of a matrix to 1 and changing the rest of column j to zero, by using elementary row operations, is known as **reducing** column j or as **pivoting** on the ij entry. Combining pivoting, the properties of the elementary row operations, and the fact:

$$|I_{n\times n}| = 1$$

provides a technique for finding the determinant of $n \times n$ matrices.

EXAMPLE. Find $|A|$ where

$$A = \begin{pmatrix} 1 & 2 & -4 \\ 5 & -3 & -7 \\ 3 & -2 & 3 \end{pmatrix}$$

First, pivot on the entry in row 1, column 1, in this case, the 1.

Multiplying row 1 by -5, then adding row 1 to row 2, we first multiply the determinant by -5, then do not change it:

$$-5|A| = \begin{vmatrix} -5 & -10 & 20 \\ 5 & -3 & -7 \\ 3 & -2 & 3 \end{vmatrix} = \begin{vmatrix} -5 & -10 & 20 \\ 0 & -13 & 13 \\ 3 & -2 & 3 \end{vmatrix}$$

Next, multiply row 1 by $\frac{3}{5}$ and add row 1 to row 3:

$$-3|A| = \begin{vmatrix} -3 & -6 & 12 \\ 0 & -13 & 13 \\ 3 & -2 & 3 \end{vmatrix} = \begin{vmatrix} -3 & -6 & 12 \\ 0 & -13 & 13 \\ 0 & -8 & 15 \end{vmatrix}$$

Next, divide row 1 by -3:

$$|A| = \begin{vmatrix} 1 & 2 & -4 \\ 0 & -13 & 13 \\ 0 & -8 & 15 \end{vmatrix}$$

Next, pivot on the entry in row 2, column 2. Multiplying row 2 by $-\frac{8}{13}$ and then adding row 2 to row 3, we get:

$$-\frac{8}{13}|A| = \begin{vmatrix} 1 & 2 & -4 \\ 0 & 8 & -8 \\ 0 & -8 & 15 \end{vmatrix} = \begin{vmatrix} 1 & 2 & -4 \\ 0 & 8 & -8 \\ 0 & 0 & 7 \end{vmatrix}$$

Next, divide row 2 by $-\frac{8}{13}$.

$$|A| = \begin{vmatrix} 1 & 2 & -4 \\ 0 & -13 & 13 \\ 0 & 0 & 7 \end{vmatrix}$$

The determinant of a triangular matrix is the product of its diagonal elements, in this case -91.

Inverses Whenever $|A|$ is not zero, that is, whenever A is nonsingular, then there is another $n \times n$ matrix, denoted A^{-1}, read ‘‘A inverse’’ with the property

$$AA^{-1} = A^{-1}A = I_{n\times n}$$

Then the $n \times n$ system of equations

$$A\mathbf{x} = \mathbf{b}$$

can be solved by multiplying both sides by A^{-1}, so

$$\mathbf{x} = I_{n\times n}\mathbf{x} = A^{-1}A\mathbf{x} = A^{-1}\mathbf{b}$$

so

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

The matrix A^{-1} may be found as follows:

1. Make a $n \times 2n$ matrix, with the first n columns the matrix A and the last n columns the identity matrix $I_{n\times n}$.
2. Pivot on each of the diagonal entries of this matrix, one after another, using the elementary row operations.
3. After pivoting n times, the matrix will have in the first n columns the identity matrix, and the last n columns will be the matrix A^{-1}.

EXAMPLE. Solve the system

$$\begin{aligned} x_1 + 2x_2 - 4x_3 &= -4 \\ 5x_1 - 3x_2 - 7x_3 &= 6 \\ 3x_1 - 2x_2 + 3x_3 &= 11 \end{aligned}$$

We must invert the matrix

$$A = \begin{pmatrix} 1 & 2 & -4 \\ 5 & -3 & -7 \\ 3 & -2 & 3 \end{pmatrix}$$

This is the same matrix used in the determinant example above. Adjoin the identity matrix to make a 3 × 6 matrix

$$\begin{pmatrix} 1 & 2 & -4 & 1 & 0 & 0 \\ 5 & -3 & -7 & 0 & 1 & 0 \\ 3 & -2 & 3 & 0 & 0 & 1 \end{pmatrix}$$

Perform the elementary row operations in exactly the same order as in the determinant example.

STEP 1. Pivot on row 1, column 1.

$$\begin{pmatrix} 1 & 2 & -4 & 1 & 0 & 0 \\ 0 & -13 & 13 & -5 & 1 & 0 \\ 0 & -8 & 15 & -3 & 0 & 1 \end{pmatrix}$$

STEP 2. Pivot on row 2, column 2.

$$\begin{pmatrix} 1 & 0 & -2 & 3/13 & 2/13 & 0 \\ 0 & 1 & -1 & 5/13 & -1/13 & 0 \\ 0 & 0 & 7 & 1/13 & -8/13 & 1 \end{pmatrix}$$

STEP 3. Pivot on row 3, column 3.

$$\begin{pmatrix} 1 & 0 & 0 & 23/91 & -2/91 & 26/91 \\ 0 & 1 & 0 & 36/91 & -15/91 & 13/91 \\ 0 & 0 & 1 & 1/91 & -8/91 & 13/91 \end{pmatrix}$$

Now, the inverse matrix appears on the right. To solve the equation,

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

so,

$$\mathbf{x} = \begin{pmatrix} 23/91 & -2/91 & 26/91 \\ 36/91 & -15/91 & 13/91 \\ 1/91 & -8/91 & 13/91 \end{pmatrix} \begin{pmatrix} -4 \\ 6 \\ 11 \end{pmatrix}$$

$$= \begin{pmatrix} (-4 \times 23 + 6 \times -2 + 11 \times 26)/91 \\ (-4 \times 36 + 6 \times -15 + 11 \times 13)/91 \\ (-4 \times 1 + 6 \times -8 + 11 \times 13)/91 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

The solution to the system is then

$$x_1 = 2 \qquad x_2 = -1 \qquad x_3 = 1$$

Special Matrices If A is a matrix of complex numbers, then it is possible to take the complex conjugate a_{ij}^* of each entry, a_{ij}. This is called the **conjugate** of A and is denoted A^*.

1. If $a_{ij} = a_{ji}$, then A is **symmetric**.
2. If $a_{ij} = -a_{ji}$, then A is skew or **antisymmetric**.
3. If $A^T = A^{-1}$, then A is **orthogonal**.
4. If $A = A^{-1}$, then A is **involutory**.
5. If $A = A^*$, then A is **hermitian.**
6. If $A = -A^*$,then A is **skew hermitian.**
7. If $A^{-1} = A^*$, then A is **unitary.**

Eigenvalues and Eigenvectors If A is a square matrix and x is a variable, then the matrix $B = A - xI$ is the **characteristic matrix,** or **eigenmatrix,** of A. The determinant $|A - xI|$ is a polynomial of degree n, called the **characteristic polynomial** of A. The roots of this polynomial, $x_1, x_2, \ldots, x_n$, are the **eigenvalues** of A.

Note that some sources define the characteristic matrix as $xI - A$. If n is odd, then this multiplies the characteristic equation by -1, but the eigenvalues are not changed.

EXAMPLE. $A = \begin{vmatrix} -2 & 5 \\ 2 & 1 \end{vmatrix} \qquad B = \begin{vmatrix} -2 - x & 5 \\ 2 & 1 - x \end{vmatrix}$

Then the characteristic polynomial is

$$\begin{aligned} |B| &= (-2 - x)(1 - x) - (2)(5) \\ &= x^2 + x - 2 - 10 \\ &= x^2 + x - 12 \\ &= (x + 4)(x - 3) \end{aligned}$$

The eigenvalues are -4 and $+3$.

A nonzero vector $\mathbf{v}$ satisfying

$$(A - x_i I)\mathbf{v} = \mathbf{0}$$

is called an **eigenvector** of A associated with the eigenvalue x_i. Eigenvectors have the special property

$$A\mathbf{v} = x_i\mathbf{v}$$

Any multiple of an eigenvector is also an eigenvector.

A matrix is nonsingular when none of its eigenvalues are zero.

Rank and Nullity It is possible that the product of a nonzero matrix A and a nonzero vector $\mathbf{v}$ is zero. This cannot happen if A is nonsingular.

The set of all vectors which become zero when multiplied by A is called the **kernel** of A. The **nullity** of A is the dimension of the kernel. It is a measure of how singular a matrix is.

If A is an $m \times n$ matrix, then the **rank** of A is defined as $n -$ nullity. Rank is at most m.

The technique of pivoting is useful in finding the rank of a matrix. The procedure is as follows:

1. Pivot on each diagonal entry in the matrix, starting with a_{11}.
2. If a row becomes all zero, exchange it with other rows to move it to the bottom of the matrix.
3. If a diagonal entry is zero but the row is not all zero, exchange the column containing the entry with a column to the right not containing a zero in that row.

When the procedure can be carried no further, the nullity is the number of rows of zeros in the matrix.

EXAMPLE. Find the rank and nullity of the 3 × 2 matrix:

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \end{pmatrix}$$

Pivoting on row 1, column 1, yields

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 0 & -3 \end{pmatrix}$$

Pivoting on row 2, column 2, yields

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Nullity is therefore 1. Rank is $3 - 1 = 2$.

If the rank of a matrix is n, so that

$$\text{Rank} + \text{nullity} = m$$

the matrix is said to be full rank.

TRIGONOMETRY

Formal Trigonometry

Angles or Rotations An **angle** is generated by the rotation of a ray, as Ox, about a fixed point O in the plane. Every angle has an **initial line** (OA) from which the rotation started (Fig. 2.1.56), and a **terminal line** (OB) where it stopped; and the counterclockwise direction of rotation is taken as positive. Since the rotating ray may revolve as often as desired, angles of any magnitude, positive or negative, may be obtained. Two angles are **congruent** if they may be superimposed so that their initial lines coincide and their terminal lines coincide; i.e., two congruent angles are either equal or differ by some multiple of 360°. Two angles are **complementary** if their sum is 90°; **supplementary** if their sum is 180°.

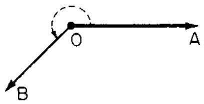


Fig. 2.1.56 Angle.

(The acute angles of a right-angled triangle are complementary.) If the initial line is placed so that it runs horizontally to the right, as in Fig. 2.1.57, then the angle is said to be an angle in the 1st, 2nd, 3rd, or 4th **quadrant** according as the terminal line lies across the region marked I, II, III, or IV.

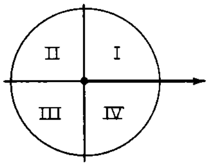


Fig. 2.1.57 Circle showing quadrants.

Units of Angular Measurement

1. *Sexagesimal measure.* (360 degrees = 1 revolution.) Denoted on many calculators by DEG. 1 degree = 1° = 1⁄90 of a right angle. The degree is usually divided into 60 equal parts called minutes (′), and each minute into 60 equal parts called seconds (″); while the second is subdivided decimally. But for many purposes it is more convenient to divide the degree itself into decimal parts, thus avoiding the use of minutes and seconds.

2. *Centesimal measure.* Used chiefly in France. Denoted on calculators by GRAD. (400 grades = 1 revolution.) 1 grade = 1⁄100 of a right angle. The grade is always divided decimally, the following terms being sometimes used: 1 ‘‘centesimal minute’’ = 1⁄100 of a grade; 1 ‘‘centesimal second’’ = 1⁄100 of a centesimal minute. In reading Continental books it is important to notice carefully which system is employed.

3. *Radian, or circular, measure.* (π radians = 180 degrees.) Denoted by RAD. 1 radian = the angle subtended by an arc whose length is equal to the length of the radius. The radian is constantly used in higher mathematics and in mechanics, and is always divided decimally. Many theorems in calculus assume that angles are being measured in radians, not degrees, and are not true without that assumption.

$$\begin{aligned}1\text{ radian} &= 57^\circ.30 - = 57^\circ.2957795131\\ &= 57^\circ17'44''.806247 = 180^\circ/\pi\\ 1^\circ &= 0.01745\ .\ .\ .\ \text{radian} = 0.01745\ 32925\text{ radian}\\ 1' &= 0.00029\ 08882\text{ radian}\\ 1'' &= 0.00000\ 48481\text{ radian}\end{aligned}$$

Table 2.1.2 Signs of the Trigonometric Functions

If x is in quadrant	I	II	III	IV
sin x and csc x are	+	+	−	−
cos x and sec x are	+	−	−	+
tan x and cot x are	+	−	+	−

Definitions of the Trigonometric Functions Let x be any angle whose initial line is OA and terminal line OP (see Fig. 2.1.58). Drop a perpendicular from P on OA or OA produced. In the right triangle OMP, the three sides are MP = ‘‘side opposite’’ O (positive if running upward); OM = ‘‘side adjacent’’ to O (positive if running to the right); OP = ‘‘hypotenuse’’ or ‘‘radius’’ (may always be taken as positive); and the six ratios between these sides are the principal trigonometric functions of the angle x; thus:

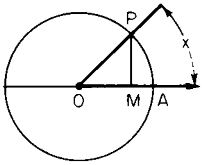


Fig. 2.1.58 Unit circle showing elements used in trigonometric functions.

$$\begin{aligned}\text{sine of } x &= \sin x = \text{opp/hyp} = MP/OP\\ \text{cosine of } x &= \cos x = \text{adj/hyp} = OM/OP\\ \text{tangent of } x &= \tan x = \text{opp/adj} = MP/OM\\ \text{cotangent of } x &= \cot x = \text{adj/opp} = OM/MP\\ \text{secant of } x &= \sec x = \text{hyp/adj} = OP/OM\\ \text{cosecant of } x &= \csc x = \text{hyp/opp} = OP/MP\end{aligned}$$

The last three are best remembered as the reciprocals of the first three:

$$\cot x = 1/\tan x \qquad \sec x = 1/\cos x \qquad \csc x = 1/\sin x$$

Trigonometric functions, the exponential functions, and complex numbers are all related by the **Euler formula:** $e^{ix} = \cos x + i \sin x$, where $i = \sqrt{-1}$. A special case of this $e^{i\pi} = -1$. Note that here x must be measured in radians.

Variations in the functions as x varies from 0 to 360° are shown in Table 2.1.3. The variations in the sine and cosine are best remembered by noting the changes in the lines MP and OM (Fig. 2.1.59) in the ‘‘unit circle’’ (i.e., a circle with radius = OP = 1), as P moves around the circumference.

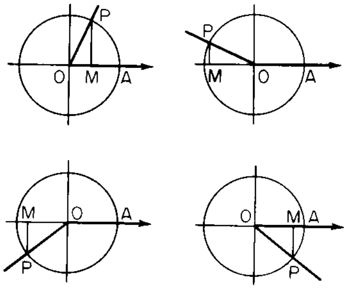


Fig. 2.1.59 Unit circle showing angles in the various quadrants.

Table 2.1.3 Ranges of the Trigonometric Functions

					Values at		
x in DEG x in RAD	0° to 90° (0 to $\pi/2$)	90° to 180° ($\pi/2$ to π)	180° to 270° (π to $3\pi/2$)	270° to 360° ($3\pi/2$ to 2π)	30° ($\pi/6$)	45° ($\pi/4$)	60° ($\pi/3$)
sin x	+0 to +1	+1 to +0	−0 to −1	−1 to −0	½	½√2	½√3
csc x	+∞ to +1	+1 to +∞	−∞ to −1	−1 to −∞	2	√2	⅔√3
cos x	+1 to +0	−0 to −1	−1 to −0	+0 to +1	½√3	½√2	½
sec x	+1 to +∞	−∞ to −1	−1 to −∞	+∞ to +1	⅔√3	√2	2
tan x	+0 to +∞	−∞ to −0	+0 to +∞	−∞ to −0	½√3	1	√3
cot x	+∞ to +0	−0 to −∞	+∞ to +0	−0 to −∞	√3	1	⅓√3

To Find Any Function of a Given Angle (Reduction to the first quadrant.) It is often required to find the functions of any angle x from a table that includes only angles between 0 and 90°. If x is not already between 0 and 360°, first ''reduce to the first revolution'' by simply adding or subtracting the proper multiple of 360° [for any function of (x) = the same function of $(x \pm n \times 360°)$]. Next **reduce to first quadrant** per table below.

If x is between	90° and 180° ($\pi/2$ and π)	180° and 270° (π and $3\pi/2$)	270° and 360° ($3\pi/2$ and 2π)
Subtract	90° from x ($\pi/2$)	180° from x (π)	270° from x ($3\pi/2$)
Then **sin x**	**$= +\cos(x - 90°)$**	**$= -\sin(x - 180°)$**	**$= -\cos(x - 270°)$**
csc x	$= +\sec(x - 90°)$	$= -\csc(x - 180°)$	$= -\sec(x - 270°)$
cos x	**$= -\sin(x - 90°)$**	**$= -\cos(x - 180°)$**	**$= +\sin(x - 270°)$**
sec x	$= -\csc(x - 90°)$	$= -\sec(x - 180°)$	$= +\csc(x - 270°)$
tan x	**$= -\cot(x - 90°)$**	**$= +\tan(x - 180°)$**	**$= -\cot(x - 270°)$**
cot x	$= -\tan(x - 90°)$	$= +\cot(x - 180°)$	$= -\tan(x - 270°)$

The ''reduced angle'' ($x - 90°$, or $x - 180°$, or $x - 270°$) will in each case be an angle between 0 and 90°, whose functions can then be found in the table.

NOTE. The formulas for sine and cosine are best remembered by aid of the unit circle.

To Find the Angle When One of Its Functions Is Given In general, there will be two angles between 0 and 360° corresponding to any given function. The rules showing how to find these angles are tabulated below.

Given	First find an *acute* angle x_0 such that	Then the required angles x_1 and x_2 will be*		
$\sin x = +a$	$\sin x_0 = a$	x_0	and	$180° - x_0$
$\cos x = +a$	$\cos x_0 = a$	x_0	and	$[360° - x_0]$
$\tan x = +a$	$\tan x_0 = a$	x_0	and	$[180° + x_0]$
$\cot x = +a$	$\cot x_0 = a$	x_0	and	$[180° + x_0]$
$\sin x = -a$	$\sin x_0 = a$	$[180° + x_0]$	and	$[360° - x_0]$
$\cos x = -a$	$\cos x_0 = a$	$180° - x_0$	and	$[180° + x_0]$
$\tan x = -a$	$\tan x_0 = a$	$180° - x_0$	and	$[360° - x_0]$
$\cot x = -a$	$\cot x_0 = a$	$180° - x_0$	and	$[360° - x_0]$

* The angles enclosed in brackets lie outside the range 0 to 180 deg and hence cannot occur as angles in a triangle.

Relations Among the Functions of a Single Angle

$$\sin^2 x + \cos^2 x = 1$$

$$\tan x = \frac{\sin x}{\cos x}$$

$$\cot x = \frac{1}{\tan x} = \frac{\cos x}{\sin x}$$

$$1 + \tan^2 x = \sec^2 x = \frac{1}{\cos^2 x}$$

$$1 + \cot^2 x = \csc^2 x = \frac{1}{\sin^2 x}$$

$$\sin x = \sqrt{1 - \cos^2 x} = \frac{\tan x}{\sqrt{1 + \tan^2 x}} = \frac{1}{\sqrt{1 + \cot^2 x}}$$

$$\cos x = \sqrt{1 - \sin^2 x} = \frac{1}{\sqrt{1 + \tan^2 x}} = \frac{\cot x}{\sqrt{1 + \cot^2 x}}$$

Functions of Negative Angles $\sin(-x) = -\sin x$; $\cos(-x) = \cos x$; $\tan(-x) = -\tan x$.

Functions of the Sum and Difference of Two Angles

$\sin(x + y) = \sin x \cos y + \cos x \sin y$
$\cos(x + y) = \cos x \cos y - \sin x \sin y$
$\tan(x + y) = (\tan x + \tan y)/(1 - \tan x \tan y)$
$\cot(x + y) = (\cot x \cot y - 1)/(\cot x + \cot y)$
$\sin(x - y) = \sin x \cos y - \cos x \sin y$
$\cos(x - y) = \cos x \cos y + \sin x \sin y$
$\tan(x - y) = (\tan x - \tan y)/(1 + \tan x \tan y)$
$\cot(x - y) = (\cot x \cot y + 1)/(\cot y - \cot x)$
$\sin x + \sin y = 2 \sin \tfrac{1}{2}(x + y) \cos \tfrac{1}{2}(x - y)$
$\sin x - \sin y = 2 \cos \tfrac{1}{2}(x + y) \sin \tfrac{1}{2}(x - y)$
$\cos x + \cos y = 2 \cos \tfrac{1}{2}(x + y) \cos \tfrac{1}{2}(x - y)$
$\cos x - \cos y = -2 \sin \tfrac{1}{2}(x + y) \sin \tfrac{1}{2}(x - y)$

$$\tan x + \tan y = \frac{\sin(x + y)}{\cos x \cos y}; \quad \cot x + \cot y = \frac{\sin(x + y)}{\sin x \sin y}$$

$$\tan x - \tan y = \frac{\sin(x - y)}{\cos x \cos y}; \quad \cot x - \cot y = \frac{\sin(y - x)}{\sin x \sin y}$$

$\sin^2 x - \sin^2 y = \cos^2 y - \cos^2 x = \sin(x + y)\sin(x - y)$
$\cos^2 x - \sin^2 y = \cos^2 y - \sin^2 x = \cos(x + y)\cos(x - y)$
$\sin(45° + x) = \cos(45° - x)$
$\tan(45° + x) = \cot(45° - x)$
$\sin(45° - x) = \cos(45° + x)$
$\tan(45° - x) = \cot(45° + x)$

In the following transformations, a and b are supposed to be positive, $c = \sqrt{a^2 + b^2}$, A = the positive acute angle for which $\tan A = a/b$, and B = the positive acute angle for which $\tan B = b/a$:

$$a \cos x + b \sin x = c \sin(A + x) = c \cos(B - x)$$
$$a \cos x - b \sin x = c \sin(A - x) = c \cos(B + x)$$

Functions of Multiple Angles and Half Angles

$\sin 2x = 2 \sin x \cos x$; $\sin x = 2 \sin \tfrac{1}{2}x \cos \tfrac{1}{2}x$
$\cos 2x = \cos^2 x - \sin^2 x = 1 - 2 \sin^2 x = 2 \cos^2 x - 1$

$$\tan 2x = \frac{2 \tan x}{1 - \tan^2 x} \qquad \cot 2x = \frac{\cot^2 x - 1}{2 \cot x}$$

$$\sin 3x = 3 \sin x - 4 \sin^3 x; \ \tan 3x = \frac{3 \tan x - \tan^3 x}{1 - 3 \tan^2 x}$$

$\cos 3x = 4 \cos^3 x - 3 \cos x$

$$\sin(nx) = n \sin x \cos^{n-1} x - (n)_3 \sin^3 x \cos^{n-3} x + (n)_5 \sin^5 x \cos^{n-5} x - \cdots$$

$$\cos(nx) = \cos^n x - (n)_2 \sin^2 x \cos^{n-2} x + (n)_4 \sin^4 x \cos^{n-4} x - \cdots$$

where $(n)_2, (n)_3, \ldots$, are the binomial coefficients.

$$\sin \tfrac{1}{2}x = \pm\sqrt{\tfrac{1}{2}(1 - \cos x)}. \ 1 - \cos x = 2 \sin^2 \tfrac{1}{2}x$$

$$\cos \tfrac{1}{2}x = \pm\sqrt{\tfrac{1}{2}(1 + \cos x)}. \ 1 + \cos x = 2 \cos^2 \tfrac{1}{2}x$$

$$\tan \tfrac{1}{2}x = \pm\sqrt{\frac{1 - \cos x}{1 + \cos x}} = \frac{\sin x}{1 + \cos x} = \frac{1 - \cos x}{\sin x}$$

$$\tan\left(\frac{x}{2} + 45°\right) = \pm\sqrt{\frac{1 + \sin x}{1 - \sin x}}$$

Here the + or − sign is to be used according to the sign of the left-hand side of the equation.

Approximations for sin *x*, cos *x*, and tan *x* For small values of x,

x measured in radians, the following approximations hold:

$$\sin x \approx x \qquad \tan x \approx x \qquad \cos x \approx 1 - \frac{x^2}{2}$$

The following actually hold:

$$\sin x < x < \tan x \qquad \cos x < \frac{\sin x}{x} < 1$$

As x approaches 0, $\lim [(\sin x)/x] = 1$.

Inverse Trigonometric Functions The notation $\sin^{-1} x$ (read: anti-sine of x, or inverse sine of x; sometimes written arc sin x) means the principal angle whose sine is x. Similarly for $\cos^{-1} x$, $\tan^{-1} x$, etc. (The principal angle means an angle between -90 and $+90°$ in case of $\sin^{-1}$ and $\tan^{-1}$, and between 0 and 180° in the case of $\cos^{-1}$.)

Solution of Plane Triangles

The ''parts'' of a plane triangle are its three sides a, b, c, and its three angles A, B, C (A being opposite a). Two triangles are **congruent** if all their corresponding parts are equal. Two triangles are **similar** if their corresponding angles are equal, that is, $A_1 = A_2$, $B_1 = B_2$, and $C_1 = C_2$. Similar triangles may differ in scale, but they satisfy $a_1/a_2 = b_1/b_2 = c_1/c_2$.

Two different triangles may have two corresponding sides and the angle opposite one of those sides equal (Fig. 2.1.60), and still not be congruent. This is the **angle-side-side theorem**.

Otherwise, a triangle is uniquely determined by any three of its parts, as long as those parts are not all angles. To ''solve'' a triangle means to find the unknown parts from the known. The fundamental formulas are

$$\text{Law of sines: } \frac{a}{b} = \frac{\sin A}{\sin B}$$

$$\text{Law of cosines: } c^2 = a^2 + b^2 - 2ab \cos C$$

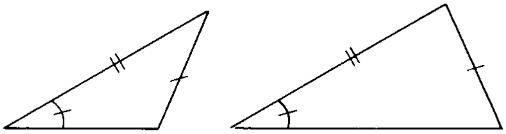


Fig. 2.1.60 Triangles with an angle, an adjacent side, and an opposite side given.

Right Triangles Use the definitions of the trigonometric functions, selecting for each unknown part a relation which connects that unknown with known quantities; then solve the resulting equations. Thus, in Fig. 2.1.61, if $C = 90°$, then $A + B = 90°$, $c^2 = a^2 + b^2$,

$$\sin A = a/c \qquad \cos A = b/c$$
$$\tan A = a/b \qquad \cot A = b/a$$

If A is very small, use $\tan \frac{1}{2}A = \sqrt{(c - b)/(c + b)}$.

Oblique Triangles There are four cases. It is highly desirable in all these cases to draw a sketch of the triangle approximately to scale before commencing the computation, so that any large numerical error may be readily detected.

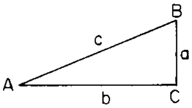


Fig. 2.1.61 Right triangle.

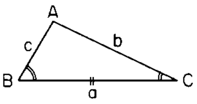


Fig. 2.1.62 Triangle with two angles and the included side given.

Case 1. Given Two Angles (provided their sum is < 180°) and One Side (say a, Fig. 2.1.62). The third angle is known since $A + B + C = 180°$. To find the remaining sides, use

$$b = \frac{a \sin B}{\sin A} \qquad c = \frac{a \sin C}{\sin A}$$

Or, drop a perpendicular from either B or C on the opposite side, and solve by right triangles.

Check: $c \cos B + b \cos C = a$.

Case 2. Given Two Sides (say a and b) and the Included Angle (C); and Suppose $a > b$ (Fig. 2.1.63).

Method 1: Find c from $c^2 = a^2 + b^2 - 2ab \cos C$; then find the smaller angle, B, from $\sin B = (b/c) \sin C$; and finally, find A from $A = 180° - (B + C)$.

Check: $a \cos B + b \cos A = c$.

Method 2: Find $\frac{1}{2}(A - B)$ from the law of tangents:

$$\tan \tfrac{1}{2}(A - B) = [(a - b)/(a + b)] \cot \tfrac{1}{2}C$$

and $\frac{1}{2}(A + B)$ from $\frac{1}{2}(A + B) = 90° - C/2$; hence $A = \frac{1}{2}(A + B) + \frac{1}{2}(A - B)$ and $B = \frac{1}{2}(A + B) - \frac{1}{2}(A - B)$. Then find c from $c = a \sin C/\sin A$ or $c = b \sin C/\sin B$.

Check: $a \cos B + b \cos A = c$.

Method 3: Drop a perpendicular from A to the opposite side, and solve by right triangles.

Case 3. Given the Three Sides (provided the largest is less than the sum of the other two) (Fig. 2.1.64).

Method 1: Find the largest angle A (which may be acute or obtuse) from $\cos A = (b^2 + c^2 - a^2)/2bc$ and then find B and C (which will always be acute) from $\sin B = b \sin A/a$ and $\sin C = c \sin A/a$.

Check: $A + B + C = 180°$.

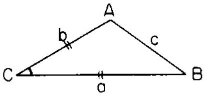


Fig. 2.1.63 Triangle with two sides and the included angle given.

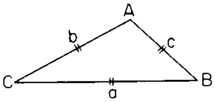


Fig. 2.1.64 Triangle with three sides given.

Method 2: Find A, B, and C from $\tan \frac{1}{2}A = r/(s - a)$, $\tan \frac{1}{2}B = r/(s - b)$, $\tan \frac{1}{2}C = r/(s - c)$, where $s = \frac{1}{2}(a + b + c)$, and $r = \sqrt{(s - a)(s - b)(s - c)/s}$. *Check:* $A + B + C = 180°$.

Method 3: If only one angle, say A, is required, use

$$\sin \tfrac{1}{2}A = \sqrt{(s - b)(s - c)/bc}$$

or

$$\cos \tfrac{1}{2}A = \sqrt{s(s - a)/bc}$$

according as $\frac{1}{2}A$ is nearer 0° or nearer 90°.

Case 4. Given Two Sides (say b and c) and the Angle Opposite One of Them (B). This is the ''ambiguous case'' in which there may be two solutions, or one, or none.

First, try to find $C = c \sin B/b$. If $\sin C > 1$, there is no solution. If $\sin C = 1$, $C = 90°$ and the triangle is a right triangle. If $\sin C < 1$, this determines two angles C, namely, an acute angle C_1, and an obtuse angle $C_2 = 180° - C_1$. Then C_1 will yield a solution when and only when $C_1 + B < 180°$ (see Case 1); and similarly C_2 will yield a solution when and only when $C_2 + B < 180°$ (see Case 1).

Other Properties of Triangles (See also Geometry, Areas, and Volumes.)

Area $= \frac{1}{2}ab \sin C = \sqrt{s(s - a)(s - b)(s - c)} = rs$ where $s = \frac{1}{2}(a + b + c)$, and r = radius of inscribed circle $= \sqrt{(s - a)(s - b)(s - c)/s}$.

Radius of circumscribed circle $= R$, where

$$2R = a/\sin A = b/\sin B = c/\sin C$$

$$r = 4R \sin \frac{A}{2} \sin \frac{B}{2} \sin \frac{C}{2} = \frac{abc}{4Rs}$$

The length of the bisector of the angle C is

$$z = \frac{2\sqrt{abs(s-c)}}{a+b} = \frac{\sqrt{ab[(a+b)^2 - c^2]}}{a+b}$$

The median from C to the middle point of c is $m = \tfrac{1}{2}\sqrt{2(a^2+b^2) - c^2}$.

Hyperbolic Functions

The **hyperbolic sine, hyperbolic cosine,** etc., of any number x, are functions of x which are closely related to the exponential e^x, and which have formal properties very similar to those of the trigonometric functions, sine, cosine, etc. Their definitions and fundamental properties are as follows:

$$\sinh x = \tfrac{1}{2}(e^x - e^{-x})$$
$$\cosh x = \tfrac{1}{2}(e^x + e^{-x})$$
$$\tanh x = \sinh x/\cosh x$$

$$\cosh x + \sinh x = e^x$$
$$\cosh x - \sinh x = e^{-x}$$

$$\operatorname{csch} x = 1/\sinh x$$
$$\operatorname{sech} x = 1/\cosh x$$
$$\coth x = 1/\tanh x$$

$$\cosh^2 x - \sinh^2 x = 1$$
$$1 - \tanh^2 x = \operatorname{sech}^2 x$$
$$1 - \coth^2 x = -\operatorname{csch}^2 x$$

$$\sinh(-x) = -\sinh x$$
$$\cosh(-x) = \cosh x$$
$$\tanh(-x) = -\tanh x$$

$$\sinh(x \pm y) = \sinh x \cosh y \pm \cosh x \sinh y$$
$$\cosh(x \pm y) = \cosh x \cosh y \pm \sinh x \sinh y$$
$$\tanh(x \pm y) = (\tanh x \pm \tanh y)/(1 \pm \tanh x \tanh y)$$

$$\sinh 2x = 2 \sinh x \cosh x$$
$$\cosh 2x = \cosh^2 x + \sinh^2 x$$
$$\tanh 2x = (2 \tanh x)/(1 + \tanh^2 x)$$

$$\sinh \tfrac{1}{2}x = \sqrt{\tfrac{1}{2}(\cosh x - 1)}$$
$$\cosh \tfrac{1}{2}x = \sqrt{\tfrac{1}{2}(\cosh x + 1)}$$
$$\tanh \tfrac{1}{2}x = (\cosh x - 1)/(\sinh x) = (\sinh x)/(\cosh x + 1)$$

The hyperbolic functions are related to the rectangular hyperbola, $x^2 - y^2 = a^2$ (Fig. 2.1.66), in much the same way that the trigonometric functions are related to the circle $x^2 + y^2 = a^2$ (Fig. 2.1.65); the analogy, however, concerns not angles but areas. Thus, in either figure, let A represent the shaded area, and let $u = A/a^2$ (a pure number). Then for the coordinates of the point P we have, in Fig. 2.1.65, $x = a \cos u$, $y = a \sin u$; and in Fig. 2.1.66, $x = a \cosh u$, $y = a \sinh u$.

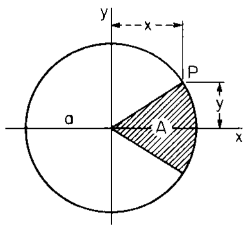


Fig. 2.1.65 Circle.

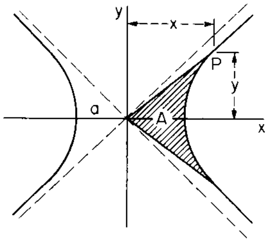


Fig. 2.1.66 Hyperbola.

The **inverse hyperbolic sine** of y, denoted by $\sinh^{-1} y$, is the number whose hyperbolic sine is y; that is, the notation $x = \sinh^{-1} y$ means $\sinh x = y$. Similarly for $\cosh^{-1} y$, $\tanh^{-1} y$, etc. These functions are closely related to the logarithmic function, and are especially valuable in the integral calculus.

$$\sinh^{-1}(y/a) = \ln(y + \sqrt{y^2 + a^2}) - \ln a$$
$$\cosh^{-1}(y/a) = \ln(y + \sqrt{y^2 - a^2}) - \ln a$$
$$\tanh^{-1}\frac{y}{a} = \tfrac{1}{2}\ln\frac{a+y}{a-y}$$
$$\coth^{-1}\frac{y}{a} = \tfrac{1}{2}\ln\frac{y+a}{y-a}$$

ANALYTICAL GEOMETRY

The Point and the Straight Line

Rectangular Coordinates (Fig. 2.1.67) Let $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$. Then, distance

$$P_1P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

slope of $P_1P_2 = m = \tan u = (y_2 - y_1)/(x_2 - x_1)$; coordinates of midpoint are $x = \tfrac{1}{2}(x_1 + x_2)$, $y = \tfrac{1}{2}(y_1 + y_2)$; coordinates of point $1/n$th of the way from P_1 to P_2 are $x = x_1 + (1/n)(x_2 - x_1)$, $y = y_1 + (1/n)(y_2 - y_1)$.

Let m_1, m_2 be the slopes of two lines; then, if the lines are parallel, $m_1 = m_2$; if the lines are perpendicular to each other, $m_1 = -1/m_2$.

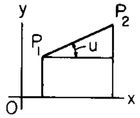


Fig. 2.1.67 Graph of straight line.

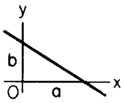


Fig. 2.1.68 Graph of straight line showing intercepts.

Equations of a Straight Line

1. *Intercept form* (Fig. 2.1.68). $x/a + y/b = 1$. (a, b = intercepts of the line on the axes.)

2. *Slope form* (Fig. 2.1.69). $y = mx + b$. ($m = \tan u =$ slope; $b =$ intercept on the y axis.)

3. *Normal form* (Fig. 2.1.70). $x \cos v + y \sin v = p$. (p = perpendicular from origin to line; v = angle from the x axis to p.)

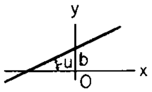


Fig. 2.1.69 Graph of straight line showing slope and vertical intercept.

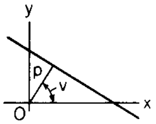


Fig. 2.1.70 Graph of straight line showing perpendicular line from origin.

4. *Parallel-intercept form* (Fig. 2.1.71). $\dfrac{y-b}{c-b} = \dfrac{x}{k}$. ($b$, c = intercepts on two parallels at distance k apart.)

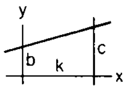


Fig. 2.1.71 Graph of straight line showing intercepts on parallel lines.

5. *General form*. $Ax + By + C = 0$. [Here $a = -C/A$, $b = -C/B$, $m = -A/B$, $\cos v = A/R$, $\sin v = B/R$, $p = -C/R$, where $R = \pm\sqrt{A^2 + B^2}$ (sign to be so chosen that p is positive).]

6. *Line through* (x_1, y_1) *with slope m*. $y - y_1 = m(x - x_1)$.

7. *Line through* (x_1, y_1) *and* (x_2, y_2). $y - y_1 = \dfrac{y_2 - y_1}{x_2 - x_1}(x - x_1)$.

8. *Line parallel to x axis*. $y = a$; to y axis: $x = b$.

Angles and Distances If u = angle from the line with slope m_1 to the line with slope m_2, then

$$\tan u = \frac{m_2 - m_1}{1 + m_2 m_1}$$

If parallel, $m_1 = m_2$.
If perpendicular, $m_1 m_2 = -1$.

If u = angle between the lines $Ax + By + C = 0$ and $A'x + B'y + C' = 0$, then

$$\cos u = \frac{AA' + BB'}{\pm\sqrt{(A^2 + B^2)(A'^2 + B'^2)}}$$

If parallel, $A/A' = B/B'$.
If perpendicular, $AA' + BB' = 0$.

The equation of a line through (x_1, y_1) and meeting a given line $y = mx + b$ at an angle u, is

$$y - y_1 = \frac{m + \tan u}{1 - m \tan u}(x - x_1)$$

The distance from (x_0, y_0) to the line $Ax + By + C = 0$ is

$$D = \left| \frac{Ax_0 + By_0 + C}{\sqrt{A^2 + B^2}} \right|$$

where the vertical bars mean ‘‘the absolute value of.’’

The distance from (x_0, y_0) to a line which passes through (x_1, y_1) and makes an angle u with the x axis is

$$D = (x_0 - x_1) \sin u - (y_0 - y_1) \cos u$$

Polar Coordinates (Fig. 2.1.72) Let (x, y) be the rectangular and (r, θ) the polar coordinates of a given point P. Then $x = r \cos \theta$; $y = r \sin \theta$; $x^2 + y^2 = r^2$.

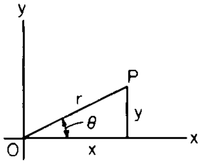


Fig. 2.1.72 Polar coordinates.

Transformation of Coordinates If origin is moved to point (x_0, y_0), the new axes being parallel to the old, $x = x_0 + x'$, $y = y_0 + y'$.

If axes are turned through the angle u, without change of origin,

$$x = x' \cos u - y' \sin u \qquad y = x' \sin u + y' \cos u$$

The Circle

The **equation of a circle** with center (a, b) and radius r is

$$(x - a)^2 + (y - b)^2 = r^2$$

If center is at the origin, the equation becomes $x^2 + y^2 = r^2$. If circle goes through the origin and center is on the x axis at point $(r, 0)$, equation becomes $x^2 + y^2 = 2rx$. The **general equation** of a circle is

$$x^2 + y^2 + Dx + Ey + F = 0$$

It has center at $(-D/2, -E/2)$, and radius $= \sqrt{(D/2)^2 + (E/2)^2 - F}$ (which may be real, null, or imaginary).

Equations of Circle in Parametric Form It is sometimes convenient to express the coordinates x and y of the moving point P (Fig. 2.1.73) in terms of an auxiliary variable, called a **parameter**. Thus, if the parameter be taken as the angle u from the x axis to the radius vector OP, then the equations of the circle in parametric form will be $x = a \cos u$; $y = a \sin u$. For every value of the parameter u, there corresponds a point (x, y) on the circle. The ordinary equation $x^2 + y^2 = a^2$ can be obtained from the parametric equations by eliminating u.

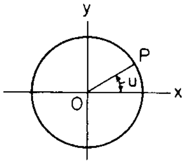


Fig. 2.1.73 Parameters of a circle.

The Parabola

The **parabola** is the locus of a point which moves so that its distance from a fixed line (called the **directrix**) is always equal to its distance from a fixed point F (called the **focus**). See Fig. 2.1.74. The point halfway from focus to directrix is the **vertex,** O. The line through the focus, perpendicular to the directrix, is the **principal axis**. The breadth of the curve at the focus is called the **latus rectum,** or **parameter,** $= 2p$, where p is the distance from focus to directrix.

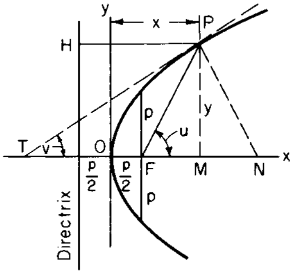


Fig. 2.1.74 Graph of parabola.

NOTE. Any section of a right circular cone made by a plane parallel to a tangent plane of the cone will be a parabola.

Equation of parabola, principal axis along the x axis, origin at vertex (Fig. 2.1.74): $y^2 = 2px$.

Polar equation of parabola, referred to F as origin and Fx as axis (Fig. 2.1.75): $r = p/(1 - \cos \theta)$.

Equation of parabola with principal axis parallel to y axis: $y = ax^2 + bx + c$. This may be rewritten, using a technique called **completing the square:**

$$\begin{aligned} y &= a\left[x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right] + c - \frac{b^2}{4a} \\ &= a\left[x + \frac{b}{2a}\right]^2 + c - \frac{b^2}{4a} \end{aligned}$$

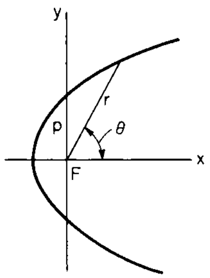


Fig. 2.1.75 Polar plot of parabola.

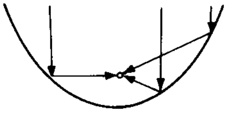


Fig. 2.1.76 Vertical parabola showing rays passing through the focus.

Then: vertex is the point $[-b/2a, c - b^2/4a]$; latus rectum is $p = 1/2a$; and focus is the point $[-b/2a, c - b^2/4a + 1/4a]$.

A parabola has the special property that lines parallel to its principal axis, when reflected off the inside ‘‘surface’’ of the parabola, will all pass through the focus (Fig. 2.1.76). This property makes parabolas useful in designing mirrors and antennas.

The Ellipse

The **ellipse** (as shown in Fig. 2.1.77), has two **foci,** F and F', and two **directrices,** DH and $D'H'$. If P is any point on the curve, $PF + PF'$ is constant, $= 2a$; and PF/PH (or PF'/PH') is also constant, $= e$, where e is the **eccentricity** ($e < 1$). Either of these properties may be taken as the definition of the curve. The relations between e and the semiaxes a and b are as shown in Fig. 2.1.78. Thus, $b^2 = a^2(1 - e^2)$, $ae = \sqrt{a^2 - b^2}$, $e^2 = 1 - (b/a)^2$. The **semilatus rectum** $= p = a(1 - e^2) = b^2/a$. Note that b is always less than a, except in the special case of the circle, in which $b = a$ and $e = 0$.

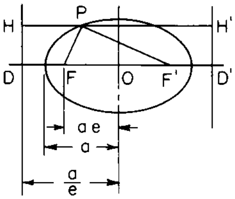


Fig. 2.1.77 Ellipse.

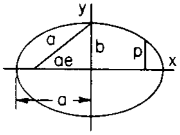


Fig. 2.1.78 Ellipse showing semiaxes.

Any section of a right circular cone made by a plane which cuts all the elements of one nappe of the cone will be an ellipse; if the plane is perpendicular to the axis of the cone, the ellipse becomes a circle.

Equation of ellipse, center at origin:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad \text{or} \qquad y = \pm\frac{b}{a}\sqrt{a^2 - x^2}$$

If $P = (x, y)$ is any point of the curve, $PF = a + ex$, $PF' = a - ex$.

Equations of the ellipse in parametric form: $x = a \cos u$, $y = b \sin u$, where u is the eccentric angle of the point $P = (x, y)$. See Fig. 2.1.81.

Polar equation, focus as origin, axes as in Fig. 2.1.79. $r = p/(1 - e \cos \theta)$.

Equation of the tangent at (x_1, y_1): $b^2x_1x + a^2y_1y = a^2b^2$.

The line $y = mx + k$ will be a tangent if $k = \pm\sqrt{a^2m^2 + b^2}$.

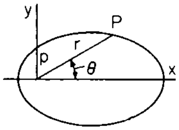


Fig. 2.1.79 Ellipse in polar form.

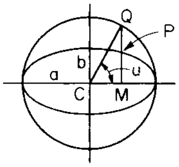


Fig. 2.1.80 Ellipse as a flattened circle.

Ellipse as a Flattened Circle, Eccentric Angle If the ordinates in a circle are diminished in a constant ratio, the resulting points will lie on an ellipse (Fig. 2.1.80). If Q traces the circle with uniform velocity, the corresponding point P will trace the ellipse, with varying velocity. The angle u in the figure is called the eccentric angle of the point P.

A consequence of this property is that if a circle is drawn with its horizontal scale different from its vertical scale, it will appear to be an ellipse. This phenomenon is common in computer graphics.

The **radius of curvature of an ellipse at any point** $P = (x, y)$ is

$$R = a^2b^2(x^2/a^4 + y^2/b^4)^{3/2} = p/\sin^3 v$$

where v is the angle which the tangent at P makes with PF or PF'. At end of major axis, $R = b^2/a = MA$; at end of minor axis, $R = a^2/b = NB$ (see Fig. 2.1.81).

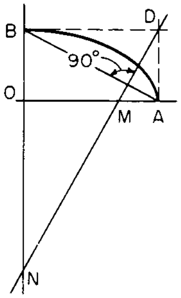


Fig. 2.1.81 Ellipse showing radius of curvature.

The Hyperbola

The **hyperbola** has two **foci,** F and F', at distances $\pm ae$ from the center, and two **directrices,** DH and $D'H'$, at distances $\pm a/e$ from the center (Fig. 2.1.82). If P is any point of the curve, $|PF - PF'|$ is constant, $= 2a$; and PF/PH (or PF'/PH') is also constant, $= e$ (called the **eccentricity**), where $e > 1$. Either of these properties may be taken as the definition of the curve. The curve has two branches which approach more and more nearly two straight lines called the **asymptotes.** Each asymptote makes with the principal axis an angle whose tangent is b/a. The relations between e, a, and b are shown in Fig. 2.1.83: $b^2 = a^2(e^2 - 1)$, $ae = \sqrt{a^2 + b^2}$, $e^2 = 1 + (b/a)^2$. The semilatus rectum, or ordinate at the focus, is $p = a(e^2 - 1) = b^2/a$.

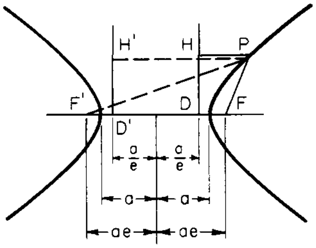


Fig. 2.1.82 Hyperbola.

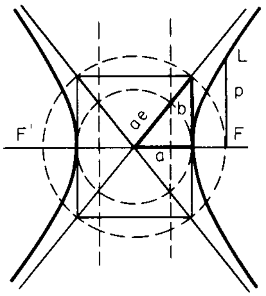


Fig. 2.1.83 Hyperbola showing the asymptotes.

Any section of a right circular cone made by a plane which cuts both nappes of the cone will be a hyperbola.

Equation of the hyperbola, center as origin:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \qquad \text{or} \qquad y = \pm\frac{b}{a}\sqrt{x^2 - a^2}$$

If $P = (x, y)$ is on the right-hand branch, $PF = ex - a$, $PF' = ex + a$. If P is on the left-hand branch, $PF = -ex + a$, $PF' = -ex - a$.

Equations of Hyperbola in Parametric Form (1) $x = a \cosh u$, $y = b \sinh u$. Here u may be interpreted as A/ab, where A is the area shaded in Fig. 2.1.84. (2) $x = a \sec v$, $y = b \tan v$, where v is an auxiliary angle of no special geometric interest.

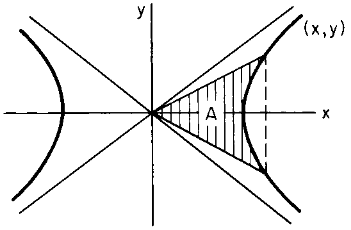


Fig. 2.1.84 Hyperbola showing parametric form.

Polar equation, referred to focus as origin, axes as in Fig. 2.1.85:

$$r = p/(1 - e \cos \theta)$$

Equation of tangent at (x_1, y_1): $b^2x_1x - a^2y_1y = a^2b^2$. The line $y = mx + k$ will be a tangent if $k = \pm\sqrt{a^2m^2 - b^2}$.

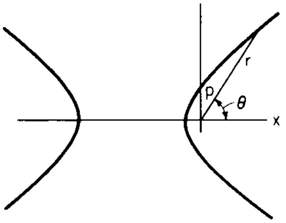


Fig. 2.1.85 Hyperbola in polar form.

The triangle bounded by the asymptotes and a variable tangent is of constant area, $= ab$.

Conjugate hyperbolas are two hyperbolas having the same asymptotes with semiaxes interchanged (Fig. 2.1.86). The equations of the hyperbola conjugate to $x^2/a^2 - y^2/b^2 = 1$ is $x^2/a^2 - y^2/b^2 = -1$.

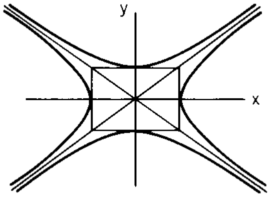


Fig. 2.1.86 Conjugate hyperbolas.

Equilateral Hyperbola $(a = b)$ Equation referred to principal axes (Fig. 2.1.87): $x^2 - y^2 = a^2$.

NOTE. $p = a$ (Fig. 2.1.87). Equation referred to asymptotes as axes (Fig. 2.1.88): $xy = a^2/2$.

Asymptotes are perpendicular. Eccentricity $= \sqrt{2}$. Any diameter is equal in length to its conjugate diameter.

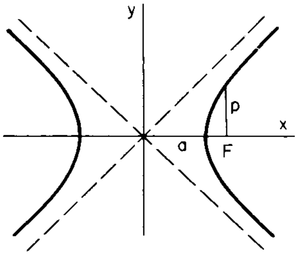


Fig. 2.1.87 Equilateral hyperbola.

The Catenary

The **catenary** is the curve in which a flexible chain or cord of uniform density will hang when supported by the two ends. Let w = weight of the chain per unit length; T = the tension at any point P; and T_h, T_v = the horizontal and vertical components of T. The horizontal component T_h is the same at all points of the curve.

The length $a = T_h/w$ is called the **parameter** of the catenary, or the distance from the lowest point O to the **directrix** DQ (Fig. 2.1.89). When a is very large, the curve is very flat.

The rectangular **equation,** referred to the lowest point as origin, is $y = a\,[\cosh (x/a) - 1]$. In case of very flat arcs (a large), $y = x^2/2a + \cdots$; $s = x + \tfrac{1}{6}x^3/a^2 + \cdots$, approx, so that in such a case the catenary closely resembles a parabola.

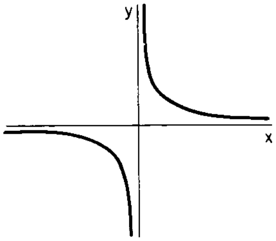


Fig. 2.1.88 Hyperbola with asymptotes as axes.

Calculus properties of the catenary are often discussed in texts on the calculus of variations (Weinstock, "Calculus of Variations," Dover; Ewing, "Calculus of Variations with Applications," Dover).

Problems on the Catenary (Fig. 2.1.89) When any two of the four quantities, x, y, s, T/w are known, the remaining two, and also the parameter a, can be found, using the following:

$$\begin{array}{ll}
a = x/z & s = a \sinh z \\
T = wa \cosh z & y/x = (\cosh z - 1)/z \\
s/x = (\sinh z)/z & wx/T = z \cosh z
\end{array}$$

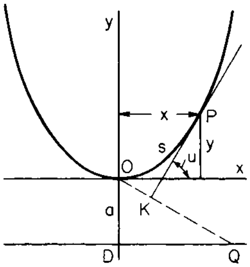


Fig. 2.1.89 Catenary.

NOTE. If $wx/T < 0.6627$, then there are two values of z, one less than 1.2, and one greater. If $wx/T > 0.6627$, then the problem has no solution.

Given the Length $2L$ of a Chain Supported at Two Points A and B Not in the Same Level, to Find a (See Fig. 2.1.90; b and c are supposed known.) Let $(\sqrt{L^2 - b^2})/c = s/x$; use $s/x = \sinh z/z$ to find z. Then $a = c/z$.

NOTE. The coordinates of the midpoint M of AB (see Fig. 2.1.90) are $x_0 = a \tanh^{-1} (b/L)$, $y_0 = (L/\tanh z) - a$, so that the position of the lowest point is determined.

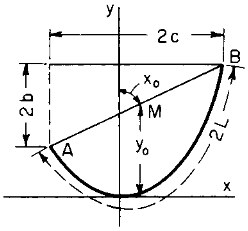


Fig. 2.1.90 Catenary with ends at unequal levels.

Other Useful Curves

The **cycloid** is traced by a point on the circumference of a circle which rolls without slipping along a straight line. Equations of cycloid, in parametric form (axes as in Fig. 2.1.91): $x = a(\text{rad}\ u - \sin u)$, $y = a(1 - \cos u)$, where a is the radius of the rolling circle, and rad u is the radian measure of the angle u through which it has rolled. The **radius of curvature** at any point P is $PC = 4a \sin(u/2) = 2\sqrt{2ay}$.

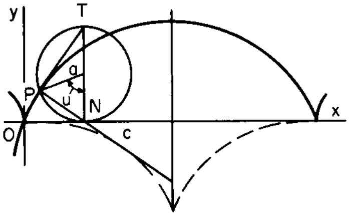


Fig. 2.1.91 Cycloid.

The **trochoid** is a more general curve, traced by any point on a radius of the rolling circle, at distance b from the center (Fig. 2.1.92). It is a prolate trochoid if $b < a$, and a curtate or looped trochoid if $b > a$. The equations in either case are $x = a\ \text{rad}\ u - b \sin u$, $y = a - b \cos u$.

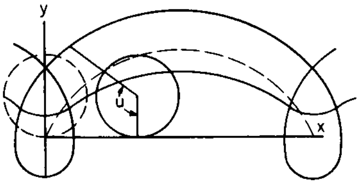


Fig. 2.1.92 Trochoid.

The **epicycloid** (or **hypocycloid**) is a curve generated by a point on the circumference of a circle of radius a which rolls without slipping on the outside (or inside) of a fixed circle of radius c (Fig. 2.1.93 and Fig. 2.1.94). For the **equations,** put $b = a$ in the equations of the epi- or hypotrochoid, below.

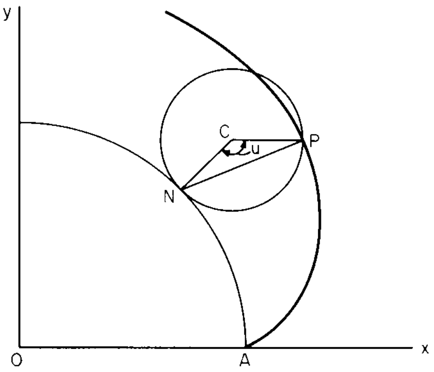


Fig. 2.1.93 Epicycloid.

Radius of curvature at any point P is

$$R = \frac{4a(c \pm a)}{c \pm 2a} \times \sin \tfrac{1}{2}u$$

At A, $R = 0$; at D, $R = \dfrac{4a(c \pm a)}{c \pm 2a}$.

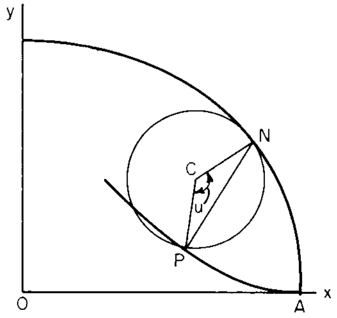


Fig. 2.1.94 Hypocycloid.

Special Cases If $a = \tfrac{1}{2}c$, the hypocycloid becomes a straight line, diameter of the fixed circle (Fig. 2.1.95). In this case the hypotrochoid traced by any point rigidly connected with the rolling circle (not necessarily on the circumference) will be an ellipse. If $a = \tfrac{1}{4}c$, the curve generated will be the four-cusped hypocycloid, or **astroid** (Fig. 2.1.96), whose equation is $x^{2/3} + y^{2/3} = c^{2/3}$. If $a = c$, the epicycloid is the **cardioid,** whose equation in polar coordinates (axes as in Fig. 2.1.97) is $r = 2c(1 + \cos \theta)$. Length of cardioid $= 16c$.

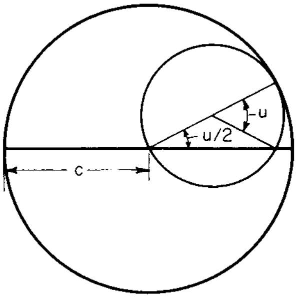


Fig. 2.1.95 Hypocycloid is straight line when the radius of inside circle is half that of the outside circle.

The **epitrochoid** (or **hypotrochoid**) is a curve traced by any point rigidly attached to a circle of radius a, at distance b from the center, when this

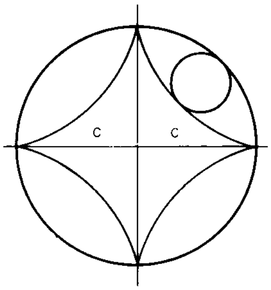


Fig. 2.1.96 Astroid.

circle rolls without slipping on the outside (or inside) of a fixed circle of radius c. The equations are

$$x = (c \pm a)\cos\left(\frac{a}{c}u\right) \pm b\cos\left[\left(1 \pm \frac{a}{c}\right)u\right]$$

$$y = (c \pm a)\sin\left(\frac{a}{c}u\right) - b\sin\left[\left(1 \pm \frac{a}{c}\right)u\right]$$

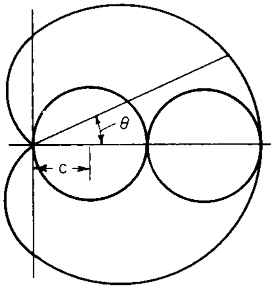


Fig. 2.1.97 Cardioid.

where u = the angle which the moving radius makes with the line of centers; take the upper sign for the epi- and the lower for the hypotrochoid. The curve is called prolate or curtate according as $b < a$ or $b > a$. When $b = a$, the special case of the epi- or hypocycloid arises.

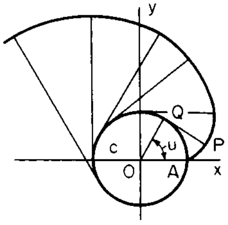


Fig. 2.1.98 Involute of circle.

The **involute of a circle** is the curve traced by the end of a taut string which is unwound from the circumference of a fixed circle, of radius c. If QP is the free portion of the string at any instant (Fig. 2.1.98), QP will be tangent to the circle at Q, and the length of QP = length of arc QA; hence the construction of the curve. The equations of the curve in parametric form (axes as in figure) are $x = c(\cos u + \text{rad } u \sin u)$, $y = c(\sin u - \text{rad } u \cos u)$, where rad u is the radian measure of the angle u which OQ makes with the x axis. Length of arc $AP = \frac{1}{2}c(\text{rad } u)^2$; radius of curvature at P is QP. Polar equations, in terms of parameter v(= angle POQ), are $r = c \sec v$, rad $\theta = \tan v - \text{rad } v$. Here, $r = OP$, and rad θ = radian measure of angle, AOP (Fig. 2.1.98).

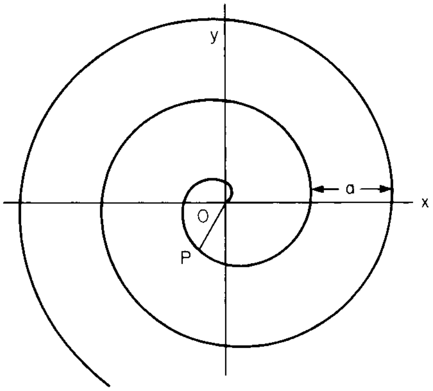


Fig. 2.1.99 Spiral of Archimedes.

The **spiral of Archimedes** (Fig. 2.1.99) is traced by a point P which, starting from O, moves with uniform velocity along a ray OP, while the ray itself revolves with uniform angular velocity about O. Polar equation: $r = k \text{ rad } \theta$, or $r = a(\theta°/360°)$. Here $a = 2\pi k$ = the distance measured along a radius, from each coil to the next.

The radius of curvature at P is $R = (k^2 + r^2)^{3/2}/(2k^2 + r^2)$.

The **logarithmic spiral** (Fig. 2.1.100) is a curve which cuts the radii from O at a constant angle v, whose cotangent is m. Polar equation: $r = ae^{m\text{rad}\theta}$. Here a is the value of r when $\theta = 0$. For large negative values of θ, the curve winds around O as an asymptotic point. If PT and PN are the tangent and normal at P, the line TON being perpendicular to OP (not shown in figure), then $ON = rm$, and $PN = r\sqrt{1 + m^2} = r/\sin v$. Radius of curvature at P is PN.

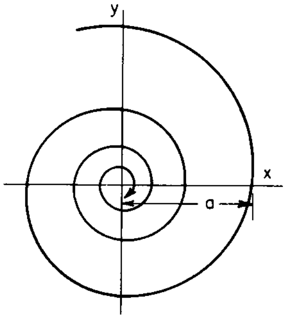


Fig. 2.1.100 Logarithmic spiral.

The **tractrix**, or Schiele's antifriction curve (Fig. 2.1.101), is a curve such that the portion PT of the tangent between the point of contact and the x axis is constant $= a$. Its equation is

$$x = \pm a\left[\cosh^{-1}\frac{a}{y} - \sqrt{1 - \left(\frac{y}{a}\right)^2}\right]$$

or, in parametric form, $x = \pm a(t - \tanh t)$, $y = a/\cosh t$. The x axis is an asymptote of the curve. Length of arc $BP = a \log_e (a/y)$.

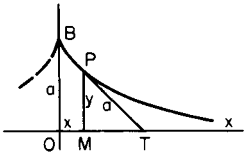


Fig. 2.1.101 Tractrix.

The tractrix describes the path taken by an object being pulled by a string moving along the x axis, where the initial position of the object is B and the opposite end of the string begins at O.

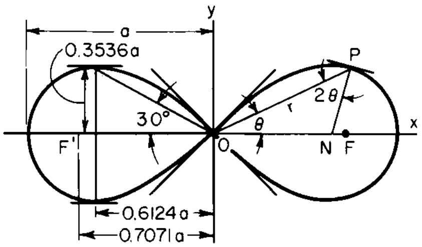


Fig. 2.1.102 Lemniscate.

The **lemniscate** (Fig. 2.1.102) is the locus of a point P the product of whose distances from two fixed points F, F' is constant, equal to $\frac{1}{2}a^2$. The distance $FF' = a\sqrt{2}$. Polar equation is $r = a\sqrt{\cos 2\theta}$. Angle between OP and the normal at P is 2θ. The two branches of the curve cross at right angles at O. Maximum y occurs when $\theta = 30°$ and $r = a/\sqrt{2}$, and is equal to $\frac{1}{4}a\sqrt{2}$. Area of one loop $= a^2/2$.

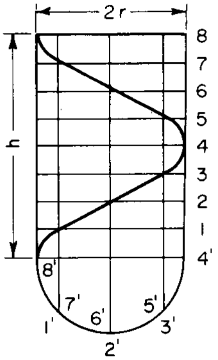


Fig. 2.1.103 Helix.

The **helix** (Fig. 2.1.103) is the curve of a screw thread on a cylinder of radius r. The curve crosses the elements of the cylinder at a constant angle, v. The pitch, h, is the distance between two coils of the helix, measured along an element of the cylinder; hence $h = 2\pi r \tan v$. Length of one coil $= \sqrt{(2\pi r)^2 + h^2} = 2\pi r/\cos v$. If the cylinder is rolled out on a plane, the development of the helix will be a straight line, with slope equal to $\tan v$.

DIFFERENTIAL AND INTEGRAL CALCULUS

Derivatives and Differentials

Derivatives and Differentials A **function** of a single variable x may be denoted by $f(x)$, $F(x)$, etc. The value of the function when x has the value x_0 is then denoted by $f(x_0)$, $F(x_0)$, etc. The **derivative** of a function $y = f(x)$ may be denoted by $f'(x)$, or by dy/dx. The value of the derivative at a given point $x = x_0$ is the **rate of change** of the function at that point; or, if the function is represented by a curve in the usual way (Fig. 2.1.104), the value of the derivative at any point shows the **slope of the curve** (i.e., the slope of the tangent to the curve) at that point (positive if the tangent points upward, and negative if it points downward, moving to the right).

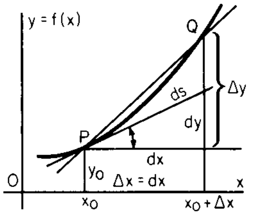


Fig. 2.1.104 Curve showing tangent and derivatives.

The **increment** Δy (read: ‘‘delta y’’) in y is the change produced in y by increasing x from x_0 to $x_0 + \Delta x$; i.e., $\Delta y = f(x_0 + \Delta x) - f(x_0)$. The **differential,** dy, of y is the value which Δy would have if the curve coincided with its tangent. (The differential, dx, of x is the same as Δx when x is the independent variable.) Note that the derivative depends only on the value of x_0, while Δy and dy depend not only on x_0 but on the value of Δx as well. The ratio $\Delta y/\Delta x$ represents the secant slope, and dy/dx the slope of tangent (see Fig. 2.1.104). If Δx is made to approach zero, the secant approaches the tangent as a limiting position, so that the derivative is

$$f'(x) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \left[\frac{\Delta y}{\Delta x} \right] = \lim_{\Delta x \to 0} \left[\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \right]$$

Also, $dy = f'(x)\, dx$.

The symbol ‘‘lim’’ in connection with $\Delta x \to 0$ means ‘‘the limit, as Δx approaches 0, of . . . .’’ (A constant c is said to be the **limit** of a variable u if, whenever any quantity m has been assigned, there is a stage in the variation process beyond which $|c - u|$ is always less than m; or, briefly, c is the limit of u if the difference between c and u can be made to become and remain as small as we please.)

To find the derivative of a given function at a given point: (1) If the function is given only by a curve, measure graphically the slope of the tangent at the point in question; (2) if the function is given by a mathematical expression, use the following rules for differentiation. These rules give, directly, the differential, dy, in terms of dx; to find the derivative, dy/dx, divide through by dx.

Rules for Differentiation (Here $u, v, w, \ldots$ represent any functions of a variable x, or may themselves be independent variables. a is a constant which does not change in value in the same discussion; $e = 2.71828$.)

1. $d(a + u) = du$
2. $d(au) = a\,du$
3. $d(u + v + w + \cdots) = du + dv + dw + \cdots$
4. $d(uv) = u\,dv + v\,du$
5. $d(uvw \ldots) = (uvw \ldots)\left(\frac{du}{u} + \frac{dv}{v} + \frac{dw}{w} + \cdots\right)$
6. $d\,\frac{u}{v} = \frac{v\,du - u\,dv}{v^2}$
7. $d(u^m) = mu^{m-1}\,du$. Thus, $d(u^2) = 2u\,du$; $d(u^3) = 3u^2\,du$; etc.
8. $d\sqrt{u} = \frac{du}{2\sqrt{u}}$
9. $d\left(\frac{1}{u}\right) = -\frac{du}{u^2}$
10. $d(e^u) = e^u\,du$
11. $d(a^u) = (\ln a)a^u\,du$
12. $d \ln u = \frac{du}{u}$
13. $d \log_{10} u = \log_{10} e\,\frac{du}{u} = (0.4343\ldots)\,\frac{du}{u}$
14. $d \sin u = \cos u\,du$
15. $d \csc u = -\cot u \csc u\,du$
16. $d \cos u = -\sin u\,du$
17. $d \sec u = \tan u \sec u\,du$
18. $d \tan u = \sec^2 u\,du$
19. $d \cot u = -\csc^2 u\,du$
20. $d \sin^{-1} u = \frac{du}{\sqrt{1 - u^2}}$
21. $d \csc^{-1} u = -\frac{du}{u\sqrt{u^2 - 1}}$
22. $d \cos^{-1} u = \frac{du}{\sqrt{1 - u^2}}$
23. $d \sec^{-1} u = \frac{du}{u\sqrt{u^2 - 1}}$
24. $d \tan^{-1} u = \frac{du}{1 + u^2}$
25. $d \cot^{-1} u = -\frac{du}{1 + u^2}$
26. $d \ln \sin u = \cot u\,du$
27. $d \ln \tan u = \frac{2\,du}{\sin 2u}$
28. $d \ln \cos u = -\tan u\,du$
29. $d \ln \cot u = -\frac{2\,du}{\sin 2u}$
30. $d \sinh u = \cosh u\,du$
31. $d\,\mathrm{csch}\,u = -\mathrm{csch}\,u \coth u\,du$
32. $d \cosh u = \sinh u\,du$
33. $d\,\mathrm{sech}\,u = -\mathrm{sech}\,u \tanh u\,du$

34. $d \tanh u = \operatorname{sech}^2 u\, du$
35. $d \coth u = -\operatorname{csch}^2 u\, du$
36. $d \sinh^{-1} u = \dfrac{du}{\sqrt{u^2+1}}$
37. $d \operatorname{csch}^{-1} u = -\dfrac{du}{u\sqrt{u^2+1}}$
38. $d \cosh^{-1} u = \dfrac{du}{\sqrt{u^2-1}}$
39. $d \operatorname{sech}^{-1} u = -\dfrac{du}{u\sqrt{1-u^2}}$
40. $d \tanh^{-1} u = \dfrac{du}{1-u^2}$
41. $d \coth^{-1} u = \dfrac{du}{1-u^2}$
42. $d(u^v) = (u^{v-1})(u \ln u\, dv + v\, du)$

Derivatives of Higher Orders The derivative of the derivative is called the second derivative; the derivative of this, the third derivative; and so on. If $y = f(x)$,

$$f'(x) = D_x y = \frac{dy}{dx}$$

$$f''(x) = D_x^2 y = \frac{d^2y}{dx^2}$$

$$f'''(x) = D_x^3 y = \frac{d^3y}{dx^3} \quad \text{etc.}$$

NOTE. If the notation d^2y/dx^2 is used, this must not be treated as a fraction, like dy/dx, but as an inseparable symbol, made up of a symbol of operation d^2/dx^2, and an operand y.

The geometric meaning of the second derivative is this: if the original function $y = f(x)$ is represented by a curve in the usual way, then at any point where $f''(x)$ is *positive*, the curve is *concave upward*, and at any point where $f''(x)$ is *negative*, the curve is *concave downward* (Fig. 2.1.105). When $f''(x) = 0$, the curve usually has a **point of inflection**.

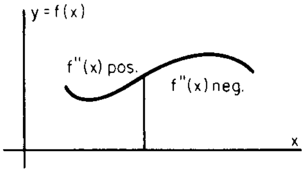


Fig. 2.1.105 Curve showing concavity.

Functions of two or more variables may be denoted by $f(x, y, \ldots)$, $F(x, y, \ldots)$, etc. The derivative of such a function $u = f(x, y, \ldots)$ formed on the assumption that x is the only variable ($y, \ldots$ being regarded for the moment as constants) is called the **partial derivative of** u **with respect to** x, and is denoted by $f_x(x, y)$ or $D_x u$, or $d_x u/dx$, or $\partial u/\partial x$. Similarly, the partial derivative of u with respect to y is $f_y(x, y)$ or $D_y u$, or $d_y u/dy$, or $\partial u/\partial y$.

NOTE. In the third notation, $d_x u$ denotes the differential of u formed on the assumption that x is the only variable. If the fourth notation, $\partial u/\partial x$, is used, this must not be treated as a fraction like du/dx; the $\partial/\partial x$ is a symbol of operation, operating on u, and the ''∂x'' must not be separated.

Partial derivatives of the second order are denoted by f_{xx}, f_{xy}, f_{yy}, or by Du, $D_x(D_y u)$, $D_y^2 u$, or by $\partial^2u/\partial x^2$, $\partial^2u/\partial x\, \partial y$, $\partial^2u/\partial y^2$, the last symbols being ''inseparable.'' Similarly for higher derivatives. Note that $f_{xy} = f_{yx}$.

If increments Δx, Δy (or dx, dy) are assigned to the independent variables x, y, the increment, Δu, produced in $u = f(x, y)$ is

$$\Delta u = f(x + \Delta x, y + \Delta y) - f(x, y)$$

while the **differential,** du, i.e., the value which Δu would have if the partial derivatives of u with respect to x and y were constant, is given by

$$du = (f_x) \cdot dx + (f_y) \cdot dy$$

Here the coefficients of dx and dy are the values of the partial derivatives of u at the point in question.

If x and y are functions of a third variable t, then the equation

$$\frac{du}{dt} = (f_x)\frac{dx}{dt} + (f_y)\frac{dy}{dt}$$

expresses the rate of change of u with respect to t, in terms of the separate rate of change of x and y with respect to t.

Implicit Functions If $f(x, y) = 0$, either of the variables x and y is said to be an implicit function of the other. To find dy/dx, either (1) solve for y in terms of x, and then find dy/dx directly; or (2) differentiate the equation through as it stands, remembering that both x and y are variables, and then divide by dx; or (3) use the formula $dy/dx = -(f_x/f_y)$, where f_x and f_y are the partial derivatives of $f(x, y)$ at the point in question.

Maxima and Minima

A **function of one variable,** as $y = f(x)$, is said to have a **maximum** at a point $x = x_0$, if at that point the slope of the curve is zero and the concavity downward (see Fig. 2.1.106); a sufficient condition for a maximum is $f'(x_0) = 0$ and $f''(x_0)$ negative. Similarly, $f(x)$ has a **minimum** if the slope is zero and the concavity upward; a sufficient condition for a minimum is $f'(x_0) = 0$ and $f''(x_0)$ positive. If $f''(x_0) = 0$ and $f'''(x_0) \neq 0$, the point x_0 will be a **point of inflection.** If $f'(x_0) = 0$ and $f''(x_0) = 0$ and $f'''(x_0) = 0$, the point x_0 will be a maximum if $f''''(x_0) < 0$, and a minimum if $f''''(x_0) > 0$. It is usually sufficient, however, in any practical case, to find the values of x which make $f'(x) = 0$, and then decide, from a general knowledge of the curve or the sign of $f'(x)$ to the right and left of x_0, which of these values (if any) give maxima or minima, without investigating the higher derivatives.

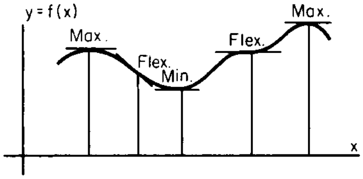


Fig. 2.1.106 Curve showing maxima and minima.

A **function of two variables,** as $u = f(x, y)$, will have a **maximum** at a point (x_0, y_0) if at that point $f_x = 0$, $f_y = 0$, and $f_{xx} < 0$, $f_{yy} < 0$; and a **minimum** if at that point $f_x = 0$, $f_y = 0$, and $f_{xx} > 0$, $f_{yy} > 0$; provided, in each case, $(f_{xx})(f_{yy}) - (f_{xy})^2$ is positive. If $f_x = 0$ and $f_y = 0$, and f_{xx} and f_{yy} have opposite signs, the point (x_0, y_0) will be a ''saddle point'' of the surface representing the function.

Indeterminate Forms

In the following paragraphs, $f(x)$, $g(x)$ denote functions which approach 0; $F(x)$, $G(x)$ functions which increase indefinitely; and $U(x)$ a function which approaches 1, when x approaches a definite quantity a. The problem in each case is to find the limit approached by certain combinations of these functions when x approaches a. The symbol → is to be read ''approaches'' or ''tends to.''

CASE 1. ''0/0.'' To find the limit of $f(x)/g(x)$ when $f(x) \to 0$ and $g(x) \to 0$, use the theorem that $\lim\,[f(x)/g(x)] = \lim\,[f'(x)/g'(x)]$, where $f'(x)$ and $g'(x)$ are the derivatives of $f(x)$ and $g(x)$. This second limit may be easier to find than the first. If $f'(x) \to 0$ and $g'(x) \to 0$, apply the same theorem a second time: $\lim [f'(x)/g'(x)] = \lim [f''(x)/g''(x)]$, and so on.

CASE 2. ''∞/∞.'' If $F(x) \to \infty$ and $G(x) \to \infty$, then $\lim [F(x)/G(x)] = \lim [F'(x)/G'(x)]$, precisely as in Case 1.

CASE 3. ''$0 \cdot \infty$.'' To find the limit of $f(x) \cdot F(x)$ when $f(x) \to 0$ and $F(x) \to \infty$, write $\lim [f(x) \cdot F(x)] = \lim\{f(x)/[1/F(x)]\}$ or $= \lim \{F(x)/[1/f(x)]\}$, then proceed as in Case 1 or Case 2.

CASE 4. The limit of combinations ''0^0'' or $[f(x)]^{g(x)}$; ''1^∞'' or $[U(x)]^{F(x)}$; ''∞^0'' or $[F(x)]^{b(x)}$ may be found since their logarithms are limits of the type evaluated in Case 3.

CASE 5. ''$\infty - \infty$.'' If $F(x) \to \infty$ and $G(x) \to \infty$, write

$$\lim [F(x) - G(x)] = \lim \frac{1/G(x) - 1/F(x)}{1/[F(x) \cdot G(x)]}$$

then proceed as in Case 1. Sometimes it is shorter to expand the functions in series. It should be carefully noticed that expressions like 0/0, ∞/∞, etc., do not represent mathematical quantities.

Curvature

The **radius of curvature** R of a plane curve at any point P (Fig. 2.1.107) is the distance, measured along the normal, on the concave side of the curve, to the **center of curvature,** C, this point being the limiting position of the point of intersection of the normals at P and a neighboring point Q, as Q is made to approach P along the curve. If the equation of the curve is $y = f(x)$,

$$R = \frac{ds}{du} = \frac{[1 + (y')^2]^{3/2}}{y''}$$

where $ds = \sqrt{dx^2 + dy^2}$ = the differential of arc, $u = \tan^{-1} [f'(x)]$ = the angle which the tangent at P makes with the x axis, and $y' = f'(x)$ and $y'' = f''(x)$ are the first and second derivatives of $f(x)$ at the point P. Note that $dx = ds \cos u$ and $dy = ds \sin u$. The **curvature,** K, at the point P, is $K = 1/R = du/ds$; i.e., the curvature is the rate at which the angle u is changing with respect to the length of arc s. If the slope of the curve is small, $K \approx f''(x)$.

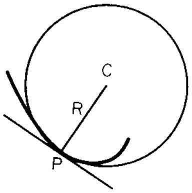


Fig. 2.1.107 Curve showing radius of curvature.

If the equation of the curve in polar coordinates is $r = f(\theta)$, where r = radius vector and θ = polar angle, then

$$R = \frac{[r^2 + (r')^2]^{3/2}}{r^2 - rr'' + 2(r')^2}$$

where $r' = f'(\theta)$ and $r'' = f''(\theta)$.

The **evolute** of a curve is the locus of its centers of curvature. If one curve is the evolute of another, the second is called the **involute** of the first.

Indefinite Integrals

An **integral** of $f(x)\,dx$ is any function whose differential is $f(x)\,dx$, and is denoted by $\int f(x)\,dx$. All the integrals of $f(x)\,dx$ are included in the expression $\int f(x)\,dx + C$, where $\int f(x)\,dx$ is any particular integral, and C is an arbitrary constant. The process of finding (when possible) an integral of a given function consists in recognizing by inspection a function which, when differentiated, will produce the given function; or in transforming the given function into a form in which such recognition is easy. The most common integrable forms are collected in the following brief table; for a more extended list, see Peirce, ''Table of Integrals,'' Ginn, or Dwight, ''Table of Integrals and other Mathematical Data,'' Macmillan, or ''CRC Mathematical Tables.''

GENERAL FORMULAS

1. $\int a\,du = a \int du = au + C$
2. $\int (u + v)\,dx = \int u\,dx + \int v\,dx$
3. $\int u\,dv = uv - \int v\,du$ (integration by parts)
4. $\int f(x)\,dx = \int f[F(y)]F'(y)\,dy,\ x = F(y)$ (change of variables)
5. $\int dy \int f(x, y)\,dx = \int dx \int f(x, y)\,dy$

FUNDAMENTAL INTEGRALS

6. $\int x^n\,dx = \dfrac{x^{n+1}}{n+1} + C$, when $n \neq -1$
7. $\int \dfrac{dx}{x} = \ln x + C = \ln cx$
8. $\int e^x\,dx = e^x + C$
9. $\int \sin x\,dx = -\cos x + C$
10. $\int \cos x\,dx = \sin x + C$
11. $\int \dfrac{dx}{\sin^2 x} = -\cot x + C$
12. $\int \dfrac{dx}{\cos^2 x} = \tan x + C$
13. $\int \dfrac{dx}{\sqrt{1 - x^2}} = \sin^{-1} x + C = -\cos^{-1} x + C$
14. $\int \dfrac{dx}{1 + x^2} = \tan^{-1} x + C = -\cot^{-1} x + C$

RATIONAL FUNCTIONS

15. $\int (a + bx)^n\,dx = \dfrac{(a + bx)^{n+1}}{(n + 1)b} + C$
16. $\int \dfrac{dx}{a + bx} = \dfrac{1}{b} \ln (a + bx) + C = \dfrac{1}{b} \ln c(a + bx)$
17. $\int \dfrac{dx}{x^n} = -\dfrac{1}{(n - 1)x^{n-1}} + C$ except when $n = 1$
18. $\int \dfrac{dx}{(a + bx)^2} = -\dfrac{1}{b(a + bx)} + C$
19. $\int \dfrac{dx}{1 - x^2} = ½ \ln \dfrac{1 + x}{1 - x} + C = \tanh^{-1} x + C$, when $x < 1$
20. $\int \dfrac{dx}{x^2 - 1} = ½ \ln \dfrac{x - 1}{x + 1} + C = -\coth^{-1} x + C$, when $x > 1$

$$\left.\begin{aligned}
&21.\ \int \frac{dx}{a + bx^2} = \frac{1}{\sqrt{ab}} \tan^{-1} \left(\sqrt{\frac{b}{a}}\,x\right) + C \\
&22.\ \int \frac{dx}{a - bx^2} = \frac{1}{2\sqrt{ab}} \ln \frac{\sqrt{ab} + bx}{\sqrt{ab} - bx} + C \\
&\qquad = \frac{1}{\sqrt{ab}} \tanh^{-1} \left(\sqrt{\frac{b}{a}}\,x\right) + C
\end{aligned}\right\} [a > 0, b > 0]$$

23. $$\left.\begin{aligned}\int \frac{dx}{a+2bx+cx^2} &= \\ &\frac{1}{\sqrt{ac-b^2}} \tan^{-1}\frac{b+cx}{\sqrt{ac-b^2}} + C\end{aligned}\right\} [ac-b^2>0]$$

$$\left.\begin{aligned}&= \frac{1}{2\sqrt{b^2-ac}} \ln \frac{\sqrt{b^2-ac}-b-cx}{\sqrt{b^2-ac}+b+cx} + C \\ &= -\frac{1}{\sqrt{b^2-ac}} \tanh^{-1}\frac{b+cx}{\sqrt{b^2-ac}} + C\end{aligned}\right\} [b^2-ac>0]$$

24. $$\int \frac{dx}{a+2bx+cx^2} = -\frac{1}{b+cx} + C, \text{ when } b^2 = ac$$

25. $$\int \frac{(m+nx)\,dx}{a+2bx+cx^2} = \frac{n}{2c} \ln (a+2bx+cx^2) + \frac{mc-nb}{c}\int \frac{dx}{a+2bx+cx^2}$$

26. In $\int \frac{f(x)\,dx}{a+2bx+cx^2}$, if $f(x)$ is a polynomial of higher than the first degree, divide by the denominator before integrating

27. $$\int \frac{dx}{(a+2bx+cx^2)^p} = \frac{1}{2(ac-b^2)(p-1)} \times \frac{b+cx}{(a+2bx+cx^2)^{p-1}} + \frac{(2p-3)c}{2(ac-b^2)(p-1)} \int \frac{dx}{(a+2bx+cx^2)^{p-1}}$$

28. $$\int \frac{(m+nx)\,dx}{(a+2bx+cx^2)^p} = -\frac{n}{2c(p-1)} \times \frac{1}{(a+2bx+cx^2)^{p-1}} + \frac{mc-nb}{c} \int \frac{dx}{(a+2bx+cx^2)^p}$$

29. $$\int x^{m-1}(a+bx)^n\,dx = \frac{x^{m-1}(a+bx)^{n+1}}{(m+n)b} - \frac{(m-1)a}{(m+n)b} \int x^{m-2}(a+bx)^n\,dx = \frac{x^m(a+bx)^n}{m+n} + \frac{na}{m+n} \int x^{m-1}(a+bx)^{n-1}\,dx$$

Irrational Functions

30. $$\int \sqrt{a+bx}\,dx = \frac{2}{3b}(\sqrt{a+bx})^3 + C$$

31. $$\int \frac{dx}{\sqrt{a+bx}} = \frac{2}{b}\sqrt{a+bx} + C$$

32. $$\int \frac{(m+nx)\,dx}{\sqrt{a+bx}} = \frac{2}{3b^2}(3mb-2an+nbx)\sqrt{a+bx} + C$$

33. $\int \frac{dx}{(m+nx)\sqrt{a+bx}}$; substitute $y = \sqrt{a+bx}$, and use 21 and 22

34. $\int \frac{f(x, \sqrt[n]{a+bx})}{F(x, \sqrt[n]{a+bx})}\,dx$; substitute $\sqrt[n]{a+bx} = y$

35. $$\int \frac{dx}{\sqrt{a^2-x^2}} = \sin^{-1}\frac{x}{a} + C = -\cos^{-1}\frac{x}{a} + C$$

36. $$\int \frac{dx}{\sqrt{a^2+x^2}} = \ln (x+\sqrt{a^2+x^2}) + C = \sinh^{-1}\frac{x}{a} + C$$

37. $$\int \frac{dx}{\sqrt{x^2-a^2}} = \ln (x+\sqrt{x^2-a^2}) + C = \cosh^{-1}\frac{x}{a} + C$$

38. $$\int \frac{dx}{\sqrt{a+2bx+cx^2}}$$
$$= \frac{1}{\sqrt{c}} \ln (b+cx+\sqrt{c}\sqrt{a+2bx+cx^2}) + C, \text{ where } c>0$$
$$= \frac{1}{\sqrt{c}} \sinh^{-1}\frac{b+cx}{\sqrt{ac-b^2}} + C, \text{ when } ac-b^2>0$$
$$= \frac{1}{\sqrt{c}} \cosh^{-1}\frac{b+cx}{\sqrt{b^2-ac}} + C, \text{ when } b^2-ac>0$$
$$= \frac{-1}{\sqrt{-c}} \sin^{-1}\frac{b+cx}{\sqrt{b^2-ac}} + C, \text{ when } c<0$$

39. $$\int \frac{(m+nx)\,dx}{\sqrt{a+2bx+cx^2}} = \frac{n}{c}\sqrt{a+2bx+cx^2} + \frac{mc-nb}{c}\int \frac{dx}{\sqrt{a+2bx+cx^2}}$$

40. $$\int \frac{x^m\,dx}{\sqrt{a+2bx+cx^2}} = \frac{x^{m-1}X}{mc} - \frac{(m-1)a}{mc}\int \frac{x^{m-2}\,dx}{X} - \frac{(2m-1)b}{mc}\int \frac{x^{m-1}}{X}\,dx \text{ when } X = \sqrt{a+2bx+cx^2}$$

41. $$\int \sqrt{a^2+x^2}\,dx = \frac{x}{2}\sqrt{a^2+x^2} + \frac{a^2}{2}\ln (x+\sqrt{a^2+x^2}) + C = \frac{x}{2}\sqrt{a^2+x^2} + \frac{a^2}{2}\sinh^{-1}\frac{x}{a} + C$$

42. $$\int \sqrt{a^2-x^2}\,dx = \frac{x}{2}\sqrt{a^2-x^2} + \frac{a^2}{2}\sin^{-1}\frac{x}{a} + C$$

43. $$\int \sqrt{x^2-a^2}\,dx = \frac{x}{2}\sqrt{x^2-a^2} - \frac{a^2}{2}\ln (x+\sqrt{x^2-a^2}) + C = \frac{x}{2}\sqrt{x^2-a^2} - \frac{a^2}{2}\cosh^{-1}\frac{x}{a} + C$$

44. $$\int \sqrt{a+2bx+cx^2}\,dx = \frac{b+cx}{2c}\sqrt{a+2bx+cx^2} + \frac{ac-b^2}{2c}\int \frac{dx}{\sqrt{a+2bx+cx^2}} + C$$

Transcendental Functions

45. $$\int a^x\,dx = \frac{a^x}{\ln a} + C$$

46. $$\int x^n e^{ax}\,dx = \frac{x^n e^{ax}}{a}\left[1 - \frac{n}{ax} + \frac{n(n-1)}{a^2x^2} - \cdots \pm \frac{n!}{a^n x^n}\right] + C$$

47. $$\int \ln x\,dx = x\ln x - x + C$$

48. $$\int \frac{\ln x}{x^2}\,dx = -\frac{\ln x}{x} - \frac{1}{x} + C$$

49. $$\int \frac{(\ln x)^n}{x}\,dx = \frac{1}{n+1}(\ln x))^{n+1} + C$$

50. $$\int \sin^2 x\,dx = -\tfrac{1}{4}\sin 2x + \tfrac{1}{2}x + C = -\tfrac{1}{2}\sin x\cos x + \tfrac{1}{2}x + C$$

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

52. $$\int \sin mx\,dx = -\frac{\cos mx}{m} + C$$

53. $$\int \cos mx\,dx = \frac{\sin mx}{m} + C$$

54. $$\int \sin mx\cos nx\,dx = -\frac{\cos (m+n)x}{2(m+n)} - \frac{\cos (m-n)x}{2(m-n)} + C$$

55. $$\int \sin mx\sin nx\,dx = \frac{\sin (m-n)x}{2(m-n)} - \frac{\sin (m+n)x}{2(m+n)} + C$$

56. $$\int \cos mx\cos nx\,dx = \frac{\sin (m-n)x}{2(m-n)} + \frac{\sin (m+n)x}{2(m+n)} + C$$

57. $\int \tan x\,dx = -\ln\cos x + C$

58. $\int \cot x\,dx = \ln\sin x + C$

59. $\int \frac{dx}{\sin x} = \ln\tan\frac{x}{2} + C$

60. $\int \frac{dx}{\cos x} = \ln\tan\left(\frac{\pi}{4} + \frac{x}{2}\right) + C$

61. $\int \frac{dx}{1 + \cos x} = \tan\frac{x}{2} + C$

62. $\int \frac{dx}{1 - \cos x} = -\cot\frac{x}{2} + C$

63. $\int \sin x\cos x\,dx = ½\sin^2 x + C$

64. $\int \frac{dx}{\sin x\cos x} = \ln\tan x + C$

65.* $\int \sin^n x\,dx = -\frac{\cos x\sin^{n-1} x}{n} + \frac{n-1}{n}\int \sin^{n-2} x\,dx$

66.* $\int \cos^n x\,dx = \frac{\sin x\cos^{n-1} x}{n} + \frac{n-1}{n}\int \cos^{n-2} x\,dx$

67. $\int \tan^n x\,dx = \frac{\tan^{n-1} x}{n-1} - \int \tan^{n-2} x\,dx$

68. $\int \cot^n x\,dx = -\frac{\cot^{n-1} x}{n-1} - \int \cot^{n-2} x\,dx$

69. $\int \frac{dx}{\sin^n x} = -\frac{\cos x}{(n-1)\sin^{n-1} x} + \frac{n-2}{n-1}\int \frac{dx}{\sin^{n-2} x}$

70. $\int \frac{dx}{\cos^n x} = \frac{\sin x}{(n-1)\cos^{n-1} x} + \frac{n-2}{n-1}\int \frac{dx}{\cos^{n-2} x}$

71.† $\int \sin^p x\cos^q x\,dx = \frac{\sin^{p+1} x\cos^{q-1} x}{p+q}$

$+ \frac{q-1}{p+q}\int \sin^p x\cos^{q-2} x\,dx = -\frac{\sin^{p-1} x\cos^{q+1} x}{p+q}$

$+ \frac{p-1}{p+q}\int \sin^{p-2} x\cos^q x\,dx$

72.† $\int \sin^{-p} x\cos^q x\,dx = -\frac{\sin^{-p+1} x\cos^{q+1} x}{p-1}$

$+ \frac{p-q-2}{p-1}\int \sin^{-p+2} x\cos^q x\,dx$

73.† $\int \sin^p x\cos^{-q} x\,dx = \frac{\sin^{p+1} x\cos^{-q+1} x}{q-1}$

$+ \frac{q-p-2}{q-1}\int \sin^p x\cos^{-q+2} x\,dx$

74. $\int \frac{dx}{a + b\cos x} = \frac{2}{\sqrt{a^2 - b^2}}\tan^{-1}\left(\sqrt{\frac{a-b}{a+b}}\tan ½x\right) + C,$

when $a^2 > b^2$,

$= \frac{1}{\sqrt{b^2 - a^2}}\ln\frac{b + a\cos x + \sin x\sqrt{b^2 - a^2}}{a + b\cos x} + C,$

when $a^2 < b^2$,

$= \frac{2}{\sqrt{b^2 - a^2}}\tanh^{-1}\left(\sqrt{\frac{b-a}{b+a}}\tan ½x\right) + C$, when $a^2 < b^2$

75. $\int \frac{\cos x\,dx}{a + b\cos x} = \frac{x}{b} - \frac{a}{b}\int \frac{dx}{a + b\cos x} + C$

76. $\int \frac{\sin x\,dx}{a + b\cos x} = -\frac{1}{b}\ln(a + b\cos x) + C$

77. $\int \frac{A + B\cos x + C\sin x}{a + b\cos x + c\sin x}\,dx = A\int \frac{dy}{a + p\cos y}$

$+ (B\cos u + C\sin u)\int \frac{\cos y\,dy}{a + p\cos y}$

$- (B\sin u - C\cos u)\int \frac{\sin y\,dy}{a + p\cos y}$, where $b = p\cos u$, $c = p\sin u$ and $x - u = y$

78. $\int e^{ax}\sin bx\,dx = \frac{a\sin bx - b\cos bx}{a^2 + b^2}e^{ax} + C$

79. $\int e^{ax}\cos bx\,dx = \frac{a\cos bx + b\sin bx}{a^2 + b^2}e^{ax} + C$

80. $\int \sin^{-1} x\,dx = x\sin^{-1} x + \sqrt{1 - x^2} + C$

81. $\int \cos^{-1} x\,dx = x\cos^{-1} x - \sqrt{1 - x^2} + C$

82. $\int \tan^{-1} x\,dx = x\tan^{-1} x - ½\ln(1 + x^2) + C$

83. $\int \cot^{-1} x\,dx = x\cot^{-1}x + ½\ln(1 + x^2) + C$

84. $\int \sinh x\,dx = \cosh x + C$

85. $\int \tanh x\,dx = \ln\cosh x + C$

86. $\int \cosh x\,dx = \sinh x + C$

87. $\int \coth x\,dx = \ln\sinh x + C$

88. $\int \mathrm{sech}\,x\,dx = 2\tan^{-1}(e^x) + C$

89. $\int \mathrm{csch}\,x\,dx = \ln\tanh(x/2) + C$

90. $\int \sinh^2 x\,dx = ½\sinh x\cosh x - ½x + C$

91. $\int \cosh^2 x\,dx = ½\sinh x\cosh x + ½x + C$

92. $\int \mathrm{sech}^2 x\,dx = \tanh x + C$

93. $\int \mathrm{csch}^2 x\,dx = -\coth x + C$

Hints on Using Integral Tables It happens with frustrating frequency that no integral table lists the integral that needs to be evaluated. When this happens, one may (a) seek a more complete integral table, (b) appeal to mathematical software, such as Mathematica, Maple, MathCad or Derive, (c) use numerical or approximate methods, such as Simpson's rule (see section ''Numerical Methods''), or (d) attempt to transform the integral into one which may be evaluated. Some hints on such transformation follow. For a more complete list and more complete explanations, consult a calculus text, such as Thomas, ''Calculus and Analytic Geometry,'' Addison-Wesley, or Anton, ''Calculus with Analytic Geometry,'' Wiley. One or more of the following ''tricks'' may be successful.

Trigonometric Substitutions

1. If an integrand contains $\sqrt{(a^2 - x^2)}$, substitute $x = a\sin u$, and $\sqrt{(a^2 - x^2)} = a\cos u$.
2. Substitute $x = a\tan u$ and $\sqrt{(x^2 + a^2)} = a\sec u$.
3. Substitute $x = a\sec u$ and $\sqrt{(x^2 - a^2)} = a\tan u$.

Completing the Square

4. Rewrite $ax^2 + bx + c = a[x + b/(2a)]^2 + (4ac - b^2)/(4a)$; then substitute $u = x + b/(2a)$ and $B = (4ac - b^2)/(4a)$.

* If n is an odd number, substitute $\cos x = z$ or $\sin x = z$.

† If p or q is an odd number, substitute $\cos x = z$ or $\sin x = z$.

PARTIAL FRACTIONS

5. For a ratio of polynomials, where the denominator has been completely factored into linear factors $p_i(x)$ and quadratic factors $q_j(x)$, and where the degree of the numerator is less than the degree of the denominator, then rewrite $r(x)/[p_1(x) \ . \ . \ . \ p_n(x)q_1(x) \ . \ . \ . \ q_m(x)] = A_1/p_1(x) + \cdots + A_n/p_n(x) + (B_1x + C_1)/q_1(x) + \cdots + (B_mx + C_m)/q_m(x)$.

INTEGRATION BY PARTS

6. Change the integral using the formula

$$\int u\,dv = uv - \int v\,du$$

where u and dv are chosen so that (a) v is easy to find from dv, and (b) $v\,du$ is easier to find than $u\,dv$.

Kasube suggests (''A Technique for Integration by Parts,'' *Am. Math. Month.*, vol. 90, no. 3, Mar. 1983): Choose u in the order of preference LIATE, that is, Logarithmic, Inverse trigonometric, Algebraic, Trigonometric, Exponential.

EXAMPLE. Find $\int x \ln x\,dx$. The logarithmic $\ln x$ has higher priority than does the algebraic x, so let $u = \ln(x)$ and $dv = x\,dx$. Then $du = (1/x)\,dx$; $v = x^2/2$, so $\int x \ln x\,dx = uv - \int v\,du = (x^2/2)\ln x - \int (x^2/2)(1/x)\,dx = (x^2/2)\ln x - \int x/2\,dx = (x^2/2)\ln x - x^2/4 + C$.

Definite Integrals The definite integral of $f(x)\,dx$ from $x = a$ to $x = b$, denoted by $\int_a^b f(x)\,dx$, is the limit (as n increases indefinitely) of a sum of n terms:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} [f(x_1)\,\Delta x + f(x_2)\,\Delta x + f(x_3)\,\Delta x + \cdots + f(x_n)\,\Delta x]$$

built up as follows: Divide the interval from a to b into n equal parts, and call each part Δx, $=(b - a)/n$; in each of these intervals take a value of x (say, $x_1, x_2, \ . \ . \ . \ , x_n$), find the value of the function $f(x)$ at each of these points, and multiply it by Δx, the width of the interval; then take the limit of the sum of the terms thus formed, when the number of terms increases indefinitely, while each individual term approaches zero.

Geometrically, $\int_a^b f(x)\,dx$ is the area bounded by the curve $y = f(x)$, the x axis, and the ordinates $x = a$ and $x = b$ (Fig. 2.1.108); i.e., briefly, the ''area under the curve, from a to b.'' The **fundamental theorem** for the evaluation of a definite integral is the following:

$$\int_a^b f(x)\,dx = \left[\int f(x)\,dx\right]_{x=b} - \left[\int f(x)\,dx\right]_{x=a}$$

i.e., the definite integral is equal to the difference between two values of any one of the indefinite integrals of the function in question. In other words, the limit of a sum can be found whenever the function can be integrated.

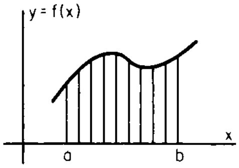


Fig. 2.1.108 Graph showing areas to be summed during integration.

Properties of Definite Integrals

$$\int_a^b = -\int_b^a; \qquad \int_a^c + \int_c^b = \int_a^b$$

MEAN-VALUE THEOREM FOR INTEGRALS

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

provided $f(x)$ does not change sign from $x = a$ to $x = b$; here X is some (unknown) value of x intermediate between a and b.

MEAN VALUE. The **mean value** of $f(x)$ with respect to x, between a and b, is

$$\bar{f} = \frac{1}{b - a}\int_a^b f(x)\,dx$$

THEOREM ON CHANGE OF VARIABLE. In evaluating $\int_{x=a}^{x=b} f(x)\,dx$, $f(x)\,dx$ may be replaced by its value in terms of a new variable t and dt, and $x = a$ and $x = b$ by the corresponding values of t, provided that throughout the interval the relation between x and t is a one-to-one correspondence (i.e., to each value of x there corresponds one and only one value of t, and to each value of t there corresponds one and only one value of x). So $\int_{x=a}^{x=b} f(x)\,dx = \int_{t=g(a)}^{t=g(b)} f(g(t))\,g'(t)\,dt$.

DIFFERENTIATION WITH RESPECT TO THE UPPER LIMIT. If b is variable, then $\int_a^b f(x)\,dx$ is a function b, whose derivative is

$$\frac{d}{db}\int_a^b f(x)\,dx = f(b)$$

DIFFERENTIATION WITH RESPECT TO A PARAMETER

$$\frac{\partial}{\partial c}\int_a^b f(x, c)\,dx = \int_a^b \frac{\partial f(x, c)}{\partial c}\,dx$$

Functions Defined by Definite Integrals The following definite integrals have received special names:

1. Elliptic integral of the first kind $= F(u, k) = \int_0^u \frac{dx}{\sqrt{1 - k^2 \sin^2 x}}$ when $k^2 < 1$.

2. Elliptic integral of the second kind $= E(u, k) = \int_0^u \sqrt{1 - k^2 \sin^2 x}\,dx$, when $k^2 < 1$.

3, 4. Complete elliptic integrals of the first and second kinds; put $u = \pi/2$ in (1) and (2).

5. The probability integral $= \frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}\,dx$.

6. The gamma function $= \Gamma(n) = \int_0^\infty x^{n-1}e^{-x}\,dx$.

Approximate Methods of Integration. Mechanical Quadrature (See also section ''Numerical Methods.'')

1. Use Simpson's rule (see also Scarborough, ''Numerical Mathematical Analyses,'' Johns Hopkins Press).

2. Expand the function in a converging power series, and integrate term by term.

3. Plot the area under the curve $y = f(x)$ from $x = a$ to $x = b$ on squared paper, and measure this area roughly by ''counting squares.''

Double Integrals The notation $\int\int f(x, y)\,dy\,dx$ means $\int[\int f(x, y)\,dy]\,dx$, the limits of integration in the inner, or first, integral being functions of x (or constants).

EXAMPLE. To find the weight of a plane area whose density, w, is variable, say $w = f(x, y)$. The weight of a typical element, $dx\,dy$, is $f(x, y)\,dx\,dy$. Keeping x and dx constant and summing these elements from, say, $y = F_1(x)$ to $y = F_2(x)$, as determined by the shape of the boundary (Fig. 2.1.109), the weight of a typical strip perpendicular to the x axis is

$$dx \int_{y=F_1(x)}^{y=F_2(x)} f(x, y)\,dy$$

Finally, summing these strips from, say, $x = a$ to $x = b$, the weight of the whole area is

$$\int_{x=a}^{x=b} \left[dx \int_{y=F_1(x)}^{y=F_2(x)} f(x, y)\,dy \right] \text{ or, briefly, } \iint f(x, y)\,dy\,dx$$

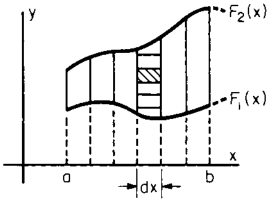


Fig. 2.1.109 Graph showing areas to be summed during double integration.

Triple Integrals The notation $\iiint f(x, y, z)\,dz\,dy\,dx$ means

$$\int \left\{ \int \left[\int f(x, y, z)\,dz \right] dy \right\} dx$$

Such integrals are known as **volume integrals**.

EXAMPLE. To find the mass of a volume which has variable density, say, $w = f(x, y, z)$. If the shape of the volume is described by $a < x < b$, $F_1(x) < y < F_2(x)$, and $G_1(x, y) < z < G_2(x, y)$, then the mass is given by

$$\int_a^b \int_{F_1(x)}^{F_2(x)} \int_{G_1(x,y)}^{G_2(x,y)} f(x, y, z)\,dz\,dy\,dx$$

SERIES AND SEQUENCES

Sequences

A **sequence** is an ordered list of numbers, $x_1, x_2, \ldots, x_n, \ldots$.

An infinite sequence is an infinitely long list. A sequence is often defined by a function $f(n)$, $n = 1, 2, \ldots$. The formula defining $f(n)$ is called the **general term** of the sequence.

The variable n is called the **index** of the sequence. Sometimes the index is taken to start with $n = 0$ instead of $n = 1$.

A sequence **converges** to a limit L if the general term $f(n)$ has limit L as n goes to infinity. If a sequence does not have a unique limit, the sequence is said to ''diverge.'' There are two fundamental ways a function can diverge: (1) It may become infinitely large, in which case the sequence is said to be ''unbounded,'' or (2) it may tend to alternate among two or more values, as in the sequence $x_n = (-1)^n$.

A sequence **alternates** if its odd-numbered terms are positive and its even-numbered terms are negative, or vice versa.

Series

A **series** is a sequence of sums. The terms of the sums are another sequence, $x_1, x_2, \ldots$. Then the series is the sequence defined by $s_n = x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$. The sequence s_n is also called the **sequence of partial sums** of the series.

If the sequence of partial sums converges (resp. diverges), then the series is said to converge (resp. diverge). If the limit of a series is S, then the sequence defined by $r_n = S - s_n$ is called the ''error sequence'' or the ''sequence of truncation errors.''

Convergence of Series THEOREM. If a series $s_n = x_1 + x_2 + \cdots + x_n$ converges, then it is necessary (but not sufficient) that the sequence x_n has limit zero. A series of partial sums of an alternating sequence is called an *alternating series*.

THEOREM. An alternating series converges whenever the sequence x_n has limit zero.

A series is a **geometric series** if its terms are of the form ar^n. The value r is called the **ratio** of the series. Usually, for geometric series, the index is taken to start with $n = 0$ instead of $n = 1$.

THEOREM. A geometric series with $x_n = ar^n$, $n = 0, 1, 2, \ldots$, converges if and only if $-1 < r < 1$, and then the limit of the series is $a/(1 - r)$. The partial sums of a geometric series are $s_n = a(1 - r^n)/(1 - r)$.

The series defined by the sequence $x_n = 1/n$, $n = 1, 2, \ldots$, is called the **harmonic series**. The harmonic series diverges.

A series with each term $x_n > 0$ is called a ''positive series.''

There are a number of tests to determine whether or not a positive series s_n converges.

1. *Comparison test*. If $c_1 + c_2 + \cdots + c_n$ is a positive series that converges, and if $0 < x_n < c_n$, then the series $x_1 + x_2 + \cdots + x_n$ also converges.

If $d_1 + d_2 + \cdots + d_n$ diverges and $x_n > d_n$, then $x_1 + x_2 + \cdots + x_n$ also diverges.

2. *Integral test*. If $f(t)$ is a strictly decreasing function and $f(n) = x_n$, then the series s_n and the integral $\int_1^\infty f(t)\,dt$ either both converge or both diverge.

3. *P test*. The series defined by $x_n = 1/n^p$ converges if $p > 1$ and diverges if $p = 1$ or $p < 1$. If $p = 1$, then this is the harmonic series.

4. *Ratio test*. If the limit of the sequence $x_{n+1}/x_n = r$, then the series diverges if $r > 1$, and it converges if $0 < r < 1$. The test is inconclusive if $r = 1$.

5. *Cauchy root test*. If L is the limit of the nth root of the nth term, $\lim x_n^{1/n}$, then the series converges if $L < 1$ and diverges if $L > 1$. If $L = 1$, then the test is inconclusive.

A **power series** is an expression of the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$ or $\sum_{i=0}^{\infty} a_ix^i$.

The range of values of x for which a power series converges is the **interval of convergence** of the power series.

General Formulas of Maclaurin and Taylor If $f(x)$ and all its derivatives are continuous in the neighborhood of the point $x = 0$ (or $x = a$), then, for any value of x in this neighborhood, the function $f(x)$ may be expressed as a power series arranged according to ascending powers of x (or of $x - a$), as follows:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n-1)}(0)}{(n-1)!}x^{n-1} + (P_n)x^n \qquad \text{(Maclaurin)}$$

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x - a)^{n-1} + (Q_n)(x - a)^n \qquad \text{(Taylor)}$$

Here $(P_n)x^n$, or $(Q_n)(x - a)^n$, is called the **remainder term;** the values of the coefficients P_n and Q_n may be expressed as follows:

$$\begin{aligned} P_n &= [f^{(n)}(sx)]/n! = [(1 - t)^{n-1}f^{(n)}(tx)]/(n - 1)! \\ Q_n &= \{f^{(n)}[a + s(x - a)]\}/n! \\ &= \{(1 - t)^{n-1}f^{(n)}[a + t(x - a)]\}/(n - 1)! \end{aligned}$$

where s and t are certain unknown numbers between 0 and 1; the s form is due to Lagrange, the t form to Cauchy.

The error due to neglecting the remainder term is less than $(\overline{P}_n)x^n$, or $(\overline{Q}_n)(x - a)^n$, where $\overline{P}_n$, or $\overline{Q}_n$, is the largest value taken on by P_n, or

Q_n, when s or t ranges from 0 to 1. If this error, which depends on both n and x, approaches 0 as n increases (for any given value of x), then the general expression with remainder becomes (for that value of x) a convergent infinite series.

The sum of the first few terms of Maclaurin's series gives a good approximation to $f(x)$ for values of x near $x = 0$; Taylor's series gives a similar approximation for values near $x = a$.

The MacLaurin series of some important functions are given below.

Power series may be differentiated term by term, so the derivative of a power series $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ is $a_1 + 2a_2x + \cdots + na_xx^{n-1}$. . . . The power series of the derivative has the same interval of convergence, except that the endpoints may or may not be included in the interval.

Series Expansions of Some Important Functions

The range of values of x for which each of the series is convergent is stated at the right of the series.

Geometrical Series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \qquad -1 < x < 1$$

$$\frac{1}{(1-x)^m} = 1 + \sum_{n=1}^{\infty} \frac{(m+n-1)!}{(m-1)!n!} x^n \qquad -1 < x < 1$$

Exponential and Logarithmic Series

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad [-\infty < x < +\infty]$$

$$a^x = e^{mx} = 1 + \frac{m}{1!}x + \frac{m^2}{2!}x^2 + \frac{m^3}{3!}x^3 + \cdots \qquad [a > 0, -\infty < x < +\infty]$$

where $m = \ln a = (2.3026)(\log_{10} a)$.

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} \cdots \qquad [-1 < x < +1]$$

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \frac{x^5}{5} - \cdots \qquad [-1 < x < +1]$$

$$\ln\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots\right) \qquad [-1 < x < +1]$$

$$\ln\left(\frac{x+1}{x-1}\right) = 2\left(\frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \frac{1}{7x^7} + \cdots\right) \qquad [x < -1 \text{ or } +1 < x]$$

$$\ln x = 2\left[\frac{x-1}{x+1} + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \cdots\right] \qquad [0 < x < \infty]$$

$$\ln(a+x) = \ln a + 2\left[\frac{x}{2a+x} + \frac{1}{3}\left(\frac{x}{2a+x}\right)^3 + \frac{1}{5}\left(\frac{x}{2a+x}\right)^5 + \cdots\right] \qquad [0 < a < +\infty, -a < x < +\infty]$$

Series for the Trigonometric Functions In the following formulas, *all angles must be expressed in radians*. If $D =$ the number of degrees in the angle, and $x =$ its radian measure, then $x = 0.017453D$.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad [-\infty < x < +\infty]$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots \qquad [-\infty < x < +\infty]$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \cdots \qquad [-\pi/2 < x < +\pi/2]$$

$$\cot x = \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - \frac{2x^5}{945} - \frac{x^7}{4725} - \cdots \qquad [-\pi < x < +\pi]$$

$$\sin^{-1} y = y + \frac{y^3}{6} + \frac{3y^5}{40} + \frac{5y^7}{112} + \cdots \qquad [-1 \le y \le +1]$$

$$\tan^{-1} y = y - \frac{y^3}{3} + \frac{y^5}{5} - \frac{y^7}{7} + \cdots \qquad [-1 \le y \le +1]$$

$$\cos^{-1} y = \tfrac{1}{2}\pi - \sin^{-1} y; \qquad \cot^{-1} y = \tfrac{1}{2}\pi - \tan^{-1} y.$$

Series for the Hyperbolic Functions (x a pure number)

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots \qquad [-\infty < x < \infty]$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots \qquad [-\infty < x < \infty]$$

$$\sinh^{-1} y = y - \frac{y^3}{6} + \frac{3y^5}{40} - \frac{5y^7}{112} + \cdots \qquad [-1 < y < +1]$$

$$\tanh^{-1} y = y + \frac{y^3}{3} + \frac{y^5}{5} + \frac{y^7}{7} + \cdots \qquad [-1 < y < +1]$$

ORDINARY DIFFERENTIAL EQUATIONS

An **ordinary differential equation** is one which contains a single independent variable, or argument, and a single dependent variable, or function, with its derivatives of various orders. A **partial differential equation** is one which contains a function of several independent variables, and its partial derivatives of various orders. The order of a differential equation is the order of the highest derivative which occurs in it. A solution of a differential equation is any relation among the variables, involving no derivatives, though possibly involving integrations which, when substituted in the given equation, will satisfy it. The general solution of an ordinary differential equation of the nth order will contain n arbitrary constants.

If specific values of the arbitrary constants are chosen, then a solution is called a **particular solution.** For most problems, all possible particular solutions to a differential equation may be found by choosing values for the constants in a general solution. In some cases, however, other solutions exist. These are called **singular solutions**.

EXAMPLE. The differential equation $(yy')^2 - a^2 - y^2 = 0$ has general solution $(x - c)^2 + y^2 = a^2$, where c is an arbitrary constant. Additionally, it has the two singular solutions $y = a$ and $y = -a$. The singular solutions form two parallel lines tangent to the family of circles given by the general solution.

The example illustrates a general property of singular solutions; at each point on a singular solution, the singular solution is tangent to some curve given in the general solution.

Methods of Solving Ordinary Differential Equations

DIFFERENTIAL EQUATIONS OF THE FIRST ORDER

1. If possible, separate the variables; i.e., collect all the x's and dx on one side, and all the y's and dy on the other side; then integrate both sides, and add the constant of integration.

2. If the equation is homogeneous in x and y, the value of dy/dx in terms of x and y will be of the form $dy/dx = f(y/x)$. Substituting $y = xt$ will enable the variables to be separated.

Solution: $\log_e x = \int \frac{dt}{f(t) - t} + C.$

3. The expression $f(x, y)\,dx + F(x, y)\,dy$ is an *exact differential* if $\dfrac{\partial f(x, y)}{\partial y} = \dfrac{\partial F(x, y)}{\partial x}$ $(= P$, say). In this case the solution of $f(x, y)\,dx + F(x, y)\,dy = 0$ is

$$\int f(x, y)\,dx + \int [F(x, y) - \int P\,dx]\,dy = C$$

or

$$\int F(x, y)\,dy + \int [f(x, y) - \int P\,dy]\,dx = C$$

4. Linear differential equation of the first order: $\dfrac{dy}{dx} + f(x) \cdot y = F(x)$.

Solution: $y = e^{-P}[\int e^{P}F(x)\,dx + C]$, where $P = \int f(x)\,dx$

5. Bernoulli's equation: $\dfrac{dy}{dx} + f(x) \cdot y = F(x) \cdot y^n$. Substituting $y^{1-n} = v$ gives $(dv/dx) + (1 - n)f(x) \cdot v = (1 - n)F(x)$, which is linear in v and x.

6. Clairaut's equation: $y = xp + f(p)$, where $p = dy/dx$. The solution consists of the family of lines given by $y = Cx + f(C)$, where C is any constant, together with the curve obtained by eliminating p between the equations $y = xp + f(p)$ and $x + f'(p) = 0$, where $f'(p)$ is the derivative of $f(p)$.

7. Riccati's equation. $p + ay^2 + Q(x)y + R(x) = 0$, where $p = dy/dx$ can be reduced to a second-order linear differential equation $(d^2u/dx^2) + Q(x)(du/dx) + R(x) = 0$ by the substitution $y = du/dx$.

8. Homogeneous equations. A function $f(x, y)$ is **homogeneous** of degree n if $f(rx, ry) = r^m f(x, y)$, for all values of r, x, and y. In practice, this means that $f(x, y)$ looks like a polynomial in the two variables x and y, and each term of the polynomial has total degree m. A differential equation is homogeneous if it has the form $f(x, y) = 0$, with f homogeneous. $(xy + x^2)\,dx + y^2\,dy = 0$ is homogeneous. Cos $(xy)\,dx + y^2\,dy = 0$ is not.

If an equation is homogeneous, then either of the substitutions $y = vx$ or $x = vy$ will transform the equation into a separable equation.

9. $dy/dx = f[(ax + by + c)/(dx + ey + g)]$ is reduced to a homogeneous equation by substituting $u = ax + by + c$, $v = dx + ey + g$, if $ae - bd = 0$, and $z = ax + by$, $w = dx + ey$ if $ae - bd = 0$.

Differential Equations of the Second Order

10. *Dependent variable missing.* If an equation does not involve the variable y, and is of the form $F(x, dy/dx, d^2y/dx^2) = 0$, then it can be reduced to a first-order equation by substituting $p = dy/dx$ and $dp/dx = d^2y/dx^2$.

11. *Independent variable missing.* If the equation is of the form $F(y, dy/dx, d^2y/dx^2) = 0$, and so is missing the variable x, then it can be reduced to a first-order equation by substituting $p = dy/dx$ and $p(dp/dy) = d^2y/dx^2$.

12. $\dfrac{d^2y}{dx^2} = -n^2y$.

Solution: $y = C_1 \sin(nx + C_2)$, or $y = C_3 \sin nx + C_4 \cos nx$.

13. $\dfrac{d^2y}{dx^2} = +n^2y$.

Solution: $y = C_1 \sinh(nx + C_2)$, or $y = C_3e^{nx} + C_4e^{-nx}$.

14. $\dfrac{d^2y}{dx^2} = f(y)$.

Solution: $x = \displaystyle\int \frac{dy}{\sqrt{C_1 + 2P}} + C_2$, where $P = \displaystyle\int f(y)\,dy$.

15. $\dfrac{d^2y}{dx^2} = f(x)$.

Solution:

$$y = \int P\,dx + C_1x + C_2 \qquad \text{where } P = \int f(x)\,dx,$$

or

$$y = xP - \int xf(x)\,dx + C_1x + C_2.$$

16. $\dfrac{d^2y}{dx^2} = f\left(\dfrac{dy}{dx}\right)$. Putting $\dfrac{dy}{dx} = z$, $\dfrac{d^2y}{dx^2} = \dfrac{dz}{dx}$,

$$x = \int \frac{dz}{f(z)} + C_1, \qquad \text{and} \qquad y = \int \frac{z\,dz}{f(z)} + C_2$$

then eliminate z from these two equations.

17. The equation for damped vibrations: $\dfrac{d^2y}{dx^2} + 2b\,\dfrac{dy}{dx} + a^2y = 0$.

Case 1. If $a^2 - b^2 > 0$, let $m = \sqrt{a^2 - b^2}$.
Solution:

$$y = C_1e^{-bx} \sin(mx + C_2) \quad \text{or} \quad y = e^{-bx}[C_3 \sin(mx) + C_4 \cos(mx)]$$

Case 2. If $a^2 - b^2 = 0$, solution is $y = e^{-bx}(C_1 + C_2x)$.

Case 3. If $a^2 - b^2 < 0$, let $n = \sqrt{b^2 - a^2}$.
Solution:

$$y = C_1e^{-bx} \sinh(nx + C_2) \qquad \text{or} \qquad y = C_3e^{-(b+n)x} + C_4e^{-(b-n)x}$$

18. $\dfrac{d^2y}{dx^2} + 2b\,\dfrac{dy}{dx} + a^2y = c$.

Solution: $y = \dfrac{c}{a^2} + y_1$, where y_1 = the solution of the corresponding equation with second member zero [see type 17 above].

19. $\dfrac{d^2y}{dx^2} + 2b\,\dfrac{dy}{dx} + a^2y = c \sin(kx)$.

Solution: $y = R \sin(kx - S) + y_1$

where $R = c/\sqrt{(a^2 - k^2)^2 + 4b^2k^2}$, $\tan S = 2bk/(a^2 - k^2)$, and y_1 = the solution of the corresponding equation with second member zero [see type 17 above].

20. $\dfrac{d^2y}{dx^2} + 2b\,\dfrac{dy}{dx} + a^2y = f(x)$.

Solution: $y = R \sin(kx - S) + y_1$

where $R = c/\sqrt{(a^2 - k^2)^2 + 4b^2k^2}$, $\tan S = 2bk/(a^2 - k^2)$, and y_1 = the solution of the corresponding equation with second member zero [see type 17 above].

If $b^2 < a^2$,

$$y_0 = \frac{1}{2\sqrt{b^2 - a^2}}\left[e^{m_1x}\int e^{-m_1x} f(x)\,dx - e^{m_2x}\int e^{-m_2x} f(x)\,dx\right]$$

where $m_1 = -b + \sqrt{b^2 - a^2}$ and $m_2 = -b - \sqrt{b^2 - a^2}$.

If $b^2 < a^2$, let $m = \sqrt{a^2 - b^2}$, then

$$y_0 = \frac{1}{m}\,e^{-bx}\left[\sin(mx)\int e^{bx}\cos(mx) \cdot f(x)\,dx - \cos(mx)\int e^{bx}\sin(mx) \cdot f(x)\,dx\right]$$

If $b^2 = a^2$, $y_0 = e^{-bx}\left[x\displaystyle\int e^{bx}f(x)\,dx - \int x \cdot e^{bx} f(x)\,dx\right]$.

Types 17 to 20 are examples of linear differential equations with constant coefficients. The solutions of such equations are often found most simply by the use of Laplace transforms. (See Franklin, ''Fourier Methods,'' pp. 198–229, McGraw-Hill.)

Linear Equations

For the linear equation of the nth order

$$A_n(x)\,d^ny/dx^n + A_{n-1}(x)\,d^{n-1}y/dx^{n-1} + \cdots + A_1(x)\,dy/dx + A_0(x)y = E(x)$$

the general solution is $y = u + c_1u_1 + c_2u_2 + \cdots + c_nu_n$. Here u, the particular integral, is any solution of the given equation, and $u_1, u_2, \ldots, u_n$ form a fundamental system of solutions of the homogeneous equation obtained by replacing $E(x)$ by zero. A set of solutions is fundamental, or independent, if its Wronskian determinant $W(x)$ is not zero, where

$$W(x) = \begin{vmatrix} u_1 & u_2 & \cdots & u_n \\ u_1' & u_2' & \cdots & u_n' \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ u_1^{(n-1)} & u_2^{(n-1)} & \cdots & u_n^{(n-1)} \end{vmatrix}$$

For any n functions, $W(x) = 0$ if some one u_i is linearly dependent on the others, as $u_n = k_1u_1 + k_2u_2 + \cdots + k_{n-1}u_{n-1}$ with the coefficients k_i constant. And for n solutions of a linear differential equation of the nth order, if $W(x) \neq 0$, the solutions are linearly independent.

Constant Coefficients To solve the homogeneous equation of the nth order $A_nd^ny/dx^n + A_{n-1}d^{n-1}y/dx^{n-1} + \cdots + A_1dy/dx + A_0y = 0$, $A_n \neq 0$, where $A_n, A_{n-1}, \ldots, A_0$ are constants, find the roots of the auxiliary equation

$$A_np^n + A_{n-1}p^{n-1} + \cdots + A_1p + A_0 = 0$$

For each simple real root r, there is a term ce^{rx} in the solution. The terms of the solution are to be added together. When r occurs twice among the n roots of the auxiliary equation, the corresponding term is $e^{rx}(c_1 + c_2x)$. When r occurs three times, the corresponding term is $e^{rx}(c_1 + c_2x + c_3x^2)$, and so forth. When there is a pair of conjugate complex roots $a + bi$ and $a - bi$, the real form of the terms in the solution is $e^{ax}(c_1 \cos bx + d_1 \sin bx)$. When the same pair occurs twice, the corresponding term is $e^{ax}[(c_1 + c_2x) \cos bx + (d_1 + d_2x) \sin bx]$, and so forth.

Consider next the general nonhomogeneous linear differential equation of order n, with constant coefficients, or

$$A_nd^ny/dx^n + A_{n-1}d^{n-1}y/dx^{n-1} + \cdots + A_1\, dy/dx + A_0y = E(x)$$

We may solve this by adding any particular integral to the complementary function, or general solution, of the homogeneous equation obtained by replacing $E(x)$ by zero. The complementary function may be found from the rules just given. And the particular integral may be found by the methods of the following paragraphs.

Undetermined Coefficients In the last equation, let the right member $E(x)$ be a sum of terms each of which is of the type k, $k \cos bx$, $k \sin bx$, ke^{ax}, kx, or more generally, kx^me^{ax}, $kx^me^{ax} \cos bx$, or $kx^me^{ax} \sin bx$. Here m is zero or a positive integer, and a and b are any real numbers. Then the form of the particular integral I may be predicted by the following rules.

CASE 1. $E(x)$ **is a single term** T. Let D be written for d/dx, so that the given equation is $P(D)y = E(x)$, where $P(D) = A_nD^n + A_{n-1}D^{n-1} + \cdots + A_1D + A_0y$. With the term T associate the simplest polynomial $Q(D)$ such that $Q(D)T = 0$. For the particular types k, etc., $Q(D)$ will be D, $D^2 + b^2$, $D^2 + b^2$, $D - a$, D^2; and for the general types kx^me^{ax}, etc., $Q(D)$ will be $(D - a)^{m+1}$, $(D^2 - 2aD + a^2 + b^2)^{m+1}$, $(D^2 - 2aD + a^2 + b^2)^{m+1}$. Thus $Q(D)$ will always be some power of a first- or second-degree factor, $Q(D) = F^V$, $F = D - a$, or $F = D^2 - 2aD + a^2 + b^2$.

Use the method described under **Constant Coefficients** to find the terms in the solution of $P(D)y = 0$ and also the terms in the solution of $Q(D)P(D)y = 0$. Then assume the particular integral I is a linear combination with unknown coefficients of those terms in the solution of $Q(D)P(D)y = 0$ which are not in the solution of $P(D)y = 0$. Thus if $Q(D) = F^q$ and F is *not* a factor of $P(D)$, assume $I = (Ax^{q-1} + Bx^{q-2} + \cdots + L)e^{ax}$ when $F = D - a$, and assume $I = (Ax^{q-1} + Bx^{q-2} + \cdots + L)e^{ax} \cos bx + (Mx^{q-1} + Nx^{q-2} + \cdots + R)e^{ax} \sin bx$ when $F = D^2 - 2aD + a^2 + b^2$. When F is a factor of $P(D)$ and the highest power of F which is a divisor of $P(D)$ is F^k, try the I above multiplied by x^k.

CASE 2. $E(x)$ **is a sum of terms**. With each term in $E(x)$, associate a polynomial $Q(D) = F^q$ as before. Arrange in one group all the terms that have the same F. The particular integral of the given equation will be the sum of solutions of equations each of which has one group on the right. For any one such equation, the form of the particular integral is given as for Case 1, with q the highest power of F associated with any term of the group on the right.

After the form has been found in Case 1 or 2, the unknown coefficients follow when we substitute back in the given differential equation, equate coefficients of like terms, and solve the resulting system of simultaneous equations.

Variation of Parameters. Whenever a fundamental system of solutions $u_1, u_2, \ldots, u_n$ for the homogeneous equation is known, a particular integral of

$$A_n(x)d^ny/dx^n + A_{n-1}(x)d^{n-1}y/dx^{n-1} + \cdots \\ + A_1(x)\, dy/dx + A_0(x)y = E(x)$$

may be found in the form $y = \Sigma v_ku_k$. In this and the next few summations, k runs from 1 to n. The v_k are functions of x, found by integrating their derivatives v_k', and these derivatives are the solutions of the n simultaneous equations $\Sigma v_k'u_k = 0$, $\Sigma v_k'u_k' = 0$, $\Sigma v_k'u''_k = 0, \cdots$, $\Sigma v_k'u_k^{(n-2)} = 0$, $A_n(x)\Sigma v_k'u_k^{(n-1)} = E(x)$. To find the v_k from $v_k = \int v_k'\, dx + c_k$, any choice of constants will lead to a particular integral. The special choice $v_k = \int_0^x v_k'\, dx$ leads to the particular integral having $y, y', y'', \ldots, y^{(n-1)}$ each equal to zero when $x = 0$.

The Cauchy-Euler Equidimensional Equation This has the form

$$k_nx^nd^ny/dx^n + k_{n-1}x^{n-1}d^{n-1}y/dx^{n-1} + \cdots \\ + k_1x\, dy/dx + k_0y = F(x)$$

The substitution $x = e^t$, which makes

$$x\, dy/dx = dy/dt$$
$$x^k\, d^ky/dx^k = (d/dt - k + 1) \cdots (d/dt - 2)(d/dt - 1)\, dy/dt$$

transforms this into a linear differential equation with constant coefficients. Its solution $y = g(t)$ leads to $y = g(\ln x)$ as the solution of the given Cauchy-Euler equation.

Bessel's Equation The general Bessel equation of order n is:

$$x^2y'' + xy' + (x^2 - n^2)y = 0$$

This equation has general solution

$$y = AJ_n(x) + BJ_{-n}(x)$$

when n is not an integer. Here, $J_n(x)$ and $J_{-n}(x)$ are Bessel functions (see section on Special Functions).

In case $n = 0$, Bessel's equation has solution

$$y = AJ_0(x) + B\left[J_0(x) \ln (x) - \sum_{k=1}^{\infty} \frac{(-1)^kH_kx^{2k}}{2^{2k}(k!)^2}\right]$$

where H_k is the kth partial sum of the harmonic series, $1 + ½ + ⅓ + \cdots + 1/k$.

In case $n = 1$, the solution is

$$y = AJ_1(x) + B\left\{J_1(x) \ln (x) + 1/x \right. \\ \left. - \left[\sum_{k=1}^{\infty} \frac{(-1)^k(H_k + H_{k-1})x^{2k-1}}{2^{2k}k!(k-1)!}\right]\right\}$$

In case $n > 1$, n is an integer, solution is

$$y = AJ_n(x) + B\left\{J_n(x) \ln (x) + \left[\sum_{k=0}^{\infty} \frac{(-1)^{k+1}(n-1)!x^{2k-n}}{2^{2k+1-n}k!(1-n)^k}\right] \right. \\ \left. + ½\left[\sum_{k=0}^{\infty} \frac{(-1)^{k+1}(H_k + H_{k+1})x^{2k+n}}{2^{2k+n}k!(k+n)!}\right]\right\}$$

Solutions to Bessel's equation may be given in several other forms, often exploiting the relation between H_k and $\ln (k)$ or the so-called *Euler constant*.

General Method of Power Series Given a general differential equation $F(x, y, y', \ldots) = 0$, the solution may be expanded as a Maclaurin series, so $y = \Sigma_{n=0}^{\infty} a_nx^n$, where $a_n = f^{(n)}(0)/n!$. The power series for y may be differentiated formally, so that $y' = \Sigma_{n=1}^{\infty} na_n x^{n-1} = \Sigma_{n=0}^{\infty} (n+1)a_{n+1}x^n$, and $y'' = \Sigma_{n=2}^{\infty} n(n-1)a_n x^{n-2} = \Sigma_{n=0}^{\infty} (n+1)(n+2)a_{n+2}x^n$.

Substituting these series into the equation $F(x, y, y', \ldots) = 0$ often gives useful recursive relationships, giving the value of a_n in terms of previous values. If approximate solutions are useful, then it may be sufficient to take the first few terms of the Maclaurin series as a solution to the equation.

EXAMPLE. Consider $y'' - y' + xy = 0$. The procedure gives $\Sigma_{n=0}^{\infty} (n+1)(n+2)a_{n+2}x^n - \Sigma_{n=0}^{\infty} (n+1)a_{n+1}x^n + x\,\Sigma_{n=0}^{\infty} a_n x^n = \Sigma_{n=0}^{\infty} (n+1)(n+2)a_{n+2}x^n - \Sigma_{n=0}^{\infty} (n+1)a_{n+1}x^n + \Sigma_{n=1}^{\infty} a_{n-1}x^n = (2a_2 - a_1)x^0 + \Sigma_{n=1}^{\infty} [(n+1)(n+2)a_{n+2} - (n+1)a_{n+1} + a_{n-1})]\,x^n = 0$. Thus $2a_2 - a_1 = 0$ and, for $n > 0$, $(n+1)(n+2)a_{n+2} - (n+1)a_{n+1} + a_{n-1} = 0$. Thus, a_0 and a_1 may be determined arbitrarily, but thereafter, the values of a_n are determined recursively.

PARTIAL DIFFERENTIAL EQUATIONS

Partial differential equations (PDEs) arise when there are two or more independent variables. Two notations are common for the partial derivatives involved in PDEs, the ''del'' or fraction notation, where the first partial derivative of f with respect to x would be written $\partial f/\partial x$, and the subscript notation, where it would be written f_x.

In the same way that ordinary differential equations often involve arbitrary constants, solutions to PDEs often involve arbitrary functions.

EXAMPLE. $f_{xy} = 0$ has as its general solution $g(x) + h(y)$. The function g does not depend on y, so $g_y = 0$. Similarly, $f_x = 0$.

PDEs usually involve boundary or initial conditions dictated by the application. These are analogous to initial conditions in ordinary differential equations.

In solving PDEs, it is seldom feasible to find a general solution and then specialize that general solution to satisfy the boundary conditions, as is done with ordinary differential equations. Instead, the boundary conditions usually play a key role in the solution of a problem. A notable exception to this is the case of linear, homogeneous PDEs since they have the property that if f_1 and f_2 are solutions, then $f_1 + f_2$ is also a solution. The wave equation is one such equation, and this property is the key to the solution described in the section ''Fourier Series.''

Often it is difficult to find exact solutions to PDEs, so it is necessary to resort to approximations or numerical solutions.

Classification of PDEs

Linear A PDE is **linear** if it involves only first derivatives, and then only to the first power. The general form of a linear PDE, in two independent variables, x and y, and the dependent variable z, is $P(x, y, z)f_x + Q(x, y, z)f_y = R(x, y, z)$, and it will have a solution of the form $z = f(x, y)$ if its solution is a function, or $F(x, y, z) = 0$ if the solution is not a function.

Elliptic Laplace's equation $f_{xx} + f_{yy} = 0$ and Poisson's equation $f_{xx} + f_{yy} = g(x, y)$ are the prototypical elliptic equations. They have analogs in more than two variables. They do not explicitly involve the variable time and generally describe steady-state or equilibrium conditions, gravitational potential, where boundary conditions are distributions of mass, electrical potential, where boundary conditions are electrical charges, or equilibrium temperatures, and where boundary conditions are points where the temperature is held constant.

Parabolic $T_t = T_{xx} + T_{yy}$ represents the dynamic condition of diffusion or heat conduction, where $T(x, y, t)$ usually represents the temperature at time t at the point (x, y). Note that when the system reaches steady state, the temperature is no longer changing, so $T_t = 0$, and this becomes Laplace's equation.

Hyperbolic Wave propagation is described by equations of the type $u_{tt} = c^2(u_{xx} + u_{yy})$, where c is the velocity of waves in the medium.

VECTOR CALCULUS

Vector Fields A **vector field** is a function that assigns a vector to each point in a region. If the region is two-dimensional, then the vectors assigned are two-dimensional, and the vector field is a two-dimensional vector field, denoted $\mathbf{F}(x, y)$. In the same way, a three-dimensional vector field is denoted $\mathbf{F}(x, y, z)$.

A three-dimensional vector field can always be written:

$$\mathbf{F}(x, y, z) = f_1(x, y, z)\mathbf{i} + f_2(x, y, z)\mathbf{j} + f_3(x, y, z)\mathbf{k}$$

where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are the basis vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1), respectively. The functions f_1, f_2, and f_3 are called **coordinate functions** of $\mathbf{F}$.

Parameterized Curves If C is a curve from a point A to a point B, either in two dimensions or in three dimensions, then a **parameterization** of C is a vector-valued function $\mathbf{r}(t) = r_1(t)\mathbf{i} + r_2(t)\mathbf{j} + r_3(t)\mathbf{k}$, which satisfies $\mathbf{r}(a) = A$, $\mathbf{r}(b) = B$, and $\mathbf{r}(t)$ is on the curve C, for $a \leq t \leq b$. It is also necessary that the function $\mathbf{r}(t)$ be continuous and one-to-one. A given curve C has many different parameterizations.

The derivative of a parameterization $\mathbf{r}(t)$ is a vector-valued function $\mathbf{r}'(t) = r_1'(t)\mathbf{i} + r_2'(t)\mathbf{j} + r_3'(t)\mathbf{k}$. The derivative is the velocity function of the parameterization. It is always tangent to the curve C, and the magnitude is the speed of the parameterization.

Line Integrals If $\mathbf{F}$ is a vector field, C is a curve, and $\mathbf{r}(t)$ is a parameterization of C, then the **line integral,** or **work integral,** of $\mathbf{F}$ along C is

$$W = \int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\,dt$$

This is sometimes called the work integral because if $\mathbf{F}$ is a force field, then W is the amount of work necessary to move an object along the curve C from A to B.

Divergence and Curl The **divergence** of a vector field $\mathbf{F}$ is div $\mathbf{F} = f_{1x} + f_{2y} + f_{3z}$. If $\mathbf{F}$ represents the flow of a fluid, then the divergence at a point represents the rate at which the fluid is expanding at that point. Vector fields with div $\mathbf{F} = 0$ are called **incompressible**.

The **curl** of $\mathbf{F}$ is

$$\text{curl } \mathbf{F} = (f_{3y} - f_{2z})\mathbf{i} + (f_{1z} - f_{3x})\mathbf{j} + (f_{2x} - f_{1y})\mathbf{k}$$

If $\mathbf{F}$ is a two-dimensional vector field, then the first two terms of the curl are zero, so the curl is just

$$\text{curl } \mathbf{F} = (f_{2x} - f_{1y})\mathbf{k}$$

If $\mathbf{F}$ represents the flow of a fluid, then the curl represents the rotation of the fluid at a given point. Vector fields with curl $\mathbf{F} = \mathbf{0}$ are called **irrotational.**

Two important facts relate div, grad, and curl:

1. div (curl $\mathbf{F}$) = 0
2. curl (grad f) = $\mathbf{0}$

Conservative Vector Fields A vector field $\mathbf{F} = f_1\mathbf{i} + f_2\mathbf{j} + f_3\mathbf{k}$ is **conservative** if all of the following are satisfied:

$$f_{1y} = f_{2x} \qquad f_{1z} = f_{3x} \qquad \text{and} \qquad f_{2z} = f_{3y}$$

If $\mathbf{F}$ is a two-dimensional vector field, then the second and third conditions are always satisfied, and so only the first condition must be checked. Conservative vector fields have three important properties:

1. $\int_C \mathbf{F} \cdot d\mathbf{r}$ has the same value regardless of what curve C is chosen that connects the points A and B. This property is called **path independence.**
2. $\mathbf{F}$ is the gradient of some function $f(x, y, z)$.
3. Curl $\mathbf{F} = \mathbf{0}$.

In the special case that $\mathbf{F}$ is a conservative vector field, if $\mathbf{F}$ = grad

(f), then

$$\int_C \mathbf{F} \cdot d\mathbf{r} = f(B) - f(A)$$

THEOREMS ABOUT LINE AND SURFACE INTEGRALS

Two important theorems relate line integrals with double integrals. If R is a region in the plane and if C is the curve tracing the boundary of R in the positive (counterclockwise) direction, and if $\mathbf{F}$ is a continuous vector field with continuous first partial derivatives, line integrals on C are related to double integrals on R by Green's theorem and the divergence theorem.

Green's Theorem

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \iint_R \text{curl } (\mathbf{F}) \cdot d\mathbf{S}$$

The right-hand double integral may also be written as $\iint_R |\text{curl } (\mathbf{F})|$ dA.

Green's theorem describes the total rotation of a vector field in two different ways, on the left in terms of the boundary of the region and on the right in terms of the rotation at each point within the region.

Divergence Theorem

$$\int_C \mathbf{F} \cdot d\mathbf{N} = \iint_R \text{div } (\mathbf{F})\, dA$$

where $\mathbf{N}$ is the so-called *normal vector field* to the curve C. The divergence theorem describes the expansion of a region in two distinct ways, on the left in terms of the flux across the boundary of the region and on the right in terms of the expansion at each point within the region.

Both Green's theorem and the divergence theorem have corresponding theorems involving surface integrals and volume integrals in three dimensions.

LAPLACE AND FOURIER TRANSFORMS

Laplace Transforms The Laplace transform is used to convert equations involving a time variable t into equations involving a frequency variable s. There are essentially three reasons for doing this: (1) higher-order differential equations may be converted to purely algebraic equations, which are more easily solved; (2) boundary conditions are easily handled; and (3) the method is well-suited to the theory associated with the **Nyquist stability criteria**.

Table 2.1.4 Laplace Transforms

$f(t)$	$F(s) = \mathscr{L}(f(t))$	Name of function
1. a	a/s	
2. 1	$1/s$	
3. $u_a(t) = \begin{cases} 0 & t < a \\ 1 & t > a \end{cases}$	e^{-as}/s	Heavyside or step function
4. $\delta_a(t) = u_a'(t)$	e^{-as}	Dirac or impulse function
5. e^{at}	$1/(s-a)$	
6. $(1/r)e^{-t/r}$	$1/(rs+1)$	
7. ke^{-at}	$k/(s+a)$	
8. $\sin at$	$a/(s^2+a^2)$	
9. $\cos at$	$s/(s^2+a^2)$	
10. $e^{at} \sin bt$	$b/[(s+a)^2+b^2]$	
11. $\dfrac{e^{at}}{a+b} - \dfrac{e^{bt}}{a+b}$	$\dfrac{1}{(s-a)(s-b)}$	
12. t	$1/s^2$	
13. t^2	$2/s^3$	
14. t^n	$n!/s^{n+1}$	
15. t^a	$\Gamma(a+1)/s^{a+1}$	Gamma function (see ''Special Functions'')
16. $\sinh at$	$a/(s^2-a^2)$	
17. $\cosh at$	$s/(s^2-a^2)$	
18. $t^n e^{at}$	$n!/(s-a)^{n+1}$	
19. $t \cos at$	$(s^2-a^2)/(s^2+a^2)^2$	
20. $t \sin at$	$2as/(s^2+a^2)^2$	
21. $\sin at - at \cos at$	$2a^3/(s^2+a^2)^2$	
22. $\arctan a/s$	$(\sin at)/t$	

Table 2.1.5 Properties of Laplace Transforms

$f(t)$	$F(s) = \mathscr{L}(f(t))$	Name of rule
1. $f(t)$	$\int_0^\infty e^{-st} f(t)\, dt$	Definition
2. $f(t) + g(t)$	$F(s) + G(s)$	Addition
3. $kf(t)$	$kF(s)$	Scalar multiples
4. $f'(t)$	$sF(s) - f(0^+)$	Derivative laws
5. $f''(t)$	$s^2F(s) - sf(0^+)$ $- f'(0^+)$	
6. $f'''(t)$	$s^3F(s) - s^2f(0^+)$ $- sf'(0^+) - f''(0^+)$	
7. $\int f(t)dt$	$(1/s)F(s)$ $+ (1/s) \int f(t)\, dt\vert_{0^+}$	Integral law
8. $f(bt)$	$(1/b)F(s/b)$	Change of scale
9. $e^{at}f(t)$	$F(s-a)$	First shifting
10. $f * g(t)$	$F(s)G(s)$	Convolution
11. $u_a(t)f(t-a)$	$F(s)e^{-at}$	Second shifting
12. $-tf(t)$	$F'(s)$	Derivative in s

In Laplace-transformation mathematics the following symbols and equations are used (Tables 2.1.4 and 2.1.5):

$f(t) =$ a function of time
$s =$ a complex variable of the form $(\sigma + j\omega)$
$F(s) =$ an equation expressed in the transform variable s, resulting from operating on a function of time with the Laplace integral
$\mathscr{L} =$ an operational symbol indicating that the quantity which it prefixes is to be transformed into the frequency domain
$f(0^+) =$ the limit from the positive direction of $f(t)$ as t approaches zero
$f(0^-) =$ the limit from the negative direction of $f(t)$ as t approaches zero

Therefore, $F(s) = \mathscr{L}[f(t)]$. The Laplace integral is defined as

$$\mathscr{L} = \int_0^\infty e^{-st}\, dt. \text{ Therefore, } \mathscr{L}[f(t)] = \int_0^\infty e^{-st} f(t)\, dt$$

Direct Transforms

EXAMPLE.

$$f(t) = \sin \beta t$$

$$\mathscr{L}[f(t)] = \mathscr{L}(\sin \beta t) = \int_0^\infty \sin \beta t\, e^{-st}\, dt$$

but $$\sin \beta t = \frac{e^{j\beta t} - e^{-j\beta t}}{2j} \qquad \text{where} \quad j^2 = -1$$

$$\begin{aligned}
\mathscr{L}\,(\sin \beta t) &= \frac{1}{2j} \int_0^\infty (e^{j\beta t} - e^{-j\beta t}) e^{-st}\, dt \\
&= \frac{1}{2j} \left(\frac{-1}{s - j\beta} \right) e^{(-s+j\beta)t} \Big|_0^\infty \\
&\quad - \frac{1}{2j} \left(\frac{-1}{s + j\beta} \right) e^{(-s-j\beta)t} \Big|_0^\infty \\
&= \frac{\beta}{s^2 + \beta^2}
\end{aligned}$$

Table 2.1.4 lists the transforms of common time-variable expressions.

Some special functions are frequently encountered when using Laplace methods.

The **Heavyside,** or **step, function** $u_a(t)$ sometimes written $u(t - a)$, is zero for all $t < a$ and 1 for all $t > a$. Its value at $t = a$ is defined differently in different applications, as 0, ½, or 1, or it is simply left undefined. In Laplace applications, the value of a function at a single point does not matter. The Heavyside function describes a force which is ''off'' until time $t = a$ and then instantly goes ''on.''

The **Dirac delta function,** or **impulse function,** $\delta_a(t)$, sometimes written $\delta(t - a)$, is the derivative of the Heavyside function. Its value is always zero, except at $t = a$, where its value is ''positive infinity.'' It is sometimes described as a ''point mass function.'' The delta function describes an impulse or an instantaneous transfer of momentum.

The derivative of the Dirac delta function is called the **dipole function**. It is less frequently encountered.

The **convolution** $f * g(t)$ of two functions $f(t)$ and $g(t)$ is defined as

$$f * g(t) = \int_0^t f(u)g(t - u)\,du$$

Laplace transforms are often used to solve differential equations arising from so-called *linear systems*. Many vibrating systems and electrical circuits are linear systems. If an input function $f_i(t)$ describes the forces exerted upon a system and a response or output function $f_o(t)$ describes the motion of the system, then the **transfer function** $T(s) = F_o(s)/F_i(s)$. Linear systems have the special property that the transfer function is independent of the input function, within the elastic limits of the system. Therefore,

$$\frac{F_o(s)}{F_i(s)} = \frac{G_o(s)}{G_i(s)}$$

This gives a technique for describing the response of a system to a complicated input function if its response to a simple input function is known.

EXAMPLE. Solve $y'' + 2y' - 3y = 8e^t$ subject to initial conditions $y(0) = 2$ and $y'(0) = 0$. Let $y = f(t)$ and $Y = F(s)$. Take Laplace transforms of both sides and substitute for $y(0)$ and $y'(0)$, and get

$$s^2Y - 2s + 2(sY - 2) - 3Y = \frac{8}{s - 1}$$

Solve for Y, apply partial fractions, and get

$$\begin{aligned} Y &= \frac{2s^2 + 2s + 4}{(s + 3)(s - 1)^2} \\ &= \frac{1}{s + 3} + \frac{s + 1}{(s - 1)^2} \\ &= \frac{1}{(s + 3)} + \frac{1}{(s - 1)} + \frac{2}{(s - 1)^2} \end{aligned}$$

Using the tables of transforms to find what function has Y as its transform, we get

$$y = e^{-3t} + e^t + 2te^t$$

EXAMPLE. A vibrating system responds to an input function $f_i(t) = \sin t$ with a response $f_o(t) = \sin 2t$. Find the system response to the input $g_i(t) = \sin 2t$.

Apply the invariance of the transfer function, and get

$$\begin{aligned} G_o(s) &= \frac{F_oG_i}{F_i} \\ &= \frac{4(s^2 + 1)}{(s^2 + 4)^2} \\ &= \frac{2(2)}{s^2 + 2^2} - \frac{12}{16}\left[\frac{16}{(s^2 + 2^2)^2}\right] \end{aligned}$$

Applying formulas 8 and 21 from Table 2.1.4 of Laplace transforms,

$$g_o(t) = 2 \sin 2t - \tfrac{3}{4} \sin 2t + \tfrac{3}{2}\, t \cos 2t$$

Inversion When an equation has been transformed, an explicit solution for the unknown may be directly determined through algebraic manipulation. In automatic-control design, the equation is usually the differential equation describing the system, and the unknown is either the output quantity or the error. The solution gained from the transformed equation is expressed in terms of the complex variable s. For many design or analysis purposes, the solution in s is sufficient, but in some cases it is necessary to retransform the solution in terms of time. The process of passing from the complex-variable (frequency domain) expression to that of time (time domain) is called an **inverse transformation**. It is represented symbolically as

$$\mathscr{L}^{-1}F(s) = f(t)$$

For any $f(t)$ there is only one direct transform, $F(s)$. For any given $F(s)$ there is only one inverse transform $f(t)$. Therefore, tables are generally used for determining inverse transforms. Very complete tables of inverse transforms may be found in Gardner and Barnes, ''Transients in Linear Systems.'' As an example of the inverse procedure consider an equation of the form

$$K = \alpha x(t) + \int \frac{x(t)}{\beta}\,dt$$

It is desired to obtain an expression for $x(t)$ resulting from an instantaneous change in the quantity K. Transforming the last equation yields

$$\frac{K}{s} = X(s)\alpha + \frac{X(s)}{s\beta} + \frac{f^{-1}(0^+)}{s}$$

If $f^{-1}(0)/s = 0$

then $$X(s) = \frac{K/\alpha}{s + 1/\alpha\beta}$$

$$x(t) = \mathscr{L}^{-1}[X(s)] = \mathscr{L}^{-1}\frac{K/\alpha}{s + 1/\alpha\beta}$$

From Table 2.1.4, $x(t) = \dfrac{K}{\alpha}e^{-t/\alpha\beta}$

Fourier Coefficients Fourier coefficients are used to analyze periodic functions in terms of sines and cosines. If $f(x)$ is a function with period $2L$, then the Fourier coefficients are defined as

$$a_n = \frac{1}{L}\int_{-L}^{L} f(s) \cos\frac{n\pi s}{L}\,ds \qquad n = 0, 1, 2, \ldots$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(s) \sin\frac{n\pi s}{L}\,ds \qquad n = 1, 2, \ldots$$

Then the Fourier theorem states that

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)$$

The series on the right is called the ''Fourier series of the function $f(x)$.'' The convergence of the Fourier series is usually rapid, so that the function $f(x)$ is usually well-approximated by the sum of the first few sums of the series.

Examples of the Fourier Series If $y = f(x)$ is the curve in Figs. 2.1.110 to 2.1.112, then in Fig. 2.1.110,

$$y = \frac{h}{2} - \frac{4h}{\pi^2}\left(\cos\frac{\pi x}{c} + \frac{1}{9}\cos\frac{3\pi x}{c} + \frac{1}{25}\cos\frac{5\pi x}{c} + \cdots\right)$$

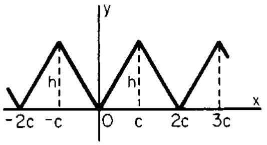


Fig. 2.1.110 Saw-tooth curve.

In Fig. 2.1.111,

$$y = \frac{4h}{\pi}\left(\sin\frac{\pi x}{c} + \frac{1}{3}\sin\frac{3\pi x}{c} + \frac{1}{5}\sin\frac{5\pi x}{c} + \cdots\right)$$

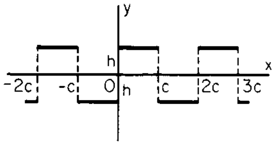


Fig. 2.1.111 Step-function curve.

In Fig. 2.1.112,

$$y = \frac{2h}{\pi}\left(\sin\frac{\pi x}{c} - \frac{1}{2}\sin\frac{2\pi x}{c} + \frac{1}{3}\sin\frac{3\pi x}{c} - \ . \ . \ .\right)$$

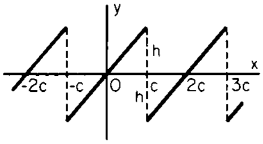


Fig. 2.1.112 Linear-sweep curve.

If the Fourier coefficients of a function $f(x)$ are known, then the coefficients of the derivative $f'(x)$ can be found, when they exist, as follows:

$$a_n' = nb_n \qquad b_n' = -na_n$$

where a_n' and b_n' are the Fourier coefficients of $f'(x)$.

The **complex Fourier coefficients** are defined by:

$$\begin{aligned} c_n &= \tfrac{1}{2}(a_n - ib_n) \\ c_0 &= \tfrac{1}{2}a_0 \\ c_n &= \tfrac{1}{2}(a_n + ib_n) \end{aligned}$$

Then the complex form of the Fourier theorem is

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L}$$

Wave Equation Fourier series are often used in the solution of the wave equation $a^2u_{xx} = u_{tt}$ where $0 < x < L$, $t > 0$, and initial conditions are $u(x, 0) = f(x)$ and $u_t(x, 0) = g(x)$. This describes the position of a vibrating string of length L, fixed at both ends, with initial position $f(x)$ and initial velocity $g(x)$. The constant a is the velocity at which waves are propagated along the string, and is given by $a^2 = T/p$, where T is the tension in the string and p is the mass per unit length of the string.

If $f(x)$ is extended to the interval $-L < x < L$ by setting $f(-x) = -f(x)$, then f may be considered periodic of period $2L$. Its Fourier coefficients are

$$a_n = 0 \qquad b_n = \int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,\pi\,dx \qquad n = 1, 2, \ . \ . \ .$$

The solution to the wave equation is

$$u(x, t) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}\cos\frac{n\pi t}{L}$$

Fourier transform A nonperiodic function $f(x)$ requires two functions to describe its Fourier transform:

$$A(w) = \int_{-\infty}^{\infty} f(x)\cos wx\,dx$$

$$B(w) = \int_{-\infty}^{\infty} f(x)\sin wx\,dx$$

Then the **Fourier integral equation** is

$$f(x) = \int_0^{\infty} A(w)\cos wx + B(w)\sin wx\,dw$$

The **complex Fourier transform** of $f(x)$ is defined as

$$F(w) = \int_{-\infty}^{\infty} f(x)\,e^{iwx}\,dx$$

Then the **complex Fourier integral equation** is

$$f(x) = \frac{1}{2\pi}\int_0^{\infty} F(w)e^{-iwx}\,dw$$

Heat Equation The Fourier transform may be used to solve the one-dimensional heat equation $u_t(x, t) = u_{xx}(x, t)$, for $t > 0$, given initial condition $u(x, 0) = f(x)$. Let $F(s)$ be the complex Fourier transform of $f(x)$, and let $U(s, t)$ be the complex Fourier transform of $u(x, t)$. Then the transform of $u_t(x, t)$ is $dU(s, t)/dt$.

Transforming $u_t(x, t) = u_{xx}(s, t)$ yields $dU/dt + s^2U = 0$ and $U(s, 0) = f(s)$. Solving this using the Laplace transform gives $U(s, t) = F(s)e^{s^2t}$.

Applying the complex Fourier integral equation, which gives $u(x, t)$ in terms of $U(s, t)$, gives

$$\begin{aligned} u(x, t) &= \frac{1}{2\pi}\int_0^{\infty} U(s, t)e^{-isx}\,ds \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty} f(y)e^{is(y-x)}e^{s^2t}\,ds\,dy \end{aligned}$$

Applying the Euler formula, $e^{ix} = \cos x + i\sin x$,

$$u(x, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty} f(y)\cos\left(s(y - x)\right)e^{s^2t}\,ds\,dy$$

SPECIAL FUNCTIONS

Gamma Function The gamma function is a generalization of the factorial function. It arises in Laplace transforms of polynomials, in continuous probability, and in the solution to certain differential equations. It is defined by the improper integral:

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\,dt$$

The integral converges for $x > 0$ and diverges otherwise. The function is extended to all negative values, except negative integers, by the relation

$$\Gamma(x + 1) = x\Gamma(x)$$

The gamma function is related to the factorial function by

$$\Gamma(n + 1) = n!$$

for all positive integers n.

An important value of the gamma function is

$$\Gamma(0.5) = \pi^{1/2}$$

Other values of the gamma function are found in CRC Standard Mathematical Tables and similar tables.

Beta Function The beta function is a function of two variables and is a generalization of the binomial coefficients. It is closely related to the gamma function. It is defined by the integral:

$$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt \qquad \text{for } x, y > 0$$

The beta function can also be represented as a trigonometric integral, by substituting $t = \sin^2 \theta$, as

$$B(x, y) = 2 \int_0^{\pi/2} (\sin \theta)^{2x-1} (\cos \theta)^{2y-1}\, d\theta$$

The beta function is related to the gamma function by the relation

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$

This relation shows that $B(x, y) = B(y, x)$.

Bernoulli Functions The Bernoulli functions are a sequence of periodic functions of period 1 used in approximation theory. Note that for any number x, $[x]$ represents the largest integer less than or equal to x. $[3.14] = 3$ and $[-1.2] = -2$. The Bernoulli functions $B_n(x)$ are defined recursively as follows:

1. $B_0(x) = 1$
2. $B_1(x) = x - [x] - ½$
3. B_{n+1} is defined so that $B'_{n+1}(x) = B_n(x)$ and so that B_{n+1} is periodic of period 1.

Bessel Functions of the First Kind Bessel functions of the first kind arise in the solution of Bessel's equation of order v:

$$x^2y'' + xy' + (x^2 - v^2)y = 0$$

When this is solved using series methods, the recursive relations define the Bessel functions of the first kind of order v:

$$J_v(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(v + k + 1)} \left(\frac{x}{2}\right)^{v+2k}$$

Chebyshev Polynomials The Chebyshev polynomials arise in the solution of PDEs of the form

$$(1 - x^2)y'' - xy' + n^2y = 0$$

and in approximation theory. They are defined as follows:

$$\begin{aligned} T_0(x) &= 1 \qquad & T_2(x) &= 2x^2 - 1 \\ T_1(x) &= x \qquad & T_3(x) &= 4x^3 - 3x \end{aligned}$$

For $n > 3$, they are defined recursively by the relation

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0$$

Chebyshev polynomials are said to be orthogonal because they have the property

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{(1 - x^2)^{1/2}}\, dx = 0 \qquad \text{for } n \neq m$$

NUMERICAL METHODS

Introduction Classical numerical analysis is based on polynomial approximation of the infinite operations of integration, differentiation, and interpolation. The objective of such analyses is to replace difficult or impossible exact computations with easier approximate computations. The challenge is to make the approximate computations short enough and accurate enough to be useful.

Modern numerical analysis includes Fourier methods, including the *f*ast *F*ourier *t*ransform (FFT) and many problems involving the way computers perform calculations. Modern aspects of the theory are changing very rapidly.

Errors Actual value = calculated value + error. There are several sources of errors in a calculation: mistakes, round-off errors, truncation errors, and accumulation errors.

Round-off errors arise from the use of a number not sufficiently accurate to represent the actual value of the number, for example, using 3.14159 to represent the irrational number π, or using 0.56 to represent 9⁄16 or 0.5625.

Truncation errors arise when a finite number of steps are used to approximate an infinite number of steps, for example, the first n terms of a series are used instead of the infinite series.

Accumulation errors occur when an error in one step is carried forward into another step. For example, if $x = 0.994$ has been previously rounded to 0.99, then $1 - x$ will be calculated as 0.01, while its true value is 0.006. An error of less than 1 percent is accumulated into an error of over 50 percent in just one step. Accumulation errors are particularly characteristic of methods involving recursion or iteration, where the same step is performed many times, with the results of one iteration used as the input for the next.

Simultaneous Linear Equations The matrix equation $Ax = b$ can be solved directly by finding A^{-1}, or it can be solved iteratively, by the method of **iteration in total steps:**

1. If necessary, rearrange the rows of the equation so that there are no zeros on the diagonal of A.
2. Take as initial approximations for the values of x_i:

$$x_1^{(0)} = \frac{b_1}{a_{11}} \qquad x_2^{(0)} = \frac{b_2}{a_{22}} \qquad \cdots \qquad x_n^{(0)} = \frac{b_n}{a_{nn}}$$

3. For successive approximations, take

$$x_i^{(k+1)} = (b_i - a_{i1}x_1^{(k)} - \cdots - a_{in}x_n^{(k)})/a_{ii}$$

Repeat step 3 until successive approximations for the values of x_i reach the specified tolerance.

A property of iteration by total steps is that it is self-correcting: that is, it can recover both from mistakes and from accumulation errors.

Zeros of Functions An iterative procedure for solving an equation $f(x) = 0$ is the **Newton-Raphson method**. The algorithm is as follows:

1. Choose a first estimate of a root x_0.
2. Let $x_{k+1} = x_k - f(x_k)/f'(x_k)$. Repeat step 2 until the estimate x_k converges to a root r.
3. If there are other roots of $f(x)$, then let $g(x) = f(x)/(x - r)$ and seek roots of $g(x)$.

False Position If two values x_0 and x_1 are known, such that $f(x_0)$ and $f(x_1)$ are opposite signs, then an iterative procedure for finding a root between x_0 and x_1 is the method of false position.

1. Let $m = [f(x_1) - f(x_0)]/(x_1 - x_0)$.
2. Let $x_2 = x_1 - f(x_1)/m$.
3. Find $f(x_2)$.
4. If $f(x_2)$ and $f(x_1)$ have the same sign, then let $x_1 = x_2$. Otherwise, let $x_0 = x_2$.
5. If x_1 is not a good enough estimate of the root, then return to step 1.

Functional Equalities To solve an equation of the form $f(x) = g(x)$, use the methods above to find roots of the equation $f(x) - g(x) = 0$.

Maxima One method for finding the maximum of a function $f(x)$ on an interval $[a, b]$ is to find the roots of the derivative $f'(x)$. The maximum of $f(x)$ occurs at a root or at an endpoint a or b.

Fibonacci Search An iterative procedure for searching for maxima works if $f(x)$ is unimodular on $[a, b]$. That is, f has only one maximum, and no other local maxima, between a and b. This procedure takes advantage of the so-called *golden ratio,* $r = 0.618034 = (\sqrt{5} - 1)/2$, which arises from the Fibonacci sequence.

1. If a is a sufficiently good estimate of the maximum, then stop. Otherwise, proceed to step 2.
2. Let $x_1 = ra + (1 - r)b$, and let $x_2 = (1 - r)a + rb$. Note $x_1 < x_2$. Find $f(x_1)$ and $f(x_2)$.
   a. If $f(x_1) = f(x_2)$, then let $a = x_1$ and $b = x_2$, and go to step 1.
   b. If $f(x_1) < f(x_2)$, then let $a = x_1$, and go to step 1.
   c. If $f(x_1) > f(x_2)$, then let $b = x_2$, and return to step 1.

In cases b and c, computation is saved since the new value of one of x_1 and x_2 will have been used in the previous step. It has been proved that the Fibonacci search is the fastest possible of the general ''cutting'' type of searches.

Steepest Ascent If $z = f(x, y)$ is to be maximized, then the method of steepest ascent takes advantage of the fact that the gradient, grad (f) always points in the direction that f is increasing the fastest.

1. Let (x_0, y_0) be an initial guess of the maximum of f.
2. Let e be an initial step size, usually taken to be small.
3. Let $(x_{k+1}, y_{k+1}) = (x_k, y_k) + e$ grad $f(x_k, y_k)/|$grad $f(x_k, y_k)|$.
4. If $f(x_{k+1}, y_{k+1})$ is not greater than $f(x_k, y_k)$, then replace e with $e/2$ (cut the step size in half) and reperform step 3.
5. If (x_k, y_k) is a sufficiently accurate estimate of the maximum, then stop. Otherwise, repeat step 3.

Minimization The theory of minimization exactly parallels the theory of maximization, since minimizing $z = f(x)$ occurs at the same value of x as maximizing $w = -f(x)$.

Numerical Differentiation In general, numerical differentiation should be avoided where possible, since differentiation tends to be very sensitive to small errors in the value of the function $f(x)$. There are several approximations to $f'(x)$, involving a ''step size'' h usually taken to be small:

$$f'(x) = \frac{f(x+h) - f(x)}{h}$$

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h}$$

$$f'(x) = \frac{f(x+2h) + f(x+h) - f(x-h) - f(x-2h)}{6h}$$

Other formulas are possible.

If a derivative is to be calculated from an equally spaced sequence of measured data, $y_1, y_2, \ldots, y_n$, then the above formulas may be adapted by taking $y_i = f(x_i)$. Then $h = x_{i+1} - x_i$ is the distance between measurements.

Since there are usually noise or measurement errors in measured data, it is often necessary to smooth the data, expecting that errors will be averaged out. Elementary smoothing is by simple averaging, where a value y_i is replaced by an average before the derivative is calculated. Examples include:

$$y_i \leftarrow \frac{y_{i+1} + y_i + y_{i-1}}{3}$$

$$y_i \leftarrow \frac{y_{i+2} + y_{i+1} + y_i + y_{i-1} + y_{i-2}}{5}$$

More information may be found in the literature under the topics linear filters, digital signal processing, and smoothing techniques.

Numerical Integration

Numerical integration requires a great deal of calculation and is usually done with the aid of a computer. All the methods described here, and many others, are widely available in packaged computer software. There is often a temptation to use whatever software is available without first checking that it really is appropriate. For this reason, it is important that the user be familiar with the methods being used and that he or she ensure that the error terms are tolerably small.

Trapezoid Rule If an interval $a \le x \le b$ is divided into subintervals $x_0, x_1, \ldots, x_n$, then the definite integral

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

may be approximated by

$$\sum_{i=1}^{n} [f(x_i) + f(x_{i-1})] \frac{x_{i+1} - x_i}{2}$$

If the values x_i are equally spaced at distance h and if f_i is written for $f(x_i)$, then the above formula reduces to

$$[f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n] \frac{h}{2}$$

The error in the trapezoid rule is given by

$$|E_n| \le \frac{(b-a)^3 |f''(t)|}{12n^2}$$

where t is some value $a \le t \le b$.

Simpson's Rule The most widely used rule for numerical integration approximates the curve with parabolas. The interval $a < x < b$ must be divided into $n/2$ subintervals, each of length $2h$, where n is an even number. Using the notation above, the integral is approximated by

$$[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{n-1} + f_n] \frac{h}{3}$$

The error term for Simpson's rule is given by $|E_n| < nh^5 |f^{(4)}(t)|/180$, where $a < t < b$.

Simpson's rule is generally more accurate than the trapezoid rule.

Ordinary Differential Equations

Modified Euler Method Consider a first-order differential equation $dy/dx = f(x, y)$ and initial condition $y = y_0$ and $x = x_0$. Take x_i equally spaced, with $x_{i+1} - x_i = h$. Then the method is:

1. Set $n = 0$.
2. $y'_n = f(x_n, y_n)$ and $y'' = f_x(x_n, y_n) + y'_n f_y(x_n, y_n)$, where f_x and f_y denote partial derivatives.
3. $y'_{n+1} = f(x_{n+1}, y_{n+1})$.

Predictor steps:

4. For $n > 0$, $y^*_{n+1} = y_{n-1} + 2hy'_n$.
5. $y'^*_{n+1} = f(x_{n+1}, y^*_{n+1})$.

Corrector steps:

6. $y^{\#}_n + 1 = y_n + [y^*_{n+1} + y'_n]h/2$.
7. $y'^{\#}_{n+1} = f(x_{n+1}, y^{\#}_{n+1})$.
8. If required accuracy is not yet obtained for y_{n+1} and y'_{n+1}, then substitute $y^{\#}$ for y^*, in all its forms, and repeat the corrector steps. Otherwise, set $n = n + 1$ and return to step 2.

Other predictor-corrector methods are described in the literature.

Runge-Kutta Methods These make up a family of widely used methods for ordinary differential equations. Given $dy/dx = f(x, y)$ and $h =$ interval size, third-order method (error proportional to h^4):

$$k_0 = hf(x_n)$$

$$k_1 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_0}{2}\right)$$

$$k_2 = hf(x_n + h, y_n + 2k_1 - k_0)$$

$$y_{n+1} = y_n + \frac{k_0 + 4k_1 + k_2}{6}$$

Higher-order Runge-Kutta methods are described in the literature. In general, higher-order methods yield smaller error terms.

2.2 COMPUTERS

by George J. Moshos

REFERENCES: Manuals from Computer Manufacturers. Knuth, ''The Art of Computer Programming,'' vols 1, 2, and 3, Addison-Wesley. Yourdon and Constantine, ''Structured Design,'' Prentice-Hall. DeMarco, ''Structured Analysis and System Specification,'' Prentice-Hall. Moshos, ''Data Communications,'' West Publishing. Date, ''An Introduction to Database Systems,'' 4th ed., Addison-Wesley. Wiener and Sincovec, ''Software Engineering with Modula-2 and ADA,'' Wiley. Hamming, ''Numerical Methods for Scientists and Engineers,'' McGraw-Hill. Bowers and Sedore, ''SCEPTRE: A Computer Program for Circuit and System Analysis,'' Prentice-Hall. Tannenbaum, ''Operating Systems,'' Prentice-Hall. Lister, ''Fundamentals of Operating Systems, 3d ed., Springer-Verlag. American National Standard Programming Language FORTRAN, ANSI X3.198-1992. Jensen and Wirth, ''PASCAL: User Manual and Report,'' Springer. *Communications, Journal,* and *Computer Surveys,* ACM Computer Society. *Computer, Spectrum,* IEEE.

COMPUTER PROGRAMMING

Machine Types

Computers are machines used for automatically processing information represented by mechanical, electrical, or optical means. They may be classified as analog or digital according to the techniques used to represent and process the information. Analog computers represent information as physically measurable, continuous quantities and process the information by components that have been interconnected to form an analogous model of the problem to be solved. Digital computers, on the other hand, represent information as discrete physical states which have been encoded into symbolic formats, and process the information by sequences of operational steps which have been preplanned to solve the given problem.

When compared to analog computers, digital computers have the advantages of greater versatility in solving scientific, engineering, and commercial problems that involve numerical and nonnumerical information; of an accuracy dictated by significant digits rather than that which can be measured; and of exact reproducibility of results that stay unvitiated by small, random fluctuations in the physical signals. In the past, multiple-purpose analog computers offered advantages of speed and cost in solving a sophisticated class of complex problems dealing with networks of differential equations, but these advantages have disappeared with the advances in solid-state computers. Other than the occasional use of analog techniques for embedding computations as part of a larger system, digital techniques now account almost exclusively for the technology used in computers.

Digital information may be represented as a series of incremental, numerical steps which may be manipulated to position control devices using stepping motors. Digital information may also be encoded into symbolic formats representing digits, alphabetic characters, arithmetic numbers, words, linguistic constructs, points, and pictures which may be processed by a variety of mechanized operators. Machines organized in this manner can handle a more general class of both numerical and nonnumerical problems and so form by far the most common type of digital machines. In fact, the term *computer* has become synonymous with this type of machine.

Digital Machines

Digital machines consist of two kinds of circuits: memory cells, which effectively act to delay signals until needed, and logical units, which perform basic Boolean operations such as AND, OR, NOT, XOR, NAND, and NOR. Memory circuits can be simply defined as units where information can be stored and retrieved on demand. Configurations assembled from the Boolean operators provide the macro operators and functions available to the machine user through encoded instructions. A typical computer might house hundreds of thousands to millions of transistors serving one or the other of these roles.

Both data and the instructions for processing the data can be stored in memory. Each unit of memory has an address at which the contents can be retrieved, or ''read.'' The *read* operation makes the contents at an address available to other parts of the computer without destroying the contents in memory. The contents at an address may be changed by a *write* operation which inserts new information after first nullifying the previous contents. Some types of memory, called read-only memory **(ROM),** can be read from but not written to. They can only be changed at the factory.

Abstractly, the address and the contents at the address serve roles analogous to a variable and the value of the variable. For example, the equation $z = x + y$ specifies that the value of x added to the value of y will produce the value of z. In a similar way, the machine instruction whose format might be:

add, address 1, address 2, address 3

will, when executed, add the contents at address 1 to the contents at address 2 and store the result at address 3. As in the equation where the variables remain unaltered while the values of the variables may be changed, the addresses in the instruction remain unaltered while the contents at the address may change.

An essential property of a digital computer is that the sequence of instructions processed to solve a problem is executed without human intervention. When an operator manually controls the sequence of computation, the machine is called a calculator. This distinction between computer and calculator, however, is arbitrary and vague with modern machines. Modern calculators offer opportunity to program a series of operations which can be executed without any required intervention. On the other hand, the computer is often programmed to interrogate the operator for a response before continuing with the solution.

Computers differ from other kinds of mechanical and electrical machines in that computers perform work on information rather than on forces and displacements. A common form of information is numbers. Numbers can be encoded into a mechanized form and processed by the four rules of arithmetic ($+$, $-$, $\times$, $\div$). But numbers are only one kind of information that can be manipulated by the computer. Given an encoded alphabet, words and languages can be formed and the computer can be used to perform such processes as information storage and retrieval, translation, and editing. Given an encoded representation of points and lines, the computer can be used to perform such functions as drawing, recognizing, editing, and displaying graphs, patterns, and pictures.

Because computers have become easily accessible, engineers and scientists from every discipline have reformatted their professional activities to mechanize those aspects which can supplant human thought and decision. In this way, mechanical processes can be viewed as augmenting human physical skills and strength, and information processes can be viewed as augmenting human mental skills and intelligence.

COMPUTER DATA STRUCTURES

Binary Notation

Digital computers represent information by strings of digits which assume one of two values: 0 or 1. These units of information are called **bits,** a word contracted from the term *binary digits*. A string of bits may represent either numerical or nonnumerical information.

In order to achieve efficiency in handling the information, the computer groups the bits together into units containing a fixed number of bits which can be referenced as discrete units. By encoding and formatting these units of information, the computer can act to process them. Units of 8 bits, called **bytes,** are common. A byte can be used to encode the basic symbolic characters which provide the computer with input-output information such as the alphabet, decimal digits, punctuation marks, and special characters.

Bit groups may be organized into larger units of 4 bytes (32 bits) called **words,** or even larger units of 8 bytes called *double words;* and sometimes into smaller units of 2 bytes called *half words*. Besides encoding numerical information and other linguistic constructs, these units are used to encode a repertoire of machine instructions. Older machines and special-purpose machines may have other word sizes.

Computers process numerical information represented as binary numbers. The binary numbering system uses a positional notation similar to the decimal system. For example, the decimal number 596.37 represents the value $5 \times 10^2 + 9 \times 10^1 + 6 \times 10^0 + 3 \times 10^{-1} + 7 \times 10^{-2}$. The value assigned to any of the 10 possible digits in the decimal system depends on its position relative to the decimal point (a weight of 10 to zero or positive exponent is assigned to the digits appearing to the left of the decimal point, and a weight of 10 to a negative exponent is applied to digits to the right of the decimal point). In a similar manner, a binary number uses a radix of 2 and two possible digits: 0 and 1. The radix point in the positional notation separates the whole from the fractional part of the number, just as in the decimal system. The binary number 1011.011 represents a value $1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3}$.

The operators available in the computer for setting up the solution of a problem are encoded into the instructions of the machine. The instruction repertoire always includes the usual arithmetic operators to handle numerical calculations. These instructions operate on data encoded in the binary system. However, this is not a serious operational problem, since the user specifies the numbers in the decimal system or by mnemonics, and the computer converts these formats into its own internal binary representation.

On occasions when one must express a number directly in the binary system, the number of digits needed to represent a numerical value becomes a handicap. In these situations, a radix of 8 or 16 (called the **octal** or **hexadecimal** system, respectively) constitutes a more convenient system. Starting with the digit to the left or with the digit to the right of the radix point, groups of 3 or 4 binary digits can be easily converted to equivalent octal or hexadecimal digits, respectively. Appending nonsignificant 0s as needed to the rightmost and leftmost part of the number to complete the set of 3 or 4 binary digits may be necessary. Table 2.2.1 lists the conversions of binary digits to their equivalent octal and hexadecimal representations. In the hexadecimal system, the letters A through F augment the set of decimal digits to represent the digits for 10 through 15. The following examples illustrate the conversion between binary numbers and octal or hexadecimal numbers using the table.

binary number	011	011	110	101	.	001	111	100
octal number	3	3	6	5	.	1	7	4

binary number	0110	1111	0101	.	0011	1110
hexadecimal number	6	F	5	.	3	E

Formats for Numerical Data

Three different formats are used to represent numerical information internal to the computer: fixed-point, encoded decimal, and floating-point.

A word or half word in fixed-point format is given as a string of 0s and 1s representing a binary number. The program infers the position of the radix point (immediately to the right of the word representing integers, and immediately to the left of the word representing fractions). Algebraic numbers have several alternate forms: 1's complement, 2's complement, and signed-magnitude. Most often 1's and 2's complement forms are adopted because they lead to a simplification in the hardware needed to perform the arithmetic operations. The sign of a 1's complement number can be changed by replacing the 0s with 1s and the 1s with 0s. To change the sign of a 2's complement number, reverse the digits as with a 1's-complement number and then add a 1 to the resulting binary number. Signed-magnitude numbers use the common representation of an explicit + or − sign by encoding the sign in the leftmost bit as a 0 or 1, respectively.

Table 2.2.1 Binary-Hexadecimal and Binary-Octal Conversion

Binary	Hexadecimal	Binary	Octal
0000	0	000	0
0001	1	001	1
0010	2	010	2
0011	3	011	3
0100	4	100	4
0101	5	101	5
0110	6	110	6
0111	7	111	7
1000	8		
1001	9		
1010	A		
1011	B		
1100	C		
1101	D		
1110	E		
1111	F		

Many computers provide an encoded-decimal representation as a convenience for applications needing a decimal system. Table 2.2.2 gives three out of over 8000 possible schemes used to encode decimal digits in which 4 bits represent each decade value. Many other codes are possible using more bits per decade, but four bits per decimal digit are common because two decimal digits can then be encoded in one byte. The particular scheme selected depends on the properties needed by the devices in the application.

The floating-point format is a mechanized version of the scientific notation ($\pm M \times 10^{\pm E}$, where $\pm M$ and $\pm E$ represent the signed mantissa and signed exponent of the number). This format makes possible the use of a machine word to encode a large range of numbers. The signed mantissa and signed exponent occupy a portion of the word. The exponent is implied as a power of 2 or 16 rather than of 10, and the radix point is implied to the left of the mantissa. After each operation, the machine adjusts the exponent so that a nonzero digit appears in the most significant digit of the mantissa. That is, the mantissa is **normalized** so that its value lies in the range of $1/b \leq M < 1$ where b is the implied base of the number system (e.g.: $1/2 \leq M < 1$ for a radix of 2, and $1/16 \leq M < 1$ for a radix of 16). Since the zero in this notation has many logical representations, the format uses a standard recognizable form for zero, with a zero mantissa and a zero exponent, in order to avoid any ambiguity.

When calculations need greater precision, floating-point numbers use

Table 2.2.2 Schemes for Encoding Decimal Digits

Decimal digit	BCD	Excess-3	4221 code
0	0000	0011	0000
1	0001	0100	0001
2	0010	0101	0010
3	0011	0110	0011
4	0100	0111	0110
5	0101	1000	1001
6	0110	1001	1100
7	0111	1010	1101
8	1000	1011	1110
9	1001	1100	1111

a two-word representation. The first word contains the exponent and mantissa as in the one-word floating point. Precision is increased by appending the extra word to the mantissa. The terms **single precision** and **double precision** make the distinction between the one- and two-word representations for floating-point numbers, although *extended precision* would be a more accurate term for the two-word form since the added word more than doubles the number of significant digits.

The equivalent decimal precision of a floating-point number depends on the number n of bits used for the unsigned mantissa and on the implied base b (binary, octal, or hexadecimal). This can be simply expressed in equivalent decimal digits p as: $0.0301\,(n - \log_2 b) < p < 0.0301\,n$. For example, a 32-bit number using 7 bits for the signed exponent of an implied base of 16, 1 bit for the sign of the mantissa, and 24 bits for the value of the mantissa gives a precision of 6.02 to 7.22 equivalent decimal digits. The fractional parts indicate that some 7-digit and some 8-digit numbers cannot be represented with a mantissa of 24 bits. On the other hand, a double-precision number formed by adding another word of 32 bits to the 24-bit mantissa gives a precision of 15.65 to 16.85 equivalent decimal digits.

The range r of possible values in floating-point notation depends on the number of bits used to represent the exponent and the implied radix. For example, for a signed exponent of 7 bits and an implied base of 16, then $16^{-64} \le r \le 16^{63}$.

Formats of Nonnumerical Data

Logical elements, also called **Boolean** elements, have two possible values which simply represent 0 or 1, true or false, yes or no, OFF or ON, etc. These values may be conveniently encoded by a single bit.

A large variety of codes are used to represent the alphabet, digits, punctuation marks, and other special symbols. The most popular ones are the 7-bit ASCII code and the 8-bit EBCDIC code. ASCII and EBCDIC find their genesis in punch-tape and punch-card technologies, respectively, where each character was encoded as a combination of punched holes in a column. Both have now evolved into accepted standards represented by a combination of 0s and 1s in a byte.

Figure 2.2.1 shows the ASCII code. (ASCII stands for American Standard Code for Information Interchange.) The possible 128 bit patterns divide the code into 96 graphic characters (although the codes 0100000 and 1111111 do not represent any printable graphic symbol) and 32 control characters which represent nonprintable characters used in communications, in controlling peripheral machines, or in expanding the code set with other characters or fonts. The graphic codes and the control codes are organized so that subsets of usable codes with fewer bits can be formed and still maintain the pattern.

	Bits		b_7	0	0	0	0	1	1	1	1
			b_6	0	0	1	1	0	0	1	1
			b_5	0	1	0	1	0	1	0	1
b_4	b_3	b_2	b_1								
0	0	0	0	<NUL>	<DLE>	<SP>	0	@	P	‘	p
0	0	0	1	<SOH>	<DC1>	!	1	A	Q	a	q
0	0	1	0	<STX>	<DC2>	″	2	B	R	b	r
0	0	1	1	<ETX>	<DC3>	#	3	C	S	c	s
0	1	0	0	<EOT>	<DC4>	$	4	D	T	d	t
0	1	0	1	<ENQ>	<NAK>	%	5	E	U	e	u
0	1	1	0	<ACK>	<SYN>	&	6	F	V	f	v
0	1	1	1	<BEL>	<ETB>	’	7	G	W	g	w
1	0	0	0	<BS>	<CAN>	(	8	H	X	h	x
1	0	0	1	<HT>	<EM>	)	9	I	Y	i	y
1	0	1	0	<LF>	<SUB>	*	:	J	Z	j	z
1	0	1	1	<VT>	<ESC>	+	;	K	[	k	{
1	1	0	0	<FF>	<FS>	,	<	L	\	l	|
1	1	0	1	<CR>	<GS>	-	=	M	]	m	}
1	1	1	0	<SO>	<RS>	.	>	N	ˆ	n	˜
1	1	1	1	<SI>	<US>	/	?	O	___	o	<DEL>

Fig. 2.2.1 ASCII code set.

Data Structure Types

The above types of numerical and nonnumerical data formats are recognized and manipulated by the hardware operations of the computer. Other more complex data structures may be programmed into the computer by building upon these primitive data types. The programmable data structures might include arrays, defined as ordered lists of elements of identical type; **sets,** defined as unordered lists of elements of identical type; **records,** defined as ordered lists of elements that need not be of the same type; **files,** defined as sequential collections of identical records; and **databases,** defined as organized collections of different records or file types.

COMPUTER ORGANIZATION

Principal Components

The principal components of a computer system shown schematically in Fig. 2.2.2 consist of a central processing unit (referred to as the **CPU** or platform), its working memory, an operator's console, file storage, and a collection of add-ons and peripheral devices. A computer system can be viewed as a library of collected data and packages of assembled sequences of instructions that can be executed in the prescribed order by the CPU to solve specific problems or perform utility functions for the users. These sequences are variously called programs, subprograms, routines, subroutines, procedures, functions, etc. Collectively they are called **software** and are directly accessible to the CPU through the working memory. The file devices act analogously to a bookshelf—they store information until it is needed. Only after a program and its data have been transferred from the file devices or from peripheral devices to the working memory can the individual instructions and data be addressed and executed to perform their intended functions.

The CPU functions to monitor the flow of data and instructions into and out of memory during program execution, control the order of instruction execution, decode the operation, locate the operand(s) needed, and perform the operation specified. Two characteristics of the memory and storage components dictate the roles they play in the computer system. They are **access time,** defined as the elapsed time between the instant a read or write operation has been initiated and the instant the

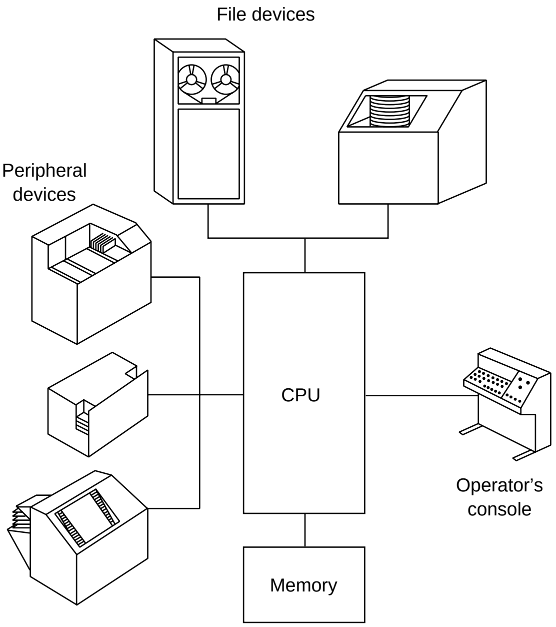


Fig. 2.2.2 Principal components of a computer system.

operation is completed, and size, defined by the number of bytes in a module. The faster the access time, the more costly per bit of memory or storage, and the smaller the module. The principal types of memory and storage components from the fastest to the slowest are registers which operate as an integral part of the CPU, cache and main memory which form the working memory, and mass and archival storage which serve for storing files.

The interrelationships among the components in a computer system and their primary performance parameters will be given in context in the following discussion. However, hundreds of manufacturers of computers and computer products have a stake in advancing the technology and adding new functionality to maintain their competitive edge. In such an environment, no performance figures stay current. With this caveat, performance figures given should not be taken as absolutes but only as an indication of how each component contributes to the performance of the total system.

Throughout the discussion (and in the computer world generally), prefixes indicating large numbers are given by the symbols k for kilo (10^3), M for mega (10^6), G for giga (10^9), and T for tera (10^{12}). For memory units, however, these symbols have a slightly altered meaning. Memories are organized in binary units whereby powers of two form the basis for all addressing schemes. According, k refers to a memory size of 1024 (2^{10}) units. Similarly M refers to 1024^2 (1,048,576), G refers to 1024^3, and T refers to 1024^4. For example, 1-Mbyte memory indicates a size of 1,048,576 bytes.

Memory

The main memory, also known as random access memory **(RAM),** is organized into fixed size bit cells (words, bytes, half words, or double words) which can be located by address and whose contents contain the instructions and data currently being executed. Typically RAM modules come in sizes of 1 to 10 Mbytes. The CPU acts to address the individual memory cells during program execution and transfers their contents to and from its internal registers.

Optionally, the working memory may contain an auxiliary memory, called **cache,** which is faster than the main memory. Cache operates on the premise that data and instructions that will shortly be needed are located near those currently being used. If the information is not found in the cache, then it is transferred from the main memory. Transfer rates between the cache and main memory are very fast and are usually made in block sizes of 16 to 64 bytes. Transfers between the cache and the registers are usually made on a word basis. Typically, cache modules come in sizes of a few kbytes to 1 Mbyte. The effective average access times offered by the combined configuration of RAM and cache results in a more powerful (faster) computer.

Central Processing Unit

The CPU makes available a repertoire of instructions which the user uses to set up the problem solutions. Although the specific format for instructions varies among machines, the following illustrates the pattern:

name: operator, operand(s)

The name designates an address whose contents contain the operator and one or more operands. The operator encodes an operation permitted by the hardware of the CPU. The operand(s) refer to the entities used in the operation which may be either data or another instruction specified by address. Some instructions have implied operand(s) and use the bits which would have been used for operand(s) to modify the operator.

To begin execution of a program, the CPU first loads the instructions serially by address into the memory either from a peripheral device, or more frequently, from storage. The internal structure of the CPU contains a number of memory registers whose number, while relatively few, depend on the machine's organization. The CPU acts to transfer the instructions one at a time from memory into a designated register where the individual bits can be interpreted and executed by the hardware. The actions of the following steps in the CPU, known as the *fetch-execute cycle,* control the order of instruction execution.

Step 1: Manually or automatically under program control load the address of the starting instruction into a register called the program register (PR).

Step 2: Fetch and copy the contents at the address in PR into a register called the program content register (PCR).

Step 3: Prepare to fetch the next instruction by augmenting PR to the next address in normal sequence.

Step 4: Interpret the instruction in PCR, retrieve the operands, execute the encoded operation, and then return to step 2. Note that the executed instruction may change the address in PR to start a different instruction sequence.

The speed of machines can be compared by calculating the average execution time of an instruction. Table 2.2.3 illustrates a typical instruction mix used in calculating the average. The instruction mix gives the relative frequency each instruction appears in a compiled list of typical programs and so depends on the types of problems one expects the machine to solve (e.g., scientific, commercial, or combination). The equation

$$t = \sum_i w_i t_i$$

expresses the average instruction execution time t as a function of the execution time t_i for instruction i having a relative frequency w_i in the instruction mix. The reciprocal of t measures the processor's performance as the average number of instructions per second (ips) it can execute.

Table 2.2.3 Instruction Mix

i	Instruction type	Weight w_i
1	Add: Floating point	0.07
2	Fixed point	0.16
3	Multiple: Floating point	0.06
4	Load/store register	0.12
5	Shift: One character	0.11
6	Branch: Conditional	0.21
7	Unconditional	0.17
8	Move 3 words in memory	0.10
	Total	1.00

For machines designed to support scientific and engineering calculations, the floating-point arithmetic operations dominate the time needed to execute an average instruction mix. The speed for such machines is given by the average number of floating-point operations which can be executed per second (flops). This measure, however, can be misleading in comparing different machine models. For example, when the machine has been configured with a cluster of processors cooperating through a shared memory, the rate of the configuration (measured in flops) represents the simple sum of the individual processors' flops rates. This does not reflect the amount of parallelism that can be realized within a given problem.

To compare the performance of different machine models, users often assemble and execute a suite of programs which characterize their particular problem load. This idea has been refined so that in 1992 two suites of **benchmark** programs representing typical scientific, mathematical, and engineering applications were standardized: Specint92 for integer operations, and Specfp92 for floating-point operations. Performance ratings for midsized computers are often reported in units calculated by a weighted average of the processing rates of these programs.

Computer performance depends on a number of interrelated factors used in their design and fabrication, among them compactness, bus size, clock frequency, instruction set size, and number of coprocessors.

The speed that energy can be transmitted through a wire, bounded theoretically at 3 × 10^{10} cm/s, limits the ultimate speed at which the electronic circuits can operate. The further apart the electronic elements are from each other, the slower the operations. Advances in integrated circuits have produced compact microprocessors operating in the nanosecond range.

The microprocessor's bus size (the width of its data path, or the number of bits that can be sent simultaneously in parallel) affect its performance in two ways: by the number of memory cells that can be directly addressed, and by the number of bits each memory reference can fetch and process at a time. For example, a 16-bit microprocessor can reference 2^{16} 16-bit memory cells and process 16 bits at a time. In order to handle the individual bits, the number of transistors that must be packed into the microprocessor goes up geometrically with the width of the data path. The earliest microprocessors were 8-bit devices, meaning that every memory reference retrieved 8 bits. To retrieve more bits, say 16, 32, or 64 bits, the 8-bit microprocessor had to make multiple references. Microprocessors have become more powerful as the packing technology has improved up to the 32-bit and 64-bit microprocessors currently available.

While normally the circuits operate asynchronously, a computer clock times the sequencing of the instructions. Clock speed is given in hertz (Hz, one cycle per second). Today's clock cycles are in the megahertz (MHz) range. Each instruction takes an integral number of cycles to complete, with one cycle being the minimum. If an instruction completes its operations in the middle of a cycle, the start of the next instruction must wait for the beginning of the next cycle.

Two schemes are used to implement the computer instruction set in the microprocessors. The more traditional complex instruction set computer (CISC) microprocessors implement by hard-wiring some 300 instruction types. Strange to say, the faster alternate-approach reduced instruction set computer (RISC) implements only about 10 to 30 percent of the instruction types by hard wiring, and implements the remaining more-complex instructions by programming them at the factory into read-only memory. Since the fewer hard-wired instructions are more frequently used and can operate faster in the simpler, more-compact RISC environment, the average instruction time is reduced.

To achieve even greater effectiveness and speed calls for more complex coordination in the execution of instructions and data. In one scheme, several microprocessors in a cluster share a common memory to form the machine organization (a multiprocessor or parallel processor). The total work which may come from a single program or independent programs is parceled out to the individual machines which operate independently but are coordinated to work in parallel with the other machines in the cluster. Faster speeds can be achieved when the individual processors can work on different parts of the problem or can be assigned to those parts of the problem solution for which they have been especially designed (e.g., input-output operations or computational operations). Two other schemes, pipelining and array processing, divide an instruction into the separate tasks that must be performed to complete its execution. A pipelining machine executes the tasks concurrently on consecutive pieces of data. An array processor executes the tasks of the different instructions in a sequence simultaneously and coordinates their completion (which might mean abandoning a partially completed instruction if it had been initiated prematurely). These schemes are usually associated with the larger and faster machines.

Operator's Console

The system operator uses the console to initiate or terminate computing tasks, to interrogate the computer to determine the status of the tasks during execution, to give and receive instructions such as mounting a particular file onto a drive or provide operating parameters during operations, and to otherwise monitor the system.

The operator's console consists of a relatively slow-speed keyboard input and a monitor display. The monitor display consists of a video scope which might be simply two-tone or could have a selection of colors or shades (up to 256) to build pictures and icons. Other important scope characteristics are the size of the screen and the resolution measured in points on the screen called **pixels.** The total number of pixels is given by the number of pixels on a horizontal line and the number of pixels on a vertical line (e.g., 1024 × 768 or 1600 × 1200). The scope has its own memory which refreshes and controls the display. For convenience and manual speed, a device called a **mouse** can be attached to the console and rolled on a flat surface which in turn moves the **cursor** on the display. This can be used to locate and select options displayed as a menu on the screen. A mouse turned upside down so the ball can be turned by the thumb performs the same function and is called a **trackball.**

File Devices

File devices serve to store libraries of directly accessible programs and data in electronically or optically readable formats. The file devices record the information in large blocks rather than by individual addresses. To be used, the blocks must first be transferred into the working memory. Depending on how selected blocks are located, file devices are categorized as sequential or direct-access. On **sequential** devices the computer locates the information by searching the file from the beginning. Direct-access devices, on the other hand, position the read-write mechanism directly at the location of the needed information. Searching on these devices works concurrently with the CPU and the other devices making up the computer configuration.

Magnetic tapes using arbitrary block sizes form commercial sequential-access products. Besides the disadvantage that the medium must be passed over sequentially to locate the beginning of the needed information, magnetic tape recording does not permit information to be changed in situ. Information can be changed only by reading the information from one tape, making the changes, and writing the changed information onto another tape.

Traditional magnetic tape recorders consist of reels of tape ½ in (12.7 mm) wide, 0.0015 in (0.0381 mm) thick, and 2400 ft (732 m) long. Information is recorded across the tape in 9-bit frames. One bit in each frame, called a **parity bit,** is used for checking purposes and is not transferred into the memory of the computer. The remaining 8 bits record the information using some standard format (e.g.: EBCDIC, modified ASCII, or an internal binary format). Lengthwise the information is recorded using standard densities such as 9600 bits/in, with gaps between blocks sufficient in size to stop the tape transport at the end of a block and before the beginning of the next block. Today's tape units use ½-in, 8-mm, or ¼-in cartridges that have a capacity up to 2.5 Tbytes of uncompacted data or 7.2 Tbytes of compacted data.

Magnetic or **optical disks** that offer a wide choice of options form the commercial direct-access devices. The recording surface consists of a platter (or platters) of recording material mounted on a common spindle rotated at high speed. The read-write heads may be permanently positioned along the radius of the platter or may be mounted on a common arm that can be moved radially to locate any specified track of information. Information is recorded on the **tracks** circumferentially using fixed-size blocks called pages or sectors. **Pages** divide the storage and memory space alike into blocks of 4096 bytes so that program transfers can be made without creating unusable space. **Sectors** nominally describe the physical division of the storage space into equal segments for easier positioning of the read-write heads.

The access time for retrieving information from a disk depends on three separately quoted factors, called seek time, latency time, and transfer time. **Seek time** gives the time needed to position the read-write heads from their current track position to the track containing the information. Average seek time is on the order of 100 ms. Since the faster fixed-head disks require no radial motion, only latency and transfer time need to be factored into the total access time for these devices. **Latency time** is the time needed to locate the start of the information along the circumferential track. This time depends on the speed of revolution of the disk, and, on average, corresponds to the time required to revolve the platter half a turn. The average latency time can be reduced by repeating the information several times around the track. Average latency time is on the order of 2 to 20 ms. **Transfer time,** usually quoted as a rate, gives the rate at which information can be transferred to memory after it has been located. There is a large variation in transfer rates depending on the disk system selected. Typical systems range from 20 kbytes/s to 20 Mbytes/s.

Disk devices are called soft or hard disks, referring to the rigidity of the platter. **Soft disks,** also called **floppy disks,** have a mountable, small, single platter that provides one or two recording surfaces. Soft or floppy might be a misnomer since many systems use diskettes about the size and rigidity of a credit card. Typical floppies have a physical size of 5¼ in or 3½ in and have a capacity of 1.2 Mbytes and 1.44 Mbytes, respectively. **Hard disks** refer to sealed devices whose physical size has been reduced to units of 1.3 to 2.5 in (33 to 63.5 mm), yet their capacity has increased. For example, disk storage of 200 Mbytes is available for small computers, and for more complex systems an array of disks is available having a capacity of from over 500 Mbytes to nearly 2 Gbytes.

Computer architects sometimes refer to file storage as **mass storage** or **archival storage,** depending on whether or not the libraries can be kept off-line from the system and mounted when needed. Disk drives with mountable platter(s) and tape drives constitute the archival storage. Sealed disks that often have fixed heads for faster access are the medium of choice for mass storage.

Peripheral Devices and Add-ons

Peripheral devices function as self-contained external units that work on line to the computer to provide or receive information or to control the flow of information. Add-ons are a special class of units whose circuits can be integrated into the circuitry of the computer hardware to augment the basic functionality of the processors. Section 15 covers the electronic technology associated with these devices.

An input device may be defined as any device that provides a machine-readable source of information. For engineering work, the most common forms of input are punched cards, punched tape, magnetic tape, magnetic ink, touch-tone dials, mark sensing, bar codes, and keyboards (usually in conjunction with a printing mechanism or video scope). Many bench instruments have been reconfigured to include digital devices to provide direct input to computers. Because of the data-handling capabilities of the computer, these instruments can be simpler, smaller, and less expensive than the hand instruments they replace. New devices have also been introduced: devices for visual measurement of distance, area, speed, and coordinate position of an object; or for inspecting color or shades of gray for computer-guided vision. Other methods of input that are finding greater acceptance include handwriting recognition, printed character recognition, voice digitizers, and picture digitizers.

Traditionally, output devices play the role of producing displays for the interpretation of results. A large variety of printers, graphical plotters, video displays, and audio sets have been developed for this purpose. Printers are distinguished by:

Type of print head (letter-quality or dot-matrix)
Type of paper feed (tractor or friction)
Allowable paper sizes
Print control (character, line, or page at a time)
Speed (measured in characters, lines, or pages per minute)
Number of fonts (especially for laser printers)

Graphic plotters and video displays offer variations in size, color capabilities, and quality. The more sophisticated video scopes offer dynamic characteristics capable of animated displays.

A variety of actuators have been developed for driving control mechanisms. Typical developments are in high-precision rack-and-pinion mechanisms and in lead screws that essentially eliminate backlash due to gear trains. For complex numerical control, programmable controllers (called PLCs) can simultaneously control and update data from multiple tasks. These electronically driven mechanisms and controllers, working with input devices, make possible systems for automatic testing of products, real-time control, and robotics with learning and adaptive capabilities.

Computer Sizes

Computer size refers not only to the physical size but also to the number of electronics elements in the system, and so reflects the performance of the system. Between the two ends of the spectrum from the largest and fastest to the smallest and slowest are machines that vary in speed and complexity. Although no nomenclature has been universally adopted that indicates computer size, the following descriptions illustrate a few generally understood terms used for some common configurations.

Personal computers (PCs) have been made possible by the advances in solid-state technology. The name applies to computers that can fit the total complement of hardware on a desktop and operate as stand-alone systems so as to provide immediate dedicated services to an individual user. This popular computer size has been generally credited for spreading computer literacy in today's society. Because of its commercial success, many peripheral devices, add-ons, and software products have been (and are continually being) developed. Laptop PCs are personal computers that have the low weight and size of a briefcase and can easily be transported when peripherals are not immediately needed.

The term **workstation** describes computer systems which have been designed to support complex engineering, scientific, or business applications in a professional environment. Although a top-of-the-line PC or a PC connected as a peripheral to another computer can function like a workstation, one can expect a machine designed as a workstation to offer higher performance than a PC and to support the more specialized peripherals and sophisticated professional software. Nevertheless, the boundary between PCs and workstations changes as the technology advances. Table 2.2.4 lists some published performance values for the spectrum of computers which have been designated as workstations. The spread in speed values represents the statistical average of reported samples distributed over one standard deviation.

Notebook PCs and the smaller sized **palmtop** PCs are portable, battery-operated machines. A typical notebook PC size would be 9 × 11 in (230 × 280 mm) in area, 1 to 2 in (25 to 50 mm) thick, and 2 to 9 lb (1 to 4 kg) in weight. They often have built-in programs stored in ROM. Having 68-pin integrated circuit cards for mass memory that can store as much as some hard disks, and being able to share programs with desktop PCs, these machines find excellent use as portable PCs in some applications and as data acquisition systems. However their undersized keyboards and small scopes limit their usefulness for sustained operations.

Table 2.2.4 Reported Performance Parameters for Workstations

	Workstation range		
	Low	Mid	High
Processor			
Clock speed, MHz	20–33	40–80	100–200
Bus size	16–32	32	64
Number of coprocessors	1–2	1–2	1–4
Instruction set	CISC		RISC
Speed rating			
Specint92	17.1–25.1	32.3–55.7	38.1–77.1
Specfp92	21.2–26.4	43.9–81.9	52.0–120.0
Mips	20.6–36.4	21.9–92.1	86.6–135.4
Mflops	2.6–6.0	4.3–20.9	30.0–50.0
Memory capacity			
Main, Mbytes	2–128		16–128
Cache, kbytes	8–128		64–256
Disk capacity			
Hard, Mbytes	10–80	80–200	200–400
Floppy, Mbytes	1.44		1.44

Computers larger than a PC or a workstation, called **mainframes** (and sometimes minis or maxis, depending on size), serve to support multiusers and multiapplications. A remotely accessible computing center may house several mainframes which either operate alone or cooperate with each other. Their high speed and large memories allow them to handle complex programs. A specific type of mainframe, used to maintain the database of a system, is called a **database machine.** Database machines act in cooperation with a number of user stations in a server-client relationship. In this, the database machine (the server) provides the data and/or the programs and shares the processing with the individual workstations (the clients).

At the upper extreme end of the computer spectrum is the **supercomputer**, the class of the fastest machines that can address large, complex scientific/engineering problems which cannot reasonably be transferred to other machines. Obviously this class of computer must have cache and main memory sizes and speeds commensurate with the speed of the platform. While mass memory sizes must also be large, computers which support large databases may often have larger memories than do supercomputers. Large, complex technical problems must be run with high-precision arithmetic. Because of this, performance is measured in double-precision flops. Supercomputer performance has moved from the current range of 10 Gflops into the Tflops range. To realize these speeds, the designers of supercomputers work at the edge of the available technology, especially in the use of multiple processors operating in parallel. Current clusters of 4 to 16 processors are being expanded to a goal of 100 and more. With multiple processors, however, performance depends as much on the time spent in communication between processors as on the computational speed of the individual processors. In the final analysis, to muster the supercomputer's inherent speed, the development of the software becomes the problem. Some users report that software must often be hand-tailored to the specific problem. The power of the machines, however, can replace years of work in analysis and experimentation.

DISTRIBUTED COMPUTING

Organization of Data Facilities

A distributed computer system can be defined as a collection of computer resources which are remotely located from each other and are interconnected to cooperate in providing their respective services. The resources include both the equipment and the software. Resources distributed to reside near the vicinity where the data is collected or used have an obvious advantage over centralization. But to provide information in a timely and reliable manner, these islands of automation must be integrated.

The size and complexity of an enterprise served by a distributed information system can vary from a single-purpose office to a multiple-plant conglomerate. An enterprise is defined as a system which has been created to accomplish a mission in its environment and whose goals involve risk. Internally it consists of organized functions and facilities which have been prepared to provide its services and accomplish its mission. When stimulated by an external entity, the enterprise acts to produce its planned response. An enterprise must handle both the flow of material (goods) and the flow of information. The information system tracks the material in the material system, but itself handles only the enterprise's information.

The technology for distributing and integrating the total information system comes under the industrial strategy known as computer-integrated business (CIB) or computer-integrated manufacturing (CIM). The following reasons have been cited for developing CIB and CIM:

Most data generated locally has only local significance.

Data integrity resides where it is generated.

The quality and consistency of operational decisions demands not only that all parts of the system work with the same data but that they can receive it in a reliable and timely manner.

If a local processor fails, it may disrupt local operations, but the remaining system should continue to function independently.

Small cohesive processors can be best managed and maintained locally.

Through standards, selection of local processes can be made from the best products in a competitive market that can be integrated into the total system.

Obsolete processors can be replaced by processors implemented by more advance technology that conform to standards without the cost of tailoring the products to the existing system.

Figure 2.2.3 depicts the total information system of an enterprise. The database consists of the organized collection of data the processors use in their operations. Because of differences in their communication requirements, the automated procedures are shown separated into those used in the office and those used on the production floor. In a business environment, the front office operations and back office operations make this separation. While all processes have a critical deadline, the production floor handles real-time operations, defined as processes which must complete their response within a critical deadline or else the results of the operations become moot. This places different constraints on the local-area networks (LANs) that serve the communication needs within the office or within the production floor. To communicate with entities outside the enterprise, the enterprise uses a wide-area network (WAN), normally made up from available public network facilities. For efficient and effective operation, the processes must be interconnected by the communications to share the data in the database and so integrate the services provided.

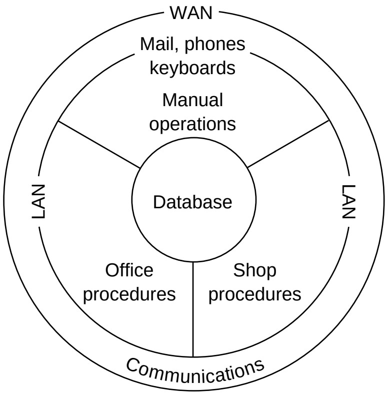


Fig. 2.2.3 Composite view of an enterprise's information system.

Communication Channels

A communication channel provides the connecting path for transmitting signals between a computing system and a remotely located application. Physically the channel may be formed by a wire line using copper, coaxial cable, or optical-fiber cable; or may be formed by a wireless line using radio, microwave, or communication satellites; or may be a combination of these lines.

Capacity, defined as the maximum rate at which information can be transmitted, characterizes a channel independent of the morphic line. Theoretically, an ideal noiseless channel that does not distort the signals has a channel capacity C given by:

$$C = 2W$$

where C is in pulses per second and W is the channel bandwidth. For digital transmission, the Hartley-Shannon theorem sets the capacity of a channel limited by the presence of gaussian noise such as the thermal noise inherent in the components. The formula:

$$C = W \log_2 (1 + S/N)$$

gives the capacity C in bits/s in terms of the signal to noise ratio S/N and the bandwidth W. Since the signal to noise ratio is normally given in decibels divisible by 3 (e.g., 12, 18, 21, 24) the following formula provides a workable approximation to the formula above:

$$C = W(S/N)_{\mathrm{db}}/3$$

where $(S/N)_{\mathrm{db}}$ is the signal-to-noise ratio expressed in decibels. Other forms of noise, signal distortions, and the methods of signal modulation reduce this theoretical capacity appreciably.

Nominal transmission speeds for electronic channels vary from 1000 bits to almost 20 Mbits per second. Fiber optics, however, form an almost noise-free medium. The transmission speed in fiber optics depends on the amount a signal spreads due to the multiple reflected paths it takes from its source to its destination. Advances in fiber technology have reduced this spread to give unbelievable rates. Effectively, the speeds available in today's optical channels make possible the transmission over a common channel, using digital techniques, of all forms of information: text, voice, and pictures.

Besides agreeing on speed, the transmitter and receiver must agree on the mode of transmission and on the timing of the signals. For stations located remotely from each other, transmission occurs by organizing the bits into groups and transferring them, one bit after another, in a serial mode. One scheme, called **asynchronous** or start-stop transmission, uses separate start and stop signals to frame a small group of bits representing a character. Separate but identical clocks at the transmitter and receiver time the signals. For transmission of larger blocks at faster rates, the stations use **synchronous** transmission which embeds the clock information within the transmitted bits.

Communication Layer Model

Figure 2.2.4 depicts two remotely located stations that must cooperate through communication in accomplishing their respective tasks. The communications substructure provides the communication services needed by the application. The application tasks themselves, however, are outside the scope of the communication substructure. The distinction here is similar to that in a telephone system which is not concerned with the application other than to provide the needed communication service. The figure shows the communication facilities packaged into a hierarchical modular layer architecture in which each node contains identical kinds of functions at the same layer level. The layer functions represent abstractions of real facilities, but need not represent specific hardware or software. The entities at a layer provide communication services to the layer above or can request the services available from the layer below. The services provided or requested are available only at service points which are identified by addresses at the boundaries that interface the adjacent layers.

The top and bottom levels of the layered structure are unique. The topmost layer interfaces and provides the communication services to the noncommunication functions performed at a node dealing with the application task (the user's program). This layer also requests communication services from the layer below. The bottom layer does not have a lower layer through which it can request communication services. This layer acts to create and recognize the physical signals transmitted between the bottom entities of the communicating partners (it arranges the actual transmission). The medium that provides the path for the transfer of signals (a wire, usually) connects the service access points at the bottom layers, but itself lies outside the layer structure.

Virtual communication occurs between peer entities, those at the same level. Peer-to-peer communication must conform to layer protocol, defined as the rules and conventions used to exchange information. Actual physical communication proceeds from the upper layers to the bottom, through the communication medium (wire), and then up through the layer structure of the cooperating node.

Since the entities at each layer both transmit and receive data, the protocol between peer layers controls both input and output data, depending on the direction of transmission. The transmitting entities accomplish this by appending control information to each data unit that they pass to the layer below. This control information is later interpreted and removed by the peer entities receiving the data unit.

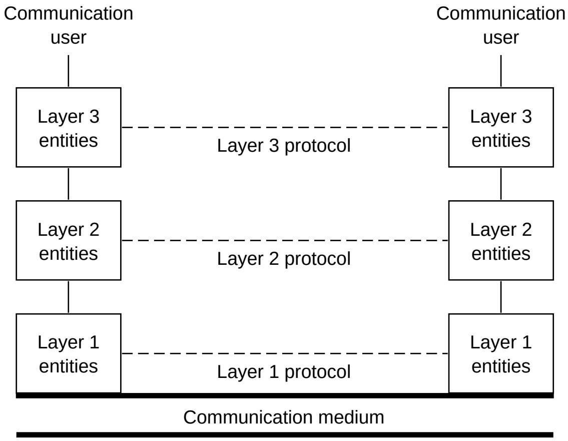


Fig. 2.2.4 Communication layer architecture.

Communication Standards

Table 2.2.5 lists a few of the hundreds of forums seeking to develop and adopt voluntary standards or to coordinate standards activities. Often users establish standards by agreement that fixes some existing practice. The ISO, however, has described a seven-layer model, called the *Reference Model for Open Systems Interconnection* (OSI), for coordinating and expediting the development of new implementation standards. The term *open systems* refers to systems that allow devices to be interconnected and to communicate with each other by conforming to common implementation standards. The ISO model is not of itself an implementation standard nor does it provide a basis for appraising existing implementations, but it partitions the communication facilities into layers of related function which can be independently standardized by different teams of experts. Its importance lies in the fact that both vendors and users have agreed to provide and accept implementation standards that conform to this model.

Table 2.2.5 Some Groups Involved with Communication Standards

CCITT	Comité Consultatif de Télégraphique et Téléphonique
ISO	International Organization for Standardization
ANSI	American National Standards Institute
EIA	Electronic Industries Association
IEEE	Institute of Electrical and Electronics Engineers
MAP/TOP	Manufacturing Automation Protocols and Technical and Office Protocols Users Group
NIST	National Institute of Standards and Technology

The following lists the names the ISO has given the layers in its ISO model together with a brief description of their roles.

Application layer provides no services to the other layers but serves as the interface for the specialized communication that may be required by the actual application, such as file transfer, message handling, virtual terminal, or job transfer.

Presentation layer relieves the node from having to conform to a particular syntactical representation of the data by converting the data formats to those needed by the layer above.

Session layer coordinates the dialogue between nodes including arranging several sessions to use the same transport layer at one time.

Transport layer establishes and releases the connections between peers to provide for data transfer services such as throughput, transit delays, connection setup delays, error rate control, and assessment of resource availability.

Network layer provides for the establishment, maintenance, and release of the route whereby a node directs information toward its destination.

Data link layer is concerned with the transfer of information that has been organized into larger blocks by creating and recognizing the block boundaries.

Physical layer generates and detects the physical signals representing the bits, and safeguards the integrity of the signals against faulty transmission or lack of synchronization.

The IEEE has formulated several implementation standards for office or production floor LANs that conform to the lower two layers of the ISO model. The functions assigned to the ISO data link layer have been distributed over two sublayers, a logical link control (LLC) upper sublayer that generates and interprets the link control commands, and a medium access control (MAC) lower sublayer that frames the data units and acquires the right to access the medium. From this structure, the IEEE has formulated three standards for the MAC sublayer and ISO physical layer combination, and a common standard for the LLC sublayer. The three standards for the bottom portion of the structure are named according to the method used to control the access to the medium: carrier sense multiple access with collision detection (CSMA/CD), token-passing bus access, and token ring access. A wide variety of options have been included for each of these standards which may be selected to tailor specific implementation standards.

CSMA/CD standardized the access method developed by the Xerox Corporation under its trademark Ethernet. The nodes in the network are attached to a common bus, schematically shown in Fig. 2.2.5*a*. All nodes hear every message transmitted, but accept only those messages addressed to themselves. When a node has a message to transmit, it listens for the line to be free of other traffic before it initiates transmission. However, more than one mode may detect the free line and may start to transmit. In this situation the signals will collide and produce a detectable change in the energy level present in the line. Even after a station detects a collision it must continue to transmit to make sure that all stations hear the collision (all data frames must be of sufficient length to be present simultaneously on the line as they pass each station). On hearing a collision, all stations that are transmitting wait a random length of time and then attempt to retransmit.

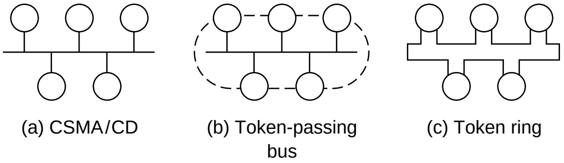


Fig. 2.2.5 LAN structures.

The stations in the **token-passing** bus access method, like the CSMA/CD method, share a common bus and communicate by broadcasting their messages to all stations. Unlike CSMA/CD, token-passing bus stations communicate in an ordered fashion as shown by the dashed line in Fig. 2.2.5*b*. By using special control frames the stations organize themselves into a logical ring by address (station 40 follows 30 which follows 20 which follows 40). The token is a special control frame which is circulated sequentially from station to station, giving the station that has the token the exclusive right to transmit any message it has ready for transmission. When a station has no message to transmit, or after it has completed transmission, it passes the token to the next station in the ring. The method features protocol procedures for restructuring the ring when ring membership changes, such as when a station intentionally or through failure leaves the ring, or a new station joins.

The **token-ring** access method connects the stations into a physical ring as shown in Fig. 2.2.5*c*. A special mechanical connector attaches the station equipment to the medium which when disconnected automatically closes the line to reestablish line continuity. The token has a priority level which may be changed by a station. When a station receives the token, it can start to circulate any data it has ready for transmission at the priority level of the token. As each station receives information from its neighbor, it regenerates the information and continues to circulate it around the ring while retaining a copy of everything destined for itself. The station that had originally sent the information retains the token until the information has been returned uncorrupted. Then it passes the token to the next station. Any station that had changed the priority level of the token has the responsibility for returning it to its previous level in a fair and orderly fashion. Protocol procedures sense failures in a station or faults in the medium.

The MAP/TOP (Manufacturing Automation Protocols and Technical and Office Protocols) Users Group started under the auspices of General Motors and Boeing Information Systems and now has a membership of many thousands of national and international corporations. The corporations in this group have made a commitment to open systems that will allow them to select the best products through standards, agreed to by the group, that will meet their respective requirements. In particular, MAP has standardized options from the IEEE token-passing bus method for production floor LAN implementation, and TOP has standardized options from the IEEE CSMA/CD for office LAN implementations. These standards have also been adopted by NIST for governmentwide use under the title Government Open Systems Interconnections Profile (GOSIP).

The Electronics Industries Association has established three interface standards, RS-232C, RS-422, and RS-423, which are frequently referenced for digital communications. These standards specify the use of multiple lines that interface the equipment at a station and the communication control equipment attached to the medium. **RS-232C** has been the primary standard for several years for low-speed voltage-oriented digital communications. RS-232C uses nonbalanced circuits sharing a common ground wire which, because of their sensitivity to noise, limits the bandwidth and length of the lines. RS-232C specifications call for a maximum line length of about 250 ft at a bandwidth of 10 kHz. **RS-423** also uses nonbalanced circuits but with individual ground wires which allows higher limits to a maximum line length of about 400 ft at a bandwidth of 100 kHz. **RS-422** uses balanced circuits with individual ground wires which allow line lengths up to 4000 ft at bandwidths of 100 kHz.

The common carriers who offer WAN communication services through their public networks have also developed **packet-switching** networks for public use. Packet switching transmits data in a purely digital format, which, when embellished, can replace the common circuit-switching technology used in analog communications such as voice. A packet is a fixed-sized block of digital data with embedded control information. The network serves to deliver the packets to their destination in an efficient and reliable manner.

CCITT has developed a set of standards, called X.25, for the three bottom ISO layers, to interface the public packet-switching networks. One of the set, named the X.21 standard, serves as a replacement to the EIA standards (RS-232C, RS-422, and RS-423) with fewer interconnecting lines whereby an expanded number of functions can be selected by coded digital means. When the equipment at the local site does not support the X.25 protocol, then a protocol converter interface, called a packet assembler/disassembler (PAD), properly structures the data for transmission over public packet-switching networks. While the upper four layers are not addressed by this interface, it is understood that end-to-end communication can take place only when the protocols between the layers at source and destination points agree or are made to conform through protocol converters.

RELATIONAL DATABASE TECHNOLOGY

Design Concepts

As computer hardware has evolved from small working memories and tape storage to large working memories and large disk storage, so has database technology moved from accessing and processing of a single, sequential file to that of multiple, random-access files. A **relational database** can be defined as an organized collection of interconnected tables or records. The records appear like the flat files of older technology. In each record the information is in columns (fields) which identify attributes, and rows (tuples) which list particular instances of the attributes. One column (or more), known as the **primary key,** identifies each row. Obviously, the primary key must be unique for each row.

If the data is to be handled in an efficient and orderly way, the records cannot be organized in a helter-skelter fashion such as simply transporting existing flat files into relational tables. To avoid problems in maintaining and using the database, redundancy should be eliminated by storing each fact at only one place so that, when making additions or deletions, one need not worry about duplicates throughout the database. This goal can be realized by organizing the records into what is known as the *third normal form*. A record is in the third normal form if and only if all nonkey attributes are mutually independent and fully dependent on the primary key.

The advantages of relational databases, assuming proper normalization, are:

Each fact can be stored exactly once.

The integrity of the data resides locally, where it is generated and can best be managed.

The tables can be physically distributed yet interconnected.

Each user can be given his/her own private view of the database without altering its physical structure.

New applications involving only a part of the total database can be developed independently.

The system can be automated to find the best path through the database for the specified data.

Each table can be used in many applications by employing simple operators without having to transfer and manipulate data superfluous to the application.

A large, comprehensive system can evolve from phased design of local systems.

New tables can be added without corrupting everyone's view of the data.

The data in each table can be protected differently for each user (read-only, write-only).

The tables can be made inaccessible to all users who do not have the right to know.

Relational Database Operators

A database system contains the structured collection of data, an on-line catalog and dictionary of data items, and facilities to access and use the data. The system allows users to:

Add new tables
Remove old tables
Insert new data into existing tables
Delete data from existing tables
Retrieve selected data
Manipulate data extracted from several tables
Create specialized reports

As might be expected, these systems include a large collection of operators and built-in functions in addition to those normally used in mathematics. Because of the similarity between database tables and mathematical sets, special set-like operators have been developed to manipulate tables. Table 2.2.6 lists eight typical table operators. The list of functions would normally also include such things as count, sum, average, find the maximum in a column, and find the minimum in a column. A rich collection of report generators offers powerful and flexible capabilities for producing tabular listings, text, graphics (bar charts, pie charts, point plots, and continuous plots), pictorial displays, and voice output.

SOFTWARE ENGINEERING

Programming Goals

Software engineering encompasses the methodologies for analyzing program requirements and for structuring programs to meet the requirements over their life cycle. The objectives are to produce programs that are:

Well documented
Easily read
Proved correct
Bug- (error-) free
Modifiable and maintainable
Implementable in modules

Control-Flow Diagrams

A control-flow diagram, popularly known as a **flowchart,** depicts all possible sequences of a program during execution by representing the control logic as a directed graph with labeled nodes. The theory associated with flowcharts has been refined so that programs can be structured to meet the above objectives. Without loss of generality, the nodes in a flowchart can be limited to the three types shown in Fig. 2.2.6. A **function** may be either a *transformer* which converts input data values into output data values or a *transducer* which converts that data's morphological form. A label placed in the rectangle specifies the function's action. A **predicate** node acts to bifurcate the path through the node. A

Table 2.2.6 Relational Database Operators

Operator	Input	Output
Select	A table and a condition	A table of all tuples that satisfy the given condition
Project	A table and an attribute	A table of all values in the specified attribute
Union	Two tables	A table of all unique tuples appearing in one table or the other
Intersection	Two tables	A table of all tuples the given tables have in common
Difference	Two tables	A table of all tuples appearing in the first and not in the second table
Join	Two tables and a condition	A table concatenating the attributes of the tuples that satisfy the given condition
Divide	A table, two attributes, and list of values	A table of values appearing in one specified attribute of the given table when the table has tuples that satisfies every value in the list in the other given attribute

question labels the diamond representing a predicate node. The answer to the question yields a binary value: 0 or 1, yes or no, ON or OFF. One of the output lines is selected accordingly. A **connector** serves to rejoin separated paths. Normally the circle representing a connector does not contain a label, but when the flowchart is used to document a computer program it may be convenient to label the connector.

Structured programming theory models all programs by their flowcharts by placing minor restrictions on their lines and nodes. Specifically, a flowchart is called a **proper program** if it has precisely one input and one output line, and for every node there exists a path from the input line through the node to the output line. The restriction prohibiting multiple input or output lines can easily be circumvented by funneling the lines through collector nodes. The other restriction simply discards unwanted program structures, since a program with a path that does not reach the output may not terminate.

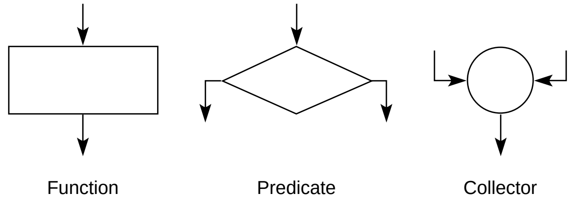


Fig. 2.2.6 Basic flowchart nodes.

Not all proper programs exhibit the desirable properties of meeting the objectives listed above. Figure 2.2.7 lists a group of proper programs whose graphs have been identified as being *well-structured* and useful as basic building blocks for creating other well-structured programs. The name assigned to each of these graph suggests the process each represents. CASE is just a convenient way of showing multiple IFTHENELSEs more compactly.

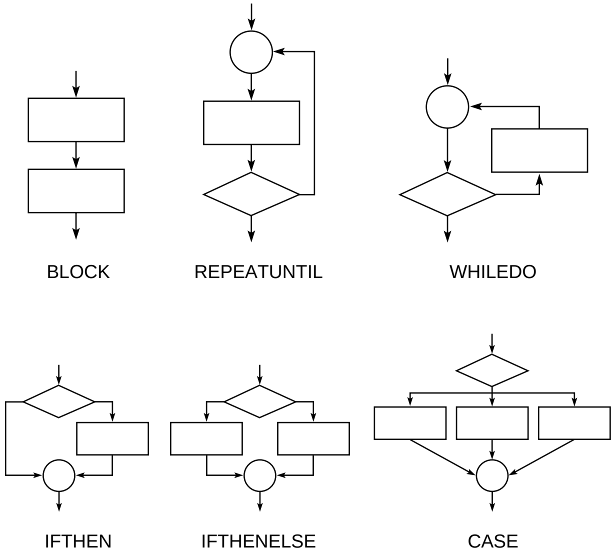


Fig. 2.2.7 Basic flowchart building blocks.

The **structured programming theorem** states: any proper program can be reconfigured to an equivalent program producing the same transformation of the data by a flowchart containing at most the graphs labeled BLOCK, IFTHENELSE, and REPEATUNTIL.

Every proper program has one input line and one output line like a function block. The synthesis of more complex well-structured programs is achieved by substituting any of the three building blocks mentioned in the theorem for a function node. In fact, any of the basic building blocks would do just as well. A program so structured will appear as a block of function nodes with a top-down control flow. Because of the top-down structure, the arrow points are not normally shown.

Figure 2.2.8 illustrates the expansion of a program to find the roots of $ax^2 + bx + c = 0$. The flowchart is shown in three levels of detail.

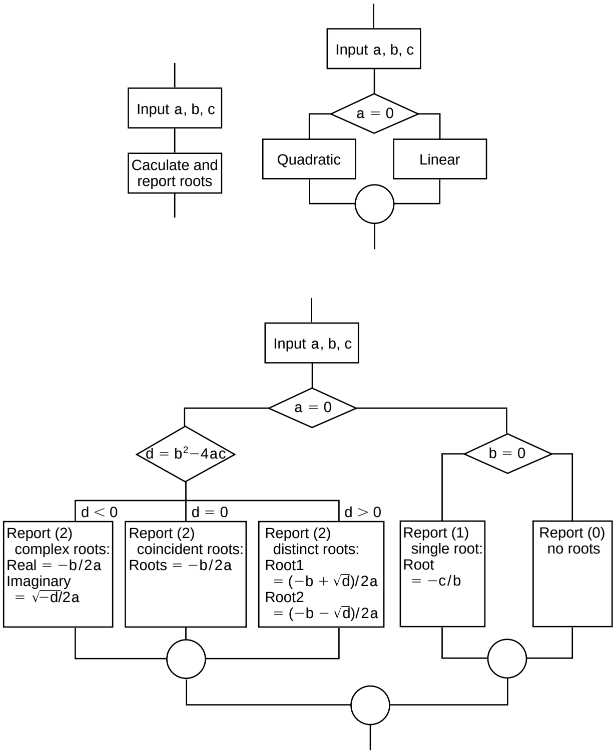

Input a, b, c
Caculate and report roots
Input a, b, c
a = 0
Quadratic
Linear
Input a, b, c
a = 0
d = b²−4ac
b = 0
d < 0
d = 0
d > 0
Report (2) complex roots: Real = −b/2a Imaginary = √−d/2a
Report (2) coincident roots: Roots = −b/2a
Report (2) distinct roots: Root1 = (−b + √d)/2a Root2 = (−b − √d)/2a
Report (1) single root: Root = −c/b
Report (0) no roots

Fig. 2.2.8 Illustration of a control-flow diagram.

Data-Flow Diagrams

Data-flow diagrams structure the actions of a program into a network by tracking the data as it passes through the program. They depict the interworkings of a system by the processes performing the work and the communication between the processes. Data-flow diagrams have proved valuable in analyzing existing or new systems to determine the system requirements and in designing systems to meet those requirements. Figure 2.2.9 shows the four basic elements used to construct a data-flow diagram. The roles each element plays in the system are:

Rectangular boxes lie outside the system and represent the input data sources or output data sinks that communicate with the system. The sources and sinks are also called *terminators*.

Circles (bubbles) represent processes or actions performed by the system in accomplishing its function.

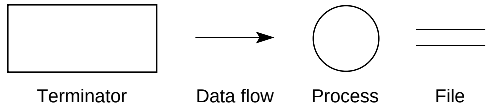


Fig. 2.2.9 Data-flow diagram elements.

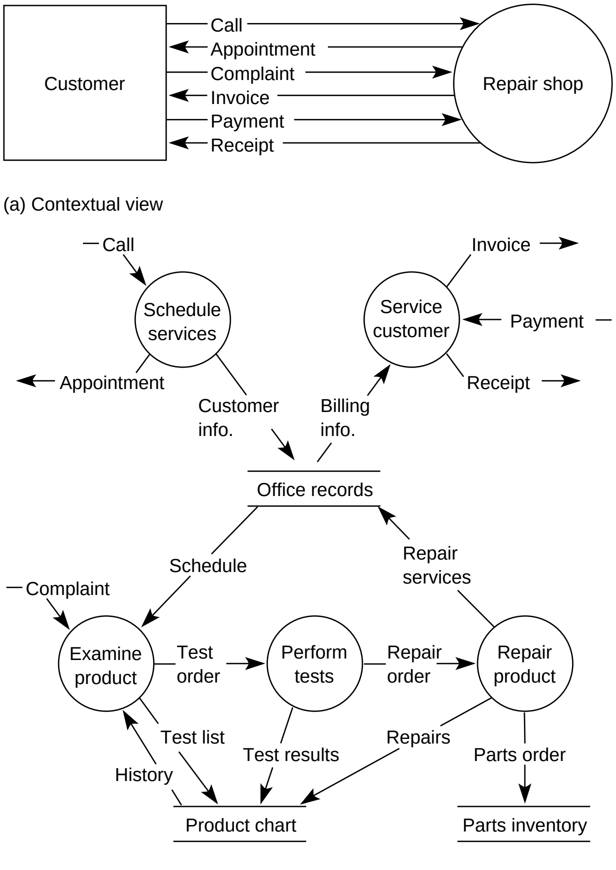
Twin parallel lines represent a data file used to collect and store data from among the processes or from a process over time which can be later recalled.

Arcs or vectors connect the other elements and represent data flows.

A label placed with each element makes clear its role in the system. The circles contain verbs and the other elements contain nouns. The arcs tie the system together. An arc between a terminator and a process represents input to or output from the system. An arc between two processes represents output from one process which is input to the other. An arc between a process and a file represents data gathered by the process and stored in the file, or retrieval of data from the file.

Analysis starts with a contextual view of the system studied in its environment. The contextual view gives the name of the system, the collection of terminators, and the data flows that provide the system inputs and outputs; all accompanied by a statement of the system objective. Details on the terminators and data they provide may also be described by text, but often the picture suffices. It is understood that the form of the input and output may not be dictated by the designer since they often involve organizations outside the system. Typical inputs in industrial systems include customer orders, payment checks, purchase orders, requests for quotations, etc. Figure 2.2.10*a* illustrates a context diagram for a repair shop.

Figure 2.2.10*b* gives many more operational details showing how the parts of the system interact to accomplish the system's objectives. The designer can restructure the internal processors and the formats of the data flows. The bubbles in a diagram can be broken down into further details to be shown in another data-flow diagram. This can be repeated level after level until the processes become manageable and understandable. To complete the system description, each bubble in the data-flow charts is accompanied by a control-flow diagram or its equivalent to describe the algorithm used to accomplish the actions and a data dictionary describing the items in the data flows and in the databases.



(a) Contextual view

(b) Behavorial view

Fig. 2.2.10 Illustration of a data-flow diagram.

The techniques of data-flow diagrams lend themselves beautifully to the analysis of existing systems. In a complex system it would be unusual for an individual to know all the details, but all system participants know their respective roles: what they receive, whence they receive it, what they do, what they send, and where they send it. By carefully structuring interviews, the complete system can be synthesized to any desired level of detail. Moreover, each system component can be verified because what is sent from one process must be received by another and what is received by a process must be used by the process. To automate the total system or parts of the system, control bubbles containing transition diagrams can be implemented to control the timing of the processes.

SOFTWARE SYSTEMS

Software Techniques

Two basic operations form the heart of nonnumerical techniques such as those found in handling large database tables. One basic operation, called **sorting,** collates the information in a table by reordering the items by their key into a specified order. The other basic operation, called **searching,** seeks to find items in a table whose keys have the same or related value as a given argument. The search operation may or may not be successful, but in either case further operations follow the search (e.g., retrieve, insert, replace).

One must recognize that computers cannot do mathematics. They can perform a few basic operations such as the four rules of arithmetic, but even in this case the operations are approximations. In fact, computers represent long integers, long rationals, and all the irrational numbers like π and e only as approximations. While computer arithmetic and the computer representation of numbers exceed the precision one commonly uses, the size of problems solved in a computer and the number of operations that are performed can produce misleading results with large computational errors.

Since the computer can handle only the four rules of arithmetic, complex **functions** must be approximated by polynomials or rational fractions. A rational fraction is a polynomial divided by another polynomial. From these curve-fitting techniques, a variety of weighted-average formulas can be developed to approximate the **definite integral** of a function. These formulas are used in the procedures for solving differential and integral equations. While differentiation can also be expressed by these techniques, it is seldom used, since the errors become unacceptable.

Taking advantage of the machine's speed and accuracy, one can solve **nonlinear equations** by trial and error. For example, one can use the Newton-Raphson method to find successive approximations to the roots of an equation. The computer is programmed to perform the calculations needed in each iteration and to terminate the procedure when it has converged on a root. More sophisticated routines can be found in the libraries for finding real, multiple, and complex roots of an equation.

Matrix techniques have been commercially programmed into libraries of prepared modules which can be integrated into programs written in all popular engineering programming languages. These libraries not only contain excellent routines for solving **simultaneous linear equations** and the **eigenvalues** of characteristic matrices, but also embody procedures guarding against ill-conditioned matrices which lead to large computational errors.

Special matrix techniques called **relaxation** are used to solve partial differential equations on the computer. A typical problem requires setting up a grid of hundreds or thousands of points to describe the region and expressing the equation at each point by finite-difference methods. The resulting matrix is very sparse with a regular pattern of nonzero elements. The form of the matrix circumvents the need for handling large arrays of numbers in the computer and avoids problems in computational accuracy normally found in dealing with extremely large matrices.

The computer is an excellent tool for handling **optimization** problems. Mathematically these problems are formulated as problems in finding the maximum or minimum of a nonlinear equation. The excellent techniques that have been developed can deal effectively with the unique complexities these problems have, such as saddle points which represent both a maximum and a minimum.

Another class of problems, called **linear programming** problems, is characterized by the linear constraint of many variables which plot into regions outlined by multidimensional planes (in the two-dimensional case, the region is a plane enclosed by straight lines). Techniques have been developed to find the optimal solution of the variables satisfying some given value or cost objective function. The solution to the problem proceeds by searching the corners of the region defined by the constraining equations to find points which represent minimum points of a cost function or maximum points of a value function.

The best known and most widely used techniques for solving statistical problems are those of linear **statistics.** These involve the techniques of **least squares** (otherwise known as **regression**). For some problems these techniques do not suffice, and more specialized techniques involving nonlinear statistics must be used, albeit a solution may not exist.

Artificial intelligence (AI) is the study and implementation of programs that model knowledge systems and exhibit aspects of intelligence in problem solving. Typical areas of application are in learning, linguistics, pattern recognition, decision making, and theorem proving. In AI, the computer serves to search a collection of heuristic rules to find a match with a current situation and to make inferences or otherwise reorganize knowledge into more useful forms. AI techniques have been utilized to build sophisticated systems, called **expert systems,** to aid in producing a timely response in problems involving a large number of complex conditions.

Operating Systems

The operating system provides the services that support the needs that computer programs have in common during execution. Any list of services would include those needed to configure the resources that will be made available to the users, to attach hardware units (e.g., memory modules, storage devices, coprocessors, and peripheral devices) to the existing configuration, to detach modules, to assign default parameters to the hardware and software units, to set up and schedule users' tasks so as to resolve conflicts and optimize throughput, to control system input and output devices, to protect the system and users' programs from themselves and from each other, to manage storage space in the file devices, to protect file devices from faults and illegal use, to account for the use of the system, and to handle in an orderly way any exception which might be encountered during program execution. A well-designed operating system provides these services in a user-friendly environment and yet makes itself and the computer operating staff transparent to the user.

The design of a computer operating system depends on the number of users which can be expected. The focus of single-user systems relies on the monitor to provide a user-friendly system through dialog menus with icons, mouse operations, and templets. Table 2.2.7 lists some popular operating systems for PCs by their trademark names. The design of a multiuser system attempts to give each user the impression that he/she is the lone user of the system. In addition to providing the accoutrements of a user-friendly system, the design focuses on the order of processing the jobs in an attempt to treat each user in a fair and equitable fashion. The basic issues for determining the order of processing center on the selection of job queues: the number of queues (a simple queue or a mix of queues), the method used in scheduling the jobs in the queue (first come–first served, shortest job next, or explicit priorities), and the internal handling of the jobs in the queue (batch, multiprogramming, or timesharing).

Table 2.2.7 Some Popular PC Operating Systems

Trademark	Supplier
DOS	Microsoft Corp.
Windows	Microsoft Corp.
OS/2	IBM Corp.
Unix	Unix Systems Laboratory Inc.
Sun/OS	Sun Microsystems Inc.
Macintosh	Apple Computer Inc.

Batch operating systems process jobs in a sequential order. Jobs are collected in batches and entered into the computer with individual job instructions which the operating system interprets to set up the job, to allocate resources needed, to process the job, and to provide the input/output. The operating system processes each job to completion in the order it appears in the batch. In the event a malfunction or fault occurs during execution, the operating system terminates the job currently being executed in an orderly fashion before initiating the next job in sequence.

Multiprogramming operating systems process several jobs concurrently. A job may be initiated any time memory and other resources which it needs become available. Many jobs may be simultaneously active in the system and maintained in a partial state of completion. The order of execution depends on the priority assignments. Jobs are executed to completion or put into a wait state until a pending request for service has been satisfied. It should be noted that, while the CPU can execute only a single program at any moment of time, operations with peripheral and storage devices can occur concurrently.

Timesharing operating systems process jobs in a way similar to multiprogramming except for the added feature that each job is given a short slice of the available time to complete its tasks. If the job has not been completed within its time slice or if it requests a service from an external device, it is put into a wait status and control passes to the next job. Effectively, the length of the time slice determines the priority of the job.

Program Preparation Facilities

For the user, the crucial part of a language system is the grammar which specifies the language syntax and semantics that give the symbols and rules used to compose acceptable statements and the meaning associated with the statements. Compared to natural languages, computer languages are more precise, have a simpler structure, and have a clearer syntax and semantics that allows no ambiguities in what one writes or what one means. For a program to be executed, it must eventually be translated into a sequence of basic machine instructions.

The statements written by a user must first be put on some machine-readable medium or typed on a keyboard for entry into the machine. The translator **(compiler)** program accepts these statements as input and translates (compiles) them into a sequence of basic machine instructions which form the executable version of the program. After that, the translated (compiled) program can be run.

During the execution of a program, a run-time program must also be present in the memory. The purpose of the run-time system is to perform services that the user's program may require. For example, in case of a program fault, the run-time system will identify the error and terminate the program in an orderly manner.

Some language systems do not have a separate compiler to produce machine-executable instructions. Instead the run-time system interprets the statements as written, converts them into a pseudo-code, and executes the coded version.

Commonly needed functions are made available as prepared modules, either as an integral part of the language or from stored libraries. The documentation of these functions must be studied carefully to assure correct selection and utilization.

Languages may be classified as procedure-oriented or problem-oriented. With **procedure-oriented** languages, all the detailed steps must be specified by the user. These languages are usually characterized as being more verbose than problem-oriented languages, but are more flexible and can deal with a wider range of problems. **Problem-oriented** languages deal with more specialized classes of problems. The elements of problem-oriented languages are usually familiar to a knowledgeable professional and so are easier to learn and use than procedure-oriented languages.

The most elementary form of a procedure-oriented language is called an **assembler.** This class of language permits a computer program to be written directly in basic computer instructions using mnemonic operators and symbolic operands. The assembler's translator converts these instructions into machine-usable form.

A further refinement of an assembler permits the use of macros. A **macro** identifies, by an assigned name and a list of formal parameters, a sequence of computer instructions written in the assembler's format and stored in its subroutine library. The macroassembler includes these macro instructions in the translated program along with the instructions written by the programmer.

Besides these basic language systems there exists a large variety of other language systems. These are called higher-level language systems since they permit more complex statements than are permitted by a macroassembler. They can also be used on machines produced by different manufacturers or on machines with different instruction repertoires.

In the field of business programming, **COBOL** (COmmon Business-Oriented Language) is the most popular. This language facilitates the handling of the complex information files found in business and data-processing problems.

Another example of an application area supported by special languages is in the field of problems involving strings of text. SNOBOL and LISP exemplify these string-manipulation or list-processing languages. Applications vary from generating concordances to sophisticated symbolic formula manipulation.

One language of historical value is **ALGOL 60.** It is a landmark in the theoretical development of computer languages. It was designed and standardized by an international committee whose goal was to formulate a language suitable for publishing computer algorithms. Its importance lies in the many language features it introduced which are now common in the more recent languages which succeeded it and in the scientific notation which was used to define it.

FORTRAN (FORmula TRANslator) was one of the first languages catering to the engineering and scientific community where algebraic formulas specify the computations used within the program. It has been standardized several times. The current version is FORTRAN 90 (ANSI X3.198-1992). Each version has expanded the language features and has removed undesirable features which lead to unstructured programs. The new features include new data types like Boolean and character strings, additional operators and functions, and new statements that support programs conforming to the requirements for structured programming.

The **PASCAL** language couples the ideas of ALGOL 60 to those of structured programming. By allowing only appropriate statement types, it guarantees that any program written in the language will be well-structured. In addition, the language introduced new data types and allows programmers to define new complex data structures based on the primitive data types.

The definition of the **Ada** language was sponsored by the Department of Defense as an all-encompassing language for the development and maintenance of very large, software-intensive projects over their life cycle. While it meets software engineering objectives in a manner similar to Pascal, it has many other features not normally found in programming languages. Like other attempts to formulate very large all-inclusive languages, it is difficult to learn and has not found popular favor. Nevertheless, its many unique features make it especially valuable in implementing programs which cannot be easily implemented in other languages (e.g., programs for parallel computations in embedded computers).

By edict, subsets of Ada were forbidden. **Modula-2** was designed to retain the inherent simplicity of PASCAL but include many of the advanced features of Ada. Its advantage lies in implementing large projects involving many programmers. The compilers for this language have rigorous interface cross-checking mechanisms to avoid poor interfaces between components. Another troublesome area is in the implicit use of global data. Modula-2 retains the Ada facilities that allow programmers to share data and avoids incorrectly modifying the data in different program units.

The **C** language was developed by AT&T's Bell Laboratories and subsequently standardized by ANSI. It has a reputation for translating programs into compact and fast code, and for allowing program segments to be precompiled. Its strength rests in the flexibility of the language; for example, it permits statements from other languages to be included in-line in a C program and it offers the largest selection of operators that mirror those available in an assembly language. Because of its flexibility, programs written in C can become unreadable.

Problem-oriented languages have been developed for every discipline. A language might deal with a specialized application within an engineering field, or it might deal with a whole gamut of applications covering one or more fields.

A class of problem-oriented languages that deserves special mention are those for solving problems in **discrete simulation.** GPSS, Simscript, and SIMULA are among the most popular. A simulation (another word for *model*) of a system is used whenever it is desirable to watch a succession of many interrelated events or when there is interplay between the system under study and outside forces. Examples are problems in human-machine interaction and in the modeling of business systems. Typical human-machine problems are the servicing of automatic equipment by a crew of operators (to study crew size and assignments, typically), or responses by shared maintenance crews to equipment subject to unpredictable (random) breakdown. Business models often involve transportation and warehousing studies. A business model could also study the interactions between a business and the rest of the economy such as competitive buying in a raw materials market or competitive marketing of products by manufacturers.

Physical or chemical systems may also be modeled. For example, to study the application of automatic control values in pipelines, the computer model consists of the control system, the valves, the piping system, and the fluid properties. Such a model, when tested, can indicate whether fluid hammer will occur or whether valve action is fast enough. It can also be used to predict pressure and temperature conditions in the fluid when subject to the valve actions.

Another class of problem-oriented languages makes the computer directly accessible to the specialist with little additional training. This is achieved by permitting the user to describe problems to the computer *in terms that are familiar in the discipline of the problem* and for which the language is designed. Two approaches are used. Figures 2.2.11 and 2.2.12 illustrate these.

One approach sets up the computer program directly from the mathematical equations. In fact, problems were formulated in this manner in the past, where analog computers were especially well-suited. Anyone familiar with analog computers finds the transitions to these languages easy. Figure 2.2.11 illustrates this approach using the MIMIC language to write the program for the solution of the initial-value problem:

$$M\ddot{y} + Z\dot{y} + Ky = 1 \qquad \text{and} \qquad \dot{y}(0) = y(0) = 0$$

MIMIC is a digital simulation language used to solve systems of ordinary differential equations. The key step in setting up the solution is to isolate the highest-order derivative on the left-hand side of the equation and equate it to an expression composed of the remaining terms. For the

MIMIC statements	Explanation
$DY2 = (1 - Z * DY1 - K * Y)/M$	Differential equation to be solved. ''*'' is used for multiplication and $DY2$, $DY1$, and Y are defined mnemonics for $\ddot{y}$, $\dot{y}$, and y.
$DY1 = \text{INT}(DY2,0.)$	INT(A,B) is used to perform integration. It forms successive values of $B + \int Adt$.
$Y = \text{INT}(DY1,0.)$	
FIN(T,10.)	T is a reserved name representing the independent variable. This statement will terminate execution when $T \geq 10$.
CON(M,K,Z)	Values must be furnished for M, K, and Z. An input with these values must appear after the END card.
PLO(T,$DY2$)	Three point plots are produced on the line printer; $\ddot{y}$, $\dot{y}$, and y vs. t.
PLO(T,$DY1$)	
PLO(T,Y)	
END	Necessary last statement.

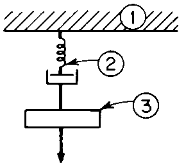
Fig. 2.2.11 Illustration of a MIMIC program.

equation above, this results in:

$$\ddot{y} = (1 - Z\dot{y} - Ky)/M$$

The highest-order derivative is derived by equating it to the expression on the right-hand side of the equation. The lower-order derivatives in the expression are generated successively by integrating the highest-order derivative. The MIMIC language permits the user to write these statements in a format closely resembling mathematical notation.

The alternate approach used in problem-oriented languages permits the setup to be described to the computer directly from the block diagram of the physical system. Figure 2.2.12 illustrates this approach using the SCEPTRE language. SCEPTRE statements are written under headings and subheadings which identify the type of component being described. This language may be applied to network problems of electrical digital-logic elements, mechanical-translation or rotational elements, or transfer-function blocks. The translator for this language develops and sets up the equations directly from this description of the network diagram, and so relieves the user from the mathematical aspects of the problem.



$K1 = 40.$
$D1 = .5$
$M1 = 10.$
$R1 = 7.32$

A node is assigned to:
- ground
- any mass
- point between two elements

The prefix of element name specifies its type; i.e., M for mass, K for spring, D for damper, and R for force.

(a)

SCEPTRE statements	Explanation
MECHANICAL DESCRIPTION	
ELEMENTS	Specifies the elements and their position in the diagram using the node numbers.
$M1$, $1 - 3 = 10.$	
$K1$, $1 - 2 = 40.$	
$D1$, $2 - 3 = .5$	
$R1$, $1 - 3 = 7.32$	
OUTPUT	Results are listed on the line printer. Prefix on the element specifies the quantity to be listed; S for displacement, V for velocity.
$SM1$,$VM1$	
RUN CONTROL	TIME is reserved name for independent variable. Statement will terminate execution of program when TIME is equal to or greater than 10.
STOPTIME = 10.	
END	Necessary statement.

(b)

Fig. 2.2.12 Illustration of SCEPTRE program. (a) Problem to be solved; (b) SCEPTRE program.

Application Packages

An application package differs from a language in that its components have been organized to solve problems in a particular application rather than to create the components themselves. The user interacts with the package by initiating the operations and providing the data. From an operational view, packages are built to minimize or simplify interactions with the users by using a menu to initiate operations and entering the data through templets.

Perhaps the most widely used application package is the **word processor.** The objective of a word processor is to allow users to compose text in an electronically stored format which can be corrected or modified, and from which a hard copy can be produced on demand. Besides the basic typewriter operations, it contains functions to manipulate text in blocks or columns, to create headers and footers, to number pages, to find and correct words, to format the data in a variety of ways, to create labels, and to merge blocks of text together. The better word processors have an integrated dictionary, a spelling checker to find and correct misspelled words, a grammar checker to find grammatical errors, and a thesaurus. They often have facilities to prepare complex mathematical equations and to include and manipulate graphical artwork, including editing color pictures. When enough page- and document formatting capability has been added, the programs are known as **desktop publishing** programs.

One of the programs that contributed to the early acceptance of personal computers was the **spread sheet** program. These programs simulate the common spread sheet with its columns and rows of interrelated data. The computerized approach has the advantage that the equations are stored so that the results of a change in data can be shown quickly after any change is made in the data. Modern spread sheet programs have many capabilities, including the ability to obtain information from other spread sheets, to produce a variety of reports, and to prepare equations which have complicated logical aspects.

Tools for **project management** have been organized into commercially available application packages. The objectives of these programs are in the planning, scheduling, and controlling the time-oriented activities describing the projects. There are two basically similar techniques used in these packages. One, called **CPM** (critical path method), assumes that the project activities can be estimated deterministically. The other, called **PERT** (project evaluation and review technology), assumes that the activities can be estimated probabilistically. Both take into account such items as the requirement that certain tasks cannot start before the completion of other tasks. The concepts of **critical path** and **float** are crucial, especially in scheduling the large projects that these programs are used for. In both cases tools are included for estimating project schedules, estimating resources needed and their schedules, and representing the project activities in graphical as well as tabular form.

A major use of the digital computer is in **data reduction,** data analysis, and visualization of data. In installations where large amounts of data are recorded and kept, it is often advisable to reduce the amount of data by ganging the data together, by averaging the data with numerical filters to reduce the amount of noise, or by converting the data to a more appropriate form for storage, analysis, visualization, or future processing. This application has been expanded to produce systems for evaluation, automatic testing, and fault diagnosis by coupling the data acquisition equipment to special peripherals that automatically measure and record the data in a digital format and report the data as meaningful, nonphysically measurable parameters associated with a mathematical model.

Computer-aided design/computer-aided manufacturing **(CAD/CAM)** is an integrated collection of software tools which have been designed to make way for innovative methods of fabricating customized products to meet customer demands. The goal of modern manufacturing is to process orders placed for different products sooner and faster, and to fabricate them without retooling. CAD has the tools for prototyping a design and setting up the factory for production. Working within a framework of agile manufacturing facilities that features automated vehicles, handling robots, assembly robots, and welding and painting robots, the factory sets itself up for production under computer control. Production starts with the receipt of an order on which customers may pick options such as color, size, shapes, and features. Manufacturing proceeds with greater flexibility, quality, and efficiency in producing an increased number of products with a reduced workforce. Effectively, CAD/CAM provides for the ultimate just-in-time (JIT) manufacturing.

Two other types of application package illustrate the versatility of data management techniques. One type ties on-line equipment to a computer for collecting real-time data from the production lines. An animated, pictorial display of the production lines forms the heart of the system, allowing supervision in a central control station to continuously track operations. The other type collects time-series data from the various activities in an enterprise. It assists in what is known as **management by exception.** It is especially useful where the detailed data is so voluminous that it is feasible to examine it only in summaries. The data elements are processed and stored in various levels of detail in a seamless fashion. The system stores the reduced data and connects it to the detailed data from which it was derived. The application package allows management, through simple computer operations, to detect a problem at a higher level and to locate and pinpoint its cause through examination of successively lower levels.