Ames, W.F.; et. al. “Mathematics”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Mathematics

William F. Ames
Georgia Institute of Technology

George Cain
Georgia Institute of Technology

Y. L. Tong
Georgia Institute of Technology

W. G. Steele
Mississippi State University

H. W. Coleman
University of Alabama

R. L. Kautz
National Institute of Standards and Technology

Dan M. Frangopol
University of Colorado

19.1 Tables

Greek Alphabet

Greek Letter	Greek Name	English Equivalent	Greek Letter	Greek Name	English Equivalent
A α	Alpha	a	N ν	Nu	n
B β	Beta	b	Ξ ξ	Xi	x
Γ γ	Gamma	g	O ο	Omicron	o
Δ δ	Delta	d	Π π	Pi	p
E ε	Epsilon	e	P ρ	Rho	r
Z ζ	Zeta	z	Σ σ ς	Sigma	s
H η	Eta	e	T τ	Tau	t
Θ θ ϑ	Theta	th	Y υ	Upsilon	u
I ι	Iota	i	Φ ϕ φ	Phi	ph
K κ	Kappa	k	X χ	Chi	ch
Λ λ	Lambda	l	Ψ ψ	Psi	ps
M μ	Mu	m	Ω ω	Omega	o

International System of Units (SI)

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built from seven *SI base units*, one for each of the seven dimensionally independent base quantities: the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes*, may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

Definitions of SI Base Units

Meter: The meter is the length of path traveled by light in vacuum during a time interval of 1/299 792 458 of a second (17th CGPM, 1983).
Kilogram: The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).
Second: The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).
Ampere: The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length (9th CGPM, 1958).
Kelvin: The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).
Mole: The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles, or specified groups of such particles (14th CGPM, 1971). Examples of the use of the mole:

- 1 mol of H_2 contains about 6.022×10^{23} H_2 molecules, or 12.044×10^{23} H atoms.
- 1 mol of HgCl has a mass of 236.04 g.
- 1 mol of Hg_2Cl_2 has a mass of 472.08 g.
- 1 mol of Hg_2^{2+} has a mass of 401.18 g and a charge of 192.97 kC.
- 1 mol of $Fe_{0.91}$ S has a mass of 82.88 g.
- 1 mol of e^- has a mass of 548.60 μg and a charge of –96.49 kC.
- 1 mol of photons whose frequency is 10^{14} Hz has energy of about 39.90 kJ.

Candela: The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979).

Names and Symbols for the SI Base Units

Physical Quantity	Name of SI Unit	Symbol for SI Unit
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

SI Derived Units with Special Names and Symbols

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Frequency[a]	hertz	Hz	s^{-1}
Force	newton	N	$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal	Pa	$N \cdot m^{-2} = m^{-1} \cdot kg \cdot s^{-2}$
Energy, work, heat	joule	J	$N \cdot m = m^2 \cdot kg \cdot s^{-2}$
Power, radiant flux	watt	W	$J \cdot s^{-1} = m^2 \cdot kg \cdot s^{-3}$
Electric charge	coulomb	C	$A \cdot s$

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Electric potential, electromotive force	volt	V	$J \cdot C^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric resistance	ohm	Ω	$V \cdot A^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Electric conductance	siemens	S	$\Omega^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Electric capacitance	farad	F	$C \cdot V^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Magnetic flux density	tesla	T	$V \cdot s \cdot m^{-2} = kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux	weber	Wb	$V \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry	H	$V \cdot A^{-1} \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Celsius temperature[b]	degree Celsius	°C	K
Luminous flux	lumen	lm	$cd \cdot sr$
Illuminance	lux	lx	$cd \cdot sr \cdot m^{-2}$
Activity (radioactive)	becquerel	Bq	s^{-1}
Absorbed dose (or radiation)	gray	Gy	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$
Dose equivalent (dose equivalent index)	sievert	Sv	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$
Plane angle	radian	rad	$1 = m \cdot m^{-1}$
Solid angle	steradian	sr	$1 = m^2 \cdot m^{-2}$

[a] For radial (circular) frequency and for angular velocity the unit rad s^{-1}, or simply s^{-1}, should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

[b] The Celsius temperature θ is defined by the equation

$$q/°C = T/K = 237.15$$

The SI unit of Celsius temperature interval is the degree Celsius, °C, which is equal to the kelvin, K. °C should be treated as a single symbol, with no space between the ° sign and the letter C. (The symbol °K, and the symbol °, should no longer be used.)

Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; megaelectronvolt, MeV; and kilotonne, kt.

Physical Quantity	Name of Unit	Symbol for Unit	Value in SI Units
Time	minute	min	60 s
Time	hour	h	3600 s
Time	day	d	86 400 s
Plane angle	degree	°	$(\pi/180)$ rad
Plane angle	minute	′	$(\pi/10\,800)$ rad
Plane angle	second	″	$(\pi/648\,000)$ rad
Length	angstrom[a]	Å	10^{-10} m
Area	barn	b	10^{-28} m^2
Volume	liter	l, L	$dm^3 = 10^{-3}$ m^3
Mass	tonne	t	$Mg = 10^3$ kg
Pressure	bar[a]	bar	10^5 Pa $= 10^5$ N $\cdot$ m^{-2}
Energy	electronvolt[b]	eV $(= e \times V)$	$\approx 1.60218 \times 10^{-19}$ J
Mass	unified atomic mass unit[b,c]	u $(= m_a(12C)/12)$	$\approx 1.66054 \times 10^{-27}$ kg

[a] The angstrom and the bar are approved by CIPM for "temporary use with SI units," until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

[b] The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants e (for the electronvolt) and N_A (for the unified atomic mass unit), which are determined by experiment.

[c] The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

Conversion Constants and Multipliers

Recommended Decimal Multiples and Submultiples

Multiple or Submultiple	Prefix	Symbol	Multiple or Submultiple	Prefix	Symbol
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ (Greek mu)
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	hecto	h	10^{-15}	femto	f
10	deca	da	10^{-18}	atto	a

Conversion Factors — Metric to English

To Obtain	Multiply	By
Inches	Centimeters	0.393 700 787 4
Feet	Meters	3.280 839 895
Yards	Meters	1.093 613 298
Miles	Kilometers	0.621 371 192 2
Ounces	Grams	$3.527\,396\,195 \times 10^{-2}$
Pounds	Kilograms	2.204 622 622
Gallons (U.S. liquid)	Liters	0.264 172 052 4
Fluid ounces	Milliliters (cc)	$3.381\,402\,270 \times 10^{-2}$
Square inches	Square centimeters	0.155 000 310 0
Square feet	Square meters	10.763 910 42
Square yards	Square meters	1.195 990 046
Cubic inches	Milliliters (cc)	$6.102\,374\,409 \times 10^{-2}$
Cubic feet	Cubic meters	35.314 666 72
Cubic yards	Cubic meters	1.307 950 619

Conversion Factors — English to Metric

To Obtain	Multiply	By[a]
Microns	Mils	**25.4**
Centimeters	Inches	**2.54**
Meters	Feet	**0.3048**
Meters	Yards	**0.9144**
Kilometers	Miles	1.609 344
Grams	Ounces	28.349 523 13
Kilograms	Pounds	0.453 592 37
Liters	Gallons (U.S. liquid)	3.785 411 784
Millimeters (cc)	Fluid ounces	29.573 529 56
Square centimeters	Square inches	6.451 6
Square meters	Square feet	0.092 903 04
Square meters	Square yards	0.836 127 36
Milliliters (cc)	Cubic inches	16.387 064
Cubic meters	Cubic feet	$2.831\,684\,659 \times 10^{-2}$
Cubic meters	Cubic yards	0.764 554 858

[a] Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Conversion Factors — General

To Obtain	Multiply	By[a]
Atmospheres	Feet of water @ 4°C	2.950×10^{-2}
Atmospheres	Inches of mercury @ 0°C	3.342×10^{-2}
Atmospheres	Pounds per square inch	6.804×10^{-2}
Btu	Foot-pounds	1.285×10^{-3}
Btu	Joules	9.480×10^{-4}
Cubic feet	Cords	**128**
Degree (angle)	Radians	57.2958
Ergs	Foot-pounds	1.356×10^{-7}
Feet	Miles	**5280**
Feet of water @ 4°C	Atmospheres	33.90
Foot-pounds	Horsepower-hours	1.98×10^{6}
Foot-pounds	Kilowatt-hours	2.655×10^{6}
Foot-pounds per minute	Horsepower	3.3×10^{4}
Horsepower	Foot-pounds per second	1.818×10^{-3}
Inches of mercury @ 0°C	Pounds per square inch	2.036
Joules	Btu	1054.8
Joules	Foot-pounds	1.355 82
Kilowatts	Btu per minute	1.758×10^{-2}
Kilowatts	Foot-pounds per minute	2.26×10^{-5}
Kilowatts	Horsepower	0.745712
Knots	Miles per hour	0.868 976 24
Miles	Feet	1.894×10^{-4}
Nautical miles	Miles	0.868 976 24
Radians	Degrees	1.745×10^{-2}
Square feet	Acres	**43 560**
Watts	Btu per minute	17.5796

[a] Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Temperature Factors

$$°\mathrm{F} = 9/5(°\mathrm{C}) + 32$$

$$\text{Fahrenheit temperature} = 1.8(\text{temperature in kelvins}) - 459.67$$

$$°\mathrm{C} = 5/9[(°\mathrm{F}) - 32]$$

$$\text{Celsius temperature} = \text{temperature in kelvins} - 273.15$$

$$\text{Fahrenheit temperature} = 1.8(\text{Celsius temperature}) + 32$$

Conversion of Temperatures

From	To		From	To	
Fahrenheit	Celcius	$t_C = \frac{t_F - 32}{1.8}$	Celsius	Fahrenheit	$t_F = (t_c \times 1.8) + 32$
				Kelvin	$T_K = t_c + 273.15$
				Rankine	$T_R = (t_c + 273.15) \times 18$
	Kelvin	$T_k = \frac{t_F - 32}{1.8} + 273.15$	Kelvin	Celsius	$t_c = T_K - 273.15$
				Rankine	$T_R = T_k \times 1.8$
	Rankine	$T_R = t_F + 459.67$	Rankine	Fahrenheit	$t_F = T_R - 459.67$
				Kelvin	$T_K = \frac{T_R}{1.8}$

Physical Constants

General

Equatorial radius of the earth = 6378.388 km = 3963.34 miles (statute)
Polar radius of the earth = 6356.912 km = 3949.99 miles (statute)
1 degree of latitude at 40° = 69 miles
1 international nautical mile = 1.150 78 miles (statute) = 1852 m = 6076.115 ft
Mean density of the earth = 5.522 g/cm^3 = 344.7 lb/ft^3
Constant of gravitation $(6.673 \pm 0.003) \times 10^{-8} \cdot \text{cm}^3 \cdot \text{g}^{-1} \cdot \text{s}^{-2}$
Acceleration due to gravity at sea level, latitude 45° = 980.6194 cm/s^2 = 32.1726 ft/s^2
Length of seconds pendulum at sea level, latitude 45° = 99.3575 cm = 39.1171 in.
1 knot (international) = 101.269 ft/min = 1.6878 ft/s = 1.1508 miles (statute)/h
1 micron = 10^{-4} cm
1 angstrom = 10^{-8} cm
Mass of hydrogen atom = $(1.673\ 39 \pm 0.0031) \times 10^{-24}$ g
Density of mercury at 0°C = 13.5955 g/mL
Density of water at 3.98°C = 1.000 000 g/mL
Density, maximum, of water, at 3.98°C = 0.999 973 g/cm^3
Density of dry air at 0°C, 760 mm = 1.2929 g/L
Velocity of sound in dry air at 0°C = 331.36 m/s – 1087.1 ft/s
Velocity of light in vacuum = $(2.997\ 925 \pm 0.000\ 002) \times 10^{10}$ cm/s
Heat of fusion of water, 0°C = 79.71 cal/g
Heat of vaporization of water, 100°C = 539.55 cal/g
Electrochemical equivalent of silver 0.001 118 g/s international amp
Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure = 6438.4696 Å
Wavelength of orange-red line of krypton 86 = 6057.802 Å

π Constants

π = 3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37511
$1/\pi$ = 0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091
π^2 = 9.8690 44010 89358 61883 44909 99876 15113 53136 99407 24079
$\log_e \pi$ = 1.14472 98858 49400 17414 34273 51353 05871 16472 94812 91531
$\log_{10} \pi$ = 0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044
$\log_{10} \sqrt{2\pi}$ = 0.39908 99341 79057 52478 25035 91507 69595 02099 34102 92128

Constants Involving *e*

e = 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69996
$1/e$ = 0.36787 94411 71442 32159 55237 70161 46086 74458 11131 03177
e^2 = 7.38905 60989 30650 22723 04274 60575 00781 31803 15570 55185
$M = \log_{10} e$ = 0.43429 44819 03251 82765 11289 18916 60508 22943 97005 80367
$1/M = \log_e 10$ = 2.30258 50929 94045 68401 79914 54684 36420 76011 01488 62877
$\log_{10} M$ = 9.63778 43113 00536 78912 29674 98645 – 10

Numerical Constants

$\sqrt{2}$ = 1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37695
$\sqrt[3]{2}$ = 1.25992 10498 94873 16476 72106 07278 22835 05702 51464 70151
$\log_e 2$ = 0.69314 71805 59945 30941 72321 21458 17656 80755 00134 36026
$\log_{10} 2$ = 0.30102 99956 63981 19521 37388 94724 49302 67881 89881 46211
$\sqrt{3}$ = 1.73205 08075 68877 29352 74463 41505 87236 69428 05253 81039
$\sqrt[3]{3}$ = 1.44224 95703 07408 38232 16383 10780 10958 83918 69253 49935
$\log_e 3$ = 1.09861 22886 68109 69139 52452 36922 52570 46474 90557 82275
$\log_{10} 3$ = 0.47712 12547 19662 43729 50279 03255 11530 92001 28864 19070

Symbols and Terminology for Physical and Chemical Quantities

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Mass	m		kg
Reduced mass	μ	$\mu = m_1 m_2/(m_1 + m_2)$	kg
Density, mass density	ρ	$\rho = m/V$	$kg \cdot m^{-3}$
Relative density	d	$d = \rho/\rho^{\theta}$	1
Surface density	ρ_A,ρ_S	$\rho_a = m/A$	$kg \cdot m^{-2}$
Specific volume	ν	$\nu = V/m = 1/\rho$	$m^3 \cdot kg^{-1}$
Momentum	p	$p = mv$	$kg \cdot m \cdot s^{-1}$
Angular momentum, action	L	$L = r \times p$	$J \cdot s$
Moment of inertia	I, J	$I = \Sigma\, m_i\, r_i^2$	$kg \cdot m^2$
Force	F	$F = d\,p/d\,t = m\,a$	N
Torque, moment of a force	T, (M)	$T = r \times F$	$N \cdot m$
Energy	E		J
Potential energy	E_p, V, Φ	$E_p = \int F \cdot ds$	J
Kinetic energy	E_k, T, K	$E_k = (1/2)\, mv^2$	J
Work	W, w	$W = \int F \cdot ds$	J
Hamilton function	H	$H(q,p) = T(q,p) + V(q)$	J
Lagrange function	L	$L(q,\dot{q}) = T(q,\dot{q}) - V(q)$	J
Pressure	p, P	$p = F/A$	Pa, $N \cdot m^{-2}$
Surface tension	γ, σ	$\gamma = dW/dA$	$N \cdot m^{-1}$, $J \cdot m^{-1}$
Weight	G (W, P)	$G = mg$	N
Gravitational constant	G	$F = Gm_1m_2/r^2$	$N \cdot m^2 \cdot kg^{-2}$
Normal stress	σ	$\sigma = F/A$	Pa
Shear stress	τ	$\tau = F/A$	Pa
Linear strain, relative elongation	ε, e	$\varepsilon = \Delta l/l$	1
Modulus of elasticity, Young's modulus	E	$E = \sigma/\varepsilon$	Pa
Shear strain	γ	$\gamma = \Delta x/d$	1
Shear modulus	G	$G = \tau/\gamma$	Pa
Volume strain, bulk strain	θ	$\theta = \Delta V/V_0$	1
Bulk modulus, compression modulus	K	$K = V_0(dp/dV)$	Pa
Viscosity, dynamic viscosity	η, μ	$\tau_{x,z} = \eta(dv_x/dz)$	$Pa \cdot s$
Fluidity	ϕ	$\phi = l/\eta$	$m \cdot kg^{-1} \cdot s$
Kinematic viscosity	ν	$\nu = \eta/\rho$	$m^2 \cdot s^{-1}$
Friction coefficient	μ, (f)	$F_{frict} = \mu F_{norm}$	1
Power	P	$P = dW/dt$	W
Sound energy flux	P, P_a	$P = dE/dt$	W
Acoustic factors			
Reflection factor	ρ	$\rho = P_r/P_0$	1
Acoustic absorption factor	α_a, (α)	$\alpha_a = 1 - \rho$	1
Transmission factor	τ	$\tau = P_t/P_0$	1
Dissipation factor	δ	$\delta = \alpha_a - \tau$	1

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Electricity and Magnetism			
Quantity of electricity, electric range	Q		C
Charge density	ρ	$\rho = Q/V$	$C \cdot m^{-3}$
Surface charge density	σ	$\sigma = Q/A$	$C \cdot m^{-2}$
Electric potential	V, ϕ	$V = dW/dQ$	V, $J \cdot C^{-1}$
Electric potential difference	U, ΔV, $\Delta\phi$	$U = V_2 - V_1$	V
Electromotive force	E	$E = \int(F/Q) \cdot ds$	V
Electric field strength	E	$E = F/Q = -\text{grad } V$	$V \cdot m^{-1}$
Electric flux	ψ	$\psi = \int D \cdot dA$	C
Electric displacement	D	$D = \varepsilon E$	$C \cdot m^{-2}$
Capacitance	C	$C = Q/U$	F, $C \cdot V^{-1}$
Permittivity	ε	$D = \varepsilon E$	$F \cdot m^{-1}$
Permittivity of vacuum	ε_0	$\varepsilon_0 = \mu_0^{-1} c_0^{-2}$	$F \cdot m^{-1}$
Relative permittivity	ε_r	$\varepsilon_r = \varepsilon/\varepsilon_0$	1
Dielectric polarization (dipole moment per volume)	P	$P = D - \varepsilon_0 E$	$C \cdot m^{-2}$
Electric susceptibility	χ_e	$\chi_e = \varepsilon_\tau - 1$	1
Electric dipole moment	p, μ	$P = Qr$	$C \cdot m$
Electric current	I	$I = dQ/dt$	A
Electric current density	j, J	$I = \int j \cdot dA$	$A \cdot m^{-2}$
Magnetic flux density, magnetic induction	B	$F = Qv \times B$	T
Magnetic flux	Φ	$\Phi = \int B \cdot dA$	Wb
Magnetic field strength	H	$B = \mu H$	$A \cdot M^{-1}$
Permeability	μ	$B = \mu H$	$N \cdot A^{-2}$, $H \cdot m^{-1}$
Permeability of vacuum	μ_0		$H \cdot m^{-1}$
Relative permeability	μ_t	$\mu_\tau = \mu/\mu_0$	1
Magnetization (magnetic dipole moment per volume)	M	$M = B/\mu_0 - H$	$A \cdot m^{-1}$
Magnetic susceptibility	χ, κ, (χ_m)	$\chi = \mu_\tau - 1$	1
Molar magnetic susceptibility	χ_m	$\chi_m = V_m \chi$	$m^3 \cdot mol^{-1}$
Magnetic dipole moment	m, μ	$E_p = -m \cdot B$	$A \cdot m^2$, $J \cdot T^{-1}$
Electrical resistance	R	$R = U/I$	Ω
Conductance	G	$G = 1/R$	S
Loss angle	δ	$\delta = (\pi/2) + \phi_I - \phi_U$	l, rad
Reactance	X	$X = (U/I) \sin \delta$	Ω
Impedance (complex impedance)	Z	$Z = R + iX$	Ω
Admittance (complex admittance)	Y	$Y = 1/Z$	S
Susceptance	B	$Y = G + iB$	S
Resistivity	ρ	$\rho = E/j$	$\Omega \cdot m$
Conductivity	κ, γ, σ	$\kappa = 1/\rho$	$S \cdot m^{-1}$
Self-inductance	L	$E = -L(dI/dt)$	H
Mutual inductance	M, L_{12}	$E_1 = L_{12}(dI_2/dt)$	H
Magnetic vector potential	A	$B = \nabla \times A$	$Wb \cdot m^{-1}$
Poynting vector	S	$S = E \times H$	$W \cdot m^{-2}$
Electromagnetic Radiation			
Wavelength	λ		m
Speed of light			
In vacuum	c_0		$m \cdot s^{-1}$
In a medium	c	$c = c_0/n$	$m \cdot s^{-1}$
Wavenumber in vacuum	$\tilde{\nu}$	$\tilde{\nu} = \nu/c_0 = 1/n\lambda$	m^{-1}
Wavenumber (in a medium)	σ	$\sigma = 1/\lambda$	m^{-1}
Frequency	ν	$\nu = c/\lambda$	Hz
Circular frequency, pulsatance	ω	$\omega = 2\pi\nu$	s^{-1}, $rad \cdot s^{-1}$
Refractive index	n	$n = c_0/c$	1
Planck constant	h		$J \cdot s$

Name	Symbol	Definition	SI Unit
		Classical Mechanics	
Planck constant/2π	$\bar{h}$	$\bar{h} = h/2\pi$	J · s
Radiant energy	Q, W		J
Radiant energy density	ρ, w	$\rho = Q/V$	J · m^{-3}
Spectral radiant energy density			
In terms of frequency	ρ_ν, w_ν	$\rho_\nu = d\rho/d\nu$	J · m^{-3} · Hz^{-1}
In terms of wavenumber	$\rho_{\tilde{\nu}}$ $w_{\tilde{\nu}}$	$\rho_\nu = d\rho / d\tilde{\nu}$	J · m^{-2}
In terms of wavelength	ρ_λ, w_λ	$\rho_\lambda = d\rho/d\lambda$	J · m^{-4}
Einstein transition probabilities			
Spontaneous emission	A_{nm}	$dN_n/dt = -A_{nm}N_n$	s^{-2}
Stimulated emission	B_{nm}	$dN_n/dt = -\rho_\nu\,(\tilde{\nu}_{nm}) \times B_{nm}N_n$	s · kg^{-1}
Stimulated absorption	B_{nm}	$dN_n/dt = \rho_{\tilde{\nu}}(\tilde{\nu}_{nm})\, B_{nm}N_n$	s · kg^{-1}
Radiant power, radiant energy per time	Φ, P	$\Phi = dQ/dt$	W
Radiant intensity	I	$I = d\Phi/d\Omega$	W · sr^{-1}
Radiant excitance (emitted radiant flux)	M	$M = d\Phi/dA_{source}$	W · m^{-2}
Irradiance (radiant flux received)	E, (I)	$E = d\Phi/dA$	W · m^{-2}
Emittance	ε	$\varepsilon = M/M_{bb}$	1
Stefan-Boltzmann constant	σ	$M_{bb} = \sigma T^4$	W · m^{-2} · K^{-4}
First radiation constant	c_1	$c_1 = 2\pi hc_0^2$	W · m^{-2}
Second radiation constant	c_2	$c_2 = hc_0/k$	K · m
Transmittance, transmission factor	τ, T	$\tau = \Phi_{tr}/\Phi_0$	1
Absorptance, absorption factor	α	$\alpha = \Phi_{abs}/\Phi_0$	1
Reflectance, reflection factor	ρ	$\rho = \Phi_{refl}/\Phi_0$	1
(Decadic) absorbance	A	$A = \lg(1 - \alpha_i)$	1
Napierian absorbance	B	$B = \ln(1 - \alpha_i)$	1
Absorption coefficient			
(Linear) decadic	a, K	$a = A/l$	m^{-1}
(Linear) napierian	α	$\alpha = B/l$	m^{-1}
Molar (decadic)	ε	$\varepsilon = a/c = A/cl$	m^2 · mol^{-1}
Molar napierian	κ	$\kappa = a/c = B/cl$	m^2 · mol^{-1}
Absorption index	k	$k = \alpha/4\pi\,\tilde{\nu}$	l
Complex refractive index	$\hat{n}$	$\hat{n} = n + ik$	1
Molar refraction	R, R_m	$R = \dfrac{(n^2-1)}{(n^2+2)} V_m$	m^3 · mol^{-1}
Angle of optical rotation	α		1, rad
		Solid State	
Lattice vector	R, R_0		m
Fundamental translation vectors for the crystal lattice	a_1; a_2; a_3, a; b; c	$R = n_1a_1 + n_2a_2 + n_3a_3$	m
(Circular) reciprocal lattice vector	G	$G \cdot R = 2\pi m$	m^{-1}
(Circular) fundamental translation vectors for the reciprocal lattice	b_1; b_2; b_3, a^*; b^*; c^*	$a_1 \cdot b_k = 2\pi\delta_{ik}$	m^{-1}
Lattice plane spacing	d		m
Bragg angle	θ	$n\lambda = 2d \sin\theta$	1, rad
Order of reflection	n		1
Order parameters			
Short range	σ		1
Long range	s		1
Burgers vector	b		m
Particle position vectort	r, R_j		m
Equilibrium position vector of an ion	R_0		m
Displacement vector of an ion	u	$u = R - R_0$	m
Debye-Waller factor	B, D		1
Debye circular wavenumber	q_D		m^{-1}
Debye circular frequency	ω_D		s^{-1}
Grüneisen parameter	γ,Γ	$\gamma = aV/\kappa C_v$	1

Name	Symbol	Definition	SI Unit
		Classical Mechanics	
Madelung constant	α, M	$E_{\text{coul}} = \dfrac{\alpha N_A z + z - e^2}{4\pi\varepsilon_0 R_0}$	1
Density of states	N_E	$N_E = dN(E)/dE$	$J^{-1} \cdot m^{-3}$
(Spectral) density of vibrational modes	N_ω, g	$N_\omega = dN(\omega)/d\omega$	$s \cdot m^{-3}$
Resistivity tensor	ρ_{ik}	$E = \rho \cdot j$	$\Omega \cdot m$
Conductivity tensor	σ_{ik}	$\sigma = \rho^{-1}$	$S \cdot m^{-1}$
Thermal conductivity tensor	λ_{ik}	$J_q = -\lambda \cdot \text{grad}\, T$	$W \cdot m^{-1} \cdot K^{-1}$
Residual resistivity	ρ_R		$\Omega \cdot m$
Relaxation time	τ	$\tau = l/v_F$	s
Lorenz coefficient	L	$L = \lambda/\sigma T$	$V^2 \cdot K^{-2}$
Hall coefficient	A_H, R_H	$E = \rho \cdot j + R_H(B \times j)$	$m^3 \cdot C^{-1}$
Thermoelectric force	E		V
Peltier coefficient	Π		V
Thomson coefficient	μ, (τ)		$V \cdot K^{-1}$
Work function	Φ	$\Phi = E_\infty - E_F$	J
Number density, number concentration	n, (p)		m^{-3}
Gap energy	E_g		J
Donor ionization energy	E_d		J
Acceptor ionization energy	E_a		J
Fermi energy	E_F, ε_F		J
Circular wave vector, propagation vector	k, q	$k = 2\pi/\lambda$	m^{-1}
Bloch function	$u_k(r)$	$\psi(r) = u_k(r) \exp(ik \cdot r)$	$m^{-3/2}$
Charge density of electrons	ρ	$\rho(r) = -e\psi^*(r)\psi(r)$	$C \cdot m^{-3}$
Effective mass	m^*		kg
Mobility	m	$\mu = v_{\text{drift}}/E$	$m^2 \cdot V^{-1} \cdot s^{-1}$
Mobility ratio	b	$b = \mu_n/\mu_p$	1
Diffusion coefficient	D	$dN/dt = -DA(dn/dx)$	$m^2 \cdot s^{-1}$
Diffusion length	L	$L = \sqrt{D\tau}$	m
Characteristic (Weiss) temperature	ϕ, ϕ_w		K
Curie temperature	T_c		K
Neel temperature	T_N		K

Elementary Algebra and Geometry

Fundamental Properties (Real Numbers)

$a + b = b + a$	Commutative law for addition
$(a + b) + c = a + (b + c)$	Associative law for addition
$a + 0 = 0 + a$	Identity law for addition
$a + (-a) = (-a) + a = 0$	Inverse law for addition
$a(bc) = (ab)c$	Associative law for multiplication
$a\left(\frac{1}{a}\right) = \left(\frac{1}{a}\right)a = 1,\ a \neq 0$	Inverse law for multiplication
$(a)(1) = (1)(a) = a$	Identity law for multiplication
$ab = ba$	Commutative law for multiplication
$a(b + c) = ab + ac$	Distributive law

Division by zero is not defined.

Exponents

For integers m and n,

$$a^n a^m = a^{n+m}$$

$$a^n / a^m = a^{n-m}$$

$$\left(a^n\right)^m = a^{nm}$$

$$(ab)^m = a^m b^m$$

$$(a/b)^m = a^m / b^m$$

Fractional Exponents

$$a^{p/q} = \left(a^{1/q}\right)^p$$

where $a^{1/q}$ is the positive qth root of a if $a > 0$ and the negative qth root of a if a is negative and q is odd. Accordingly, the five rules of exponents given above (for integers) are also valid if m and n are fractions, provided a and b are positive.

Irrational Exponents

If an exponent is irrational (e.g., $\sqrt{2}$), the quantity, such as $a^{\sqrt{2}}$, is the limit of the sequence $a^{1.4}$, $a^{1.41}$, $a^{1.414}$,…

Operations with Zero

$$0^m = 0 \qquad a^0 = 1$$

Logarithms

If x, y, and b are positive $b \neq 1$,

$$\log_b(xy) = \log_b x + \log_b y$$

$$\log_b(x/y) = \log_b x - \log_b y$$

$$\log_b x^p = p\log_b x$$

$$\log_b(1/x) = -\log_b x$$

$$\log_b b = 1$$

$$\log_b 1 = 0 \qquad \textit{Note: } b^{\log_b x} = x$$

Change of Base ($a \neq 1$)

$$\log_b x = \log_a x \log_b a$$

Factorials

The factorial of a positive integer n is the product of all the positive integers less than or equal to the integer n and is denoted $n!$. Thus,

$$n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$$

Factorial 0 is defined: 0! = 1.

Stirling's Approximation

$$\lim_{n\to\infty}\left(n/e^n\right)^n \sqrt{2\pi n} = n!$$

Binomial Theorem

For positive integer n

$$(x+y)^n = x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 + \cdots + nxy^{n-1} + y^n$$

Factors and Expansion

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$(a-b)^2 = a^2 - 2ab + b^2$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$$

$$\left(a^2 - b^2\right) = (a-b)(a+b)$$

$$\left(a^3 - b^3\right) = (a-b)\left(a^2 + ab + b^2\right)$$

$$\left(a^3 + b^3\right) = (a+b)\left(a^2 - ab + b^2\right)$$

Progression

An *arithmetic progression* is a sequence in which the difference between any term and the preceding term is a constant (d):

$$a, a+d, a+2d, \ldots, a+(n-1)d$$

If the last term is denoted l [$= a + (n - 1)d$], then the sum is

$$s = \frac{n}{2}(a+l)$$

A *geometric progression* is a sequence in which the ratio of any term to the preceding term is a constant r. Thus, for n terms,

$$a, ar, ar^2, \ldots, ar^{n-1}$$

The sum is

$$S = \frac{a - ar^n}{1-r}$$

Complex Numbers

A complex number is an ordered pair of real numbers (a, b).

Equality: $(a,b) = (c,d)$ if and only if $a = c$ and $b = d$

Addition: $(a,b) + (c,d) = (a+c, b+d)$

Multiplication: $(a,b)(c,d) = (ac - bd, ad + bc)$

The first element (a, b) is called the *real* part, the second the *imaginary* part. An alternative notation for (a, b) is $a + bi$, where $i^2 = (-1, 0)$, and $i = (0, 1)$ or $0 + 1i$ is written for this complex number as a convenience. With this understanding, i behaves as a number, that is, $(2 - 3i)(4 + i) = 8 + 2i - 12i - 3i^2 = 11 - 10i$. The conjugate of a $a + bi$ is $a - bi$, and the product of a complex number and its conjugate is $a^2 + b^2$. Thus, *quotients* are computed by multiplying numerator and denominator by the conjugate of the denominator, as illustrated below:

$$\frac{2+3i}{4+2i} = \frac{(4-2i)(2+3i)}{(4-2i)(4+2i)} = \frac{14+8i}{20} = \frac{7+4i}{10}$$

Polar Form

The complex number $x + iy$ may be represented by a plane vector with components x and y:

$$x + iy = r(\cos\theta + i\sin\theta)$$

(See Figure 19.1.1.). Then, given two complex numbers $z_1 = \tau_1(\cos\theta_1 + i \sin\theta_1)$ and $z_2 = \tau_2(\cos\theta_2 + i \sin\theta_2)$, the product and quotient are:

Product: $z_1 z_2 = r_1 r_2\left[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)\right]$

Quotient: $z_1/z_2 = (r_1/r_2)\left[\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)\right]$

Powers: $z^n = \left[r(\cos\theta + i\sin\theta)\right]^n = r^n[\cos n\theta + i\sin n\theta]$

Roots: $z^{1/n} = \left[r(\cos\theta + i\sin\theta)\right]^{1/n}$

$$= r^{1/n}\left[\cos\frac{\theta + k\cdot 360}{n} + i\sin\frac{\theta + k\cdot 360}{n}\right]$$

$k = 0,\ 1,\ 2,\ \ldots,\ n-1$

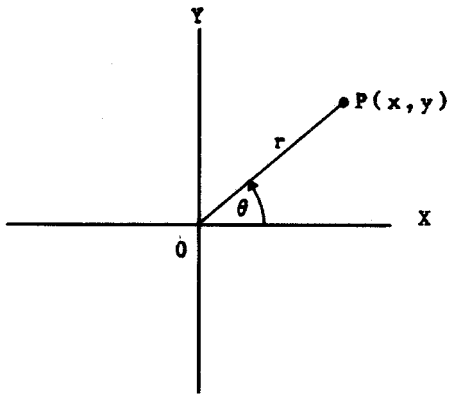

FIGURE 19.1.1 Polar form of complex number.

Permutations

A permutation is an ordered arrangement (sequence) of all or part of a set of objects. The number of permutations of n objects taken r at a time is

$$p(n,r) = n(n-1)(n-2)\cdots(n-r+1)$$

$$= \frac{n!}{(n-r)!}$$

A permutation of positive integers is "even" or "odd" if the total number of inversions is an even integer or an odd integer, respectively. Inversions are counted relative to each integer j in the permutation by counting the number of integers that follow j and are less than j. These are summed to give the total number of inversions. For example, the permutation 4132 has four inversions: three relative to 4 and one relative to 3. This permutation is therefore even.

Combinations

A combination is a selection of one or more objects from among a set of objects regardless of order. The number of combinations of n different objects taken r at a time is

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

Algebraic Equations

Quadratic

If $ax^2 + bx + c = 0$, and $a \neq 0$, then roots are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Cubic

To solve $\mathbf{x}^2 + bx^2 + cx + d = 0$, let $x = y - b/3$. Then the *reduced cubic* is obtained:

$$y^3 + py + q = 0$$

where $p = c - (1/3)b^2$ and $q = d - (1/3)bc + (2/27)b^3$. Solutions of the original cubic are then in terms of the reduced curbic roots y_1, y_2, y_3:

$$x_1 = y_1 - (1/3)b \qquad x_2 = y_2 - (1/3)b \qquad x_3 = y_3 - (1/3)b$$

The three roots of the reduced cubic are

$$y_1 = (A)^{1/3} + (B)^{1/3}$$

$$y_2 = W(A)^{1/3} + W^2(B)^{1/3}$$

$$y_3 = W^2(A)^{1/3} + W(B)^{1/3}$$

where

$$A = -\tfrac{1}{2}q + \sqrt{(1/27)p^3 + \tfrac{1}{4}q^2}$$

$$B = -\tfrac{1}{2}q - \sqrt{(1\ 27)p^3 + \tfrac{1}{4}q^2}$$

$$W = \frac{-1 + i\sqrt{3}}{2}, \qquad W^2 = \frac{-1 - i\sqrt{3}}{2}$$

When $(1/27)p^3 + (1/4)q^2$ is negative, A is complex; in this case A should be expressed in trigonometric form: $A = r(\cos\theta + i \sin\theta)$ where θ is a first or second quadrant angle, as q is negative or positive. The three roots of the reduced cubic are

$$y_1 = 2(r)^{1/3}\cos(\theta/3)$$

$$y_2 = 2(r)^{1/3}\cos\left(\frac{\theta}{3} + 120°\right)$$

$$y_3 = 2(r)^{1/3}\cos\left(\frac{\theta}{3} + 240°\right)$$

Geometry

Figures 19.1.2 to 19.1.12 are a collection of common geometric figures. Area (A), volume (V), and other measurable features are indicated.

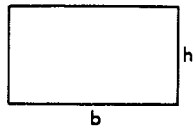

FIGURE 19.1.2 Rectangle. $A = bh$.

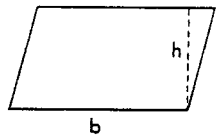

FIGURE 19.1.3 Parallelogram. $A = bh$.

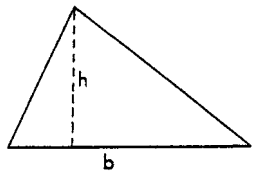

FIGURE 19.1.4 Triangle. $A = 1/2\ bh$.

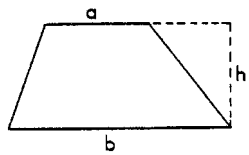

FIGURE 19.1.5 Trapezoid. $A = 1/2\ (a + b)h$.

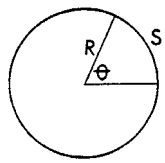

FIGURE 19.1.6 Circle. $A = \pi R^2$; circumference $= 2\pi R$, arc length $S = R\ \theta$ (θ in radians).

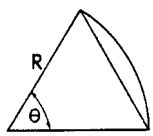

FIGURE 19.1.7 Sector of circle. $A_{\text{sector}} = 1/2R^2\theta$; $A_{\text{segment}} = 1/2R^2\ (\theta - \sin\theta)$.

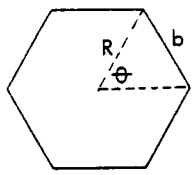

FIGURE 19.1.8 Regular polygon of n sides. $A = (n/4)b^2\ \text{ctn}(\pi/n)$; $R = (b/2)\ \csc(\pi/n)$.

Table of Derivatives

In the following table, a and n are constants, e is the base of the natural logarithms, and u and v denote functions of x.

Additional Relations with Derivatives

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t) \qquad \frac{d}{dt}\int_t^a f(x)\,dx = -f(t)$$

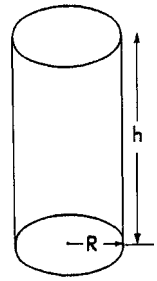

FIGURE 19.1.9 Right circular cylinder. $V = \pi R^2 h$; lateral surface area = $2\pi Rh$.

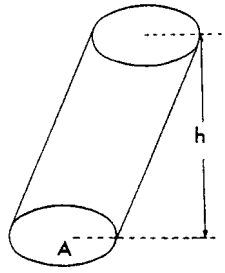

FIGURE 19.1.10 Cylinder (or prism) with parallel bases. $V = Ah$.

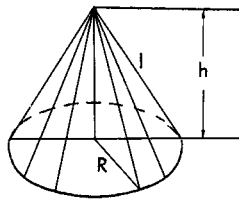

FIGURE 19.1.11 Right circular cone. $V = 1/3\ \pi R^2 h$; lateral surface area = $\pi Rl = \pi R\sqrt{R^2 + h^2}$.

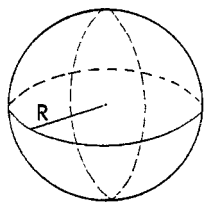

FIGURE 19.1.12 Sphere $V = 4/3\ \pi R^3$; surface area = $4\pi R^2$.

If $x = f(y)$, then $\dfrac{dy}{dx} = \dfrac{1}{dx/dy}$

If $y = f(u)$ and $u = g(x)$, then $\dfrac{dy}{dx} = \dfrac{dy}{du} \cdot \dfrac{du}{dx}$ (chain rule)

If $x = f(t)$ and $y = g(t)$, then $\dfrac{dy}{dx} = \dfrac{g'(t)}{f'(t)}$, and $\dfrac{d^2y}{dx^2} = \dfrac{f'(t)g''(t) - g'(t)f''(t)}{\left[f'(t)\right]^3}$

(Note: Exponent in denominator is 3.)

1. $\frac{d}{dx}(a) = 0$
2. $\frac{d}{dx}(x) = 1$
3. $\frac{d}{dx}(au) = a\frac{du}{dx}$
4. $\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$
5. $\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$
6. $\frac{d}{dx}(u/v) = \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$
7. $\frac{d}{dx}(u^n) = nu^{n-1}\frac{du}{dx}$
8. $\frac{d}{dx}e^u = e^u\frac{du}{dx}$
9. $\frac{d}{dx}a^u = (\log_e a)a^u\frac{du}{dx}$
10. $\frac{d}{dx}\log_e u = (1/u)\frac{du}{dx}$
11. $\frac{d}{dx}\log_a u = (\log_a e)(1/u)\frac{du}{dx}$
12. $\frac{d}{dx}u^v = vu^{v-1}\frac{du}{dx} + u^v(\log_e u)\frac{dv}{dx}$
13. $\frac{d}{dx}\sin u = \cos u\frac{du}{dx}$
14. $\frac{d}{dx}\cos u = -\sin u\frac{du}{dx}$
15. $\frac{d}{dx}\tan u = \sec^2 u\frac{du}{dx}$
16. $\frac{d}{dx}\operatorname{ctn} u = -\csc^2 u\frac{du}{dx}$
17. $\frac{d}{dx}\sec u = \sec u\tan u\frac{du}{dx}$
18. $\frac{d}{dx}\csc u = -\csc u\operatorname{ctn} u\frac{du}{dx}$
19. $\frac{d}{dx}\sin^{-1} u = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$, $\left(-\frac{1}{2}\pi \le \sin^{-1} u \le \frac{1}{2}\pi\right)$
20. $\frac{d}{dx}\cos^{-1} u = \frac{-1}{\sqrt{1-u^2}}\frac{du}{dx}$, $\left(0 \le \cos^{-1} u \le \pi\right)$
21. $\frac{d}{dx}\tan^{-1} u = \frac{1}{1+u^2}\frac{du}{dx}$
22. $\frac{d}{dx}\operatorname{ctn}^{-1} u = \frac{-1}{1+u^2}\frac{du}{dx}$
23. $\frac{d}{dx}\sec^{-1} u = \frac{1}{u\sqrt{u^2-1}}\frac{du}{dx}$,
$\left(-\pi \le \sec^{-1} u < -\frac{1}{2}\pi;\quad 0 \le \sec^{-1} u \le \frac{1}{2}\pi\right)$
24. $\frac{d}{dx}\csc^{-1} u = \frac{-1}{u\sqrt{u^2-1}}\frac{du}{dx}$,
$\left(-\pi < \csc^{-1} u \le -\frac{1}{2}\pi;\quad 0 < \csc^{-1} u \le \frac{1}{2}\pi\right)$
25. $\frac{d}{dx}\sinh u = \cosh u\frac{du}{dx}$
26. $\frac{d}{dx}\cosh u = \sinh u\frac{du}{dx}$
27. $\frac{d}{dx}\tanh u = \operatorname{sech}^2 u\frac{du}{dx}$
28. $\frac{d}{dx}\operatorname{ctnh} u = -\operatorname{csch}^2 u\frac{du}{dx}$
29. $\frac{d}{dx}\operatorname{sech} u = -\operatorname{sech} u\tanh u\frac{du}{dx}$
30. $\frac{d}{dx}\operatorname{csch} u = -\operatorname{csch} u\operatorname{ctnh} u\frac{du}{dx}$
31. $\frac{d}{dx}\sin^{-1} u = \frac{1}{\sqrt{u^2+1}}\frac{du}{dx}$
32. $\frac{d}{dx}\cosh^{-1} u = \frac{1}{\sqrt{u^2-1}}\frac{du}{dx}$
33. $\frac{d}{dx}\tanh^{-1} u = \frac{1}{1-u^2}\frac{du}{dx}$
34. $\frac{d}{dx}\operatorname{ctnh}^{-1} u = \frac{-1}{u^2-1}\frac{du}{dx}$
35. $\frac{d}{dx}\operatorname{sech}^{-1} u = \frac{-1}{u\sqrt{1-u^2}}\frac{du}{dx}$
36. $\frac{d}{dx}\operatorname{csch}^{-1} u = \frac{-1}{u\sqrt{u^2+1}}\frac{du}{dx}$

Integrals

Elementary Forms (Add an arbitrary constant to each integral)

1. $\int a\,dx = ax$
2. $\int a \cdot f(x)\,dx = a\int f(x)\,dx$
3. $\int \phi(y)\,dx = \int \frac{\phi(y)}{y'}\,dy, \quad \text{where } y' = \frac{dy}{dx}$
4. $\int (u+v)\,dx = \int u\,dx + \int v\,dx, \quad \text{where } u \text{ and } v \text{ are any functions of } x$
5. $\int u\,dv = u\int dv - \int v\,du = uv - \int v\,du$
6. $\int u\,\frac{dv}{dx}\,dx = uv - \int v\,\frac{du}{dx}\,dx$
7. $\int x^n\,dx = \frac{x^{n+1}}{n+1}, \quad \text{except } n = -1$
8. $\int \frac{f'(x)\,dx}{f(x)} = \log f(x), \quad \left[df(x) = f'(x)\,dx\right]$
9. $\int \frac{dx}{x} = \log x$
10. $\int \frac{f'(x)\,dx}{2\sqrt{f(x)}} = \sqrt{f(x)}, \quad \left[df(x) = f'(x)\,dx\right]$
11. $\int e^x\,dx = e^x$
12. $\int e^{ax}\,dx = e^{ax}/a$
13. $\int b^{ax}\,dx = \frac{b^{ax}}{a\log b}, \quad (b > 0)$
14. $\int \log x\,dx = x\log x - x$
15. $\int a^x \log a\,dx = a^x, \quad (a > 0)$
16. $\int \frac{dx}{a^2 + x^2} = \frac{1}{a}\tan^{-1}\frac{x}{a}$
17. $\int \frac{dx}{a^2 - x^2} = \begin{cases} \frac{1}{a}\tan^{-1}\frac{x}{a} \\ \quad \text{or} \\ \frac{1}{2a}\log\frac{a+x}{a-x}, \quad \left(a^2 > x^2\right) \end{cases}$
18. $\int \frac{dx}{x^2 - a^2} = \begin{cases} -\frac{1}{a}\operatorname{ctnh}^{-1}\frac{x}{a} \\ \quad \text{or} \\ \frac{1}{2a}\log\frac{x-a}{x+a}, \quad \left(x^2 > a^2\right) \end{cases}$

19. $$\int \frac{dx}{\sqrt{a^2 - x^2}} = \begin{cases} \sin^{-1}\dfrac{x}{|a|} \\ \text{or} \\ -\cos^{-1}\dfrac{x}{|a|}, \quad \left(a^2 > x^2\right) \end{cases}$$

20. $$\int \frac{dx}{\sqrt{x^2 \pm a^2}} = \log\left(x + \sqrt{x^2 \pm a^2}\right)$$

21. $$\int \frac{dx}{x\sqrt{x^2 - a^2}} = \frac{1}{|a|}\sec^{-1}\frac{x}{|a|}$$

22. $$\int \frac{dx}{x\sqrt{a^2 \pm x^2}} = -\frac{1}{a}\log\left(\frac{a + \sqrt{a^2 \pm x^2}}{x}\right)$$

Forms Containing ($a + bx$)

For forms containing $a + bx$, but not listed in the table, the substitution $u = (a + bx)\,x$ may prove helpful.

23. $$\int (a+bx)^n\,dx = \frac{(a+bx)^{n+1}}{(n+1)b}, \quad (n \neq -1)$$

24. $$\int x(a+bx)^n\,dx = \frac{1}{b^2(n+2)}(a+bx)^{n+2} - \frac{a}{b^2(n+1)}(a+bx)^{n+1}, \quad (n \neq -1, -2)$$

25. $$\int x^2(a+bx)^n\,dx = \frac{1}{b^3}\left[\frac{(a+bx)^{n+3}}{n+3} - 2a\frac{(a+bx)^{n+2}}{n+2} + a^2\frac{(a+bx)^{n+1}}{n+1}\right]$$

26. $$\int x^m(a+bx)^n\,dx = \begin{cases} \dfrac{x^{m+1}(a+bx)^n}{m+n+1} + \dfrac{an}{m+n+1}\displaystyle\int x^m(a+bx)^{n-1}\,dx \\ \text{or} \\ \dfrac{1}{a(n+1)}\left[-x^{m+1}(a+bx)^{n+1} + (m+n+2)\displaystyle\int x^m(a+bx)^{n+1}\,dx\right] \\ \text{or} \\ \dfrac{1}{b(m+n+1)}\left[x^m(a+bx)^{n+1} - ma\displaystyle\int x^{m-1}(a+bx)^n\,dx\right] \end{cases}$$

27. $$\int \frac{dx}{a+bx} = \frac{1}{b}\log(a+bx)$$

28. $$\int \frac{dx}{(a+bx)^2} = -\frac{1}{b(a+bx)}$$

29. $$\int \frac{dx}{(a+bx)^3} = -\frac{1}{2b(a+bx)^2}$$

30. $$\int \frac{x\,dx}{a+bx} = \begin{cases} \dfrac{1}{b^2}\left[a + bx - a\log(a+bx)\right] \\ \text{or} \\ \dfrac{x}{b} - \dfrac{a}{b^2}\log(a+bx) \end{cases}$$

31. $$\int \frac{x\,dx}{(a+bx)^2} = \frac{1}{b^2}\left[\log(a+bx) + \frac{a}{a+bx}\right]$$

32. $\int \frac{x\,dx}{(a+bx)^n} = \frac{1}{b^2}\left[\frac{-1}{(n-2)(a+bx)^{n-2}} + \frac{a}{(n-1)(a+bx)^{n-1}}\right], \quad n \neq 1, 2$

33. $\int \frac{x^2\,dx}{a+bx} = \frac{1}{b^3}\left[\frac{1}{2}(a+bx)^2 - 2a(a+bx) + a^2\log(a+bx)\right]$

34. $\int \frac{x^2\,dx}{(a+bx)^2} = \frac{1}{b^3}\left[a+bx-2a\log(a+bx) - \frac{a^2}{a+bx}\right]$

35. $\int \frac{x^2\,dx}{(a+bx)^3} = \frac{1}{b^3}\left[\log(a+bx) + \frac{2a}{a+bx} - \frac{a^2}{2(a+bx)^2}\right]$

36. $\int \frac{x^2\,dx}{(a+bx)^n} = \frac{1}{b^3}\left[\frac{-1}{(n-3)(a+bx)^{n-3}} + \frac{2a}{(n-2)(a+bx)^{n-2}} - \frac{a}{(n-1)(a+bx)^{n-1}}\right], \quad n \neq 1, 2, 3$

37. $\int \frac{dx}{x(a+bx)} = -\frac{1}{a}\log\frac{a+bx}{x}$

38. $\int \frac{dx}{x(a+bx)^2} = \frac{1}{a(a+bx)} - \frac{1}{a^2}\log\frac{a+bx}{x}$

39. $\int \frac{dx}{x(a+bx)^3} = \frac{1}{a^3}\left[\frac{1}{2}\left(\frac{2a+bx}{a+bx}\right)^2 + \log\frac{x}{a+bx}\right]$

40. $\int \frac{dx}{x^2(a+bx)} = -\frac{1}{ax} + \frac{b}{a^2}\log\frac{a+bx}{x}$

41. $\int \frac{dx}{x^3(a+bx)} = \frac{2bx-a}{2a^2x^2} + \frac{b^2}{a^3}\log\frac{x}{a+bx}$

42. $\int \frac{dx}{x^2(a+bx)^2} = -\frac{a+2bx}{a^2x(a+bx)} + \frac{2b^2}{a^3}\log\frac{a+bx}{x}$

The Fourier Transforms

For a piecewise continuous function $F(x)$ over a finite interval $0 \le x \le \pi$, the *finite Fourier cosine transform* of $F(x)$ is

$$f_c(n) = \int_0^\pi F(x)\cos nx\,dx \quad (n = 0,\ 1,\ 2,\ \ldots) \tag{19.1.1}$$

If x ranges over the interval $0 \le x \le L$, the substitution $x' = \pi x/L$ allows the use of this definition also. The inverse transform is written

$$\overline{F}(x) = \frac{1}{\pi}f_c(0) + \frac{2}{\pi}\sum_{n=1}^{\infty} f_c(n)\cos nx \quad (0 < x < \pi) \tag{19.1.2}$$

where $\overline{F}(x) = [F(x+0) + F(x-0)]/2$. We observe that $\overline{F}(x) = F(x)$ at points of continuity. The formula

$$\begin{aligned} f_c^{(2)}(n) &= \int_0^\pi F''(x)\cos nx\,dx \\ &= -n^2 f_c(n) - F'(0) + (-1)^n F'(\pi) \end{aligned} \tag{19.1.3}$$

makes the finite Fourier cosine transform useful in certain boundary value problems.

Analogously, the *finite Fourier sine transform* of $F(x)$ is

$$f_s(n) = \int_0^{\pi} F(x)\sin nx\, dx \quad (n = 1,\ 2,\ 3,\ \ldots) \tag{19.1.4}$$

and

$$\bar{F}(x) = \frac{2}{\pi}\sum_{n=1}^{\infty} f_s(n)\sin nx \quad (0 < x < \pi) \tag{19.1.5}$$

Corresponding to Equation (19.1.6), we have

$$\begin{aligned} f_s^{(2)}(n) &= \int_0^{\pi} F''(x)\sin nx\, dx \\ &= -n^2 f_s(n) - nF(0) - n(-1)^n F(\pi) \end{aligned} \tag{19.1.6}$$

Fourier Transforms

If $F(x)$ is defined for $x \geq 0$ and is piecewise continuous over any finite interval, and if

$$\int_0^{\infty} F(x)\, dx$$

is absolutely convergent, then

$$f_c(\alpha) = \sqrt{\frac{2}{\pi}}\int_0^{\infty} F(x)\cos(\alpha x)\, dx \tag{19.1.7}$$

is the *Fourier cosine transform* of $F(x)$. Furthermore,

$$\bar{F}(x) = \sqrt{\frac{2}{\pi}}\int_0^{\infty} f_c(\alpha)\cos(\alpha x)\, d\alpha \tag{19.1.8}$$

If $\lim_{x\to\infty} d^nF/dx^n = 0$, an important property of the Fourier cosine transform,

$$\begin{aligned} f_c^{(2r)}(\alpha) &= \sqrt{\frac{2}{\pi}}\int_0^{\infty}\left(\frac{d^{2r}F}{dx^{2r}}\right)\cos(\alpha x)\, dx \\ &= -\sqrt{\frac{2}{\pi}}\sum_{n=0}^{r-1}(-1)^n a_{2r-2n-1}\alpha^{2n} + (-1)^r\alpha^{2r} f_c(\alpha) \end{aligned} \tag{19.1.9}$$

where $\lim_{x\to 0} d^rF/dx^r = a_r$, makes it useful in the solution of many problems.

Under the same conditions,

$$f_s(\alpha) = \sqrt{\frac{2}{\pi}} \int_0^\infty F(x) \sin(\alpha x)\, dx \tag{19.1.10}$$

defines the *Fourier sine transform* of *F(x)*, and

$$\overline{F}(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty f_s(\alpha) \sin(\alpha x)\, d\alpha \tag{19.1.11}$$

Corresponding to Equation (19.1.9) we have

$$\begin{aligned} f_s^{(2r)}(\alpha) &= \sqrt{\frac{2}{\pi}} \int_0^\infty \frac{d^{2r}F}{dx^{2r}} \sin(\alpha x)\, dx \\ &= -\sqrt{\frac{2}{\pi}} \sum_{n=1}^{r} (-1)^n \alpha^{2n-1} a_{2r-2n} + (-1)^{r-1} \alpha^{2r} f_s(\alpha) \end{aligned} \tag{19.1.12}$$

Similarly, if *F(x)* is defined for $-\infty < x < \infty$, and if $\int_{-\infty}^{\infty} F(x)\, dx$ is absolutely convergent, then

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(x) e^{i\alpha x}\, dx \tag{19.1.13}$$

is the *Fourier transform* of *F(x)*, and

$$\overline{F}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\alpha) e^{-i\alpha x}\, d\alpha \tag{19.1.14}$$

Also, if

$$\lim_{|x| \to \infty} \left| \frac{d^n F}{dx^n} \right| = 0 \quad (n = 1,\ 2,\ \ldots,\ r-1)$$

then

$$f^{(r)}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F^{(r)}(x) e^{i\alpha x}\, dx = (-i\alpha)^r f(\alpha) \tag{19.1.15}$$

Finite Sine Transforms

	$f_s(n)$	$F(x)$
1.	$f_s(n)=\int_0^{\pi} F(x)\sin nx\,dx \quad (n=1,\ 2,\ \ldots)$	$F(x)$
2.	$(-1)^{n+1} f_s(n)$	$F(\pi - x)$
3.	$\dfrac{1}{n}$	$\dfrac{\pi - x}{\pi}$
4.	$\dfrac{(-1)^{n+1}}{n}$	$\dfrac{x}{\pi}$
5.	$\dfrac{1-(-1)^n}{n}$	1
6.	$\dfrac{2}{n^2}\sin\dfrac{n\pi}{2}$	$\begin{cases} x & \text{when } 0 < x < \pi/2 \\ \pi - x & \text{when } \pi/2 < x < \pi \end{cases}$
7.	$\dfrac{(-1)^{n+1}}{n^3}$	$\dfrac{x\left(\pi^2 - x^2\right)}{6\pi}$
8.	$\dfrac{1-(-1)^n}{n^3}$	$\dfrac{x(\pi - x)}{2}$
9.	$\dfrac{\pi^2(-1)^{n-1}}{n} - \dfrac{2\left[1-(-1)^n\right]}{n^3}$	x^2
10.	$\pi(-1)^n\left(\dfrac{6}{n^3} - \dfrac{\pi^2}{n}\right)$	x^3
11.	$\dfrac{n}{n^2+c^2}\left[1-(-1)^n e^{c\pi}\right]$	e^{cx}
12.	$\dfrac{n}{n^2+c^2}$	$\dfrac{\sinh c(\pi - x)}{\sinh c\pi}$
13.	$\dfrac{n}{n^2-k^2} \quad (k \neq 0,\ 1,\ 2,\ \ldots)$	$\dfrac{\sinh k(\pi - x)}{\sinh k\pi}$
14.	$\begin{cases} \dfrac{\pi}{2} & \text{when } n = m \\ 0 & \text{when } n \neq m \end{cases} \quad (m=1,\ 2,\ \ldots)$	$\sin mx$
15.	$\dfrac{n}{n^2-k^2}\left[1-(-1)^n\cos k\pi\right] \quad (k \neq 1,\ 2,\ \ldots)$	$\cos kx$
16.	$\begin{cases} \dfrac{n}{n^2-m^2}\left[1-(-1)^{n+m}\right] & \text{when } n \neq m = 1,\ 2,\ \ldots \\ 0 & \text{when } n = m \end{cases}$	$\cos mx$
17.	$\dfrac{n}{\left(n^2-k^2\right)^2} \quad (k \neq 0,\ 1,\ 2,\ \ldots)$	$\dfrac{\pi\sin kx}{2k\sin^2 kx} - \dfrac{x\cos k(\pi - x)}{2k\sin k\pi}$
18.	$\dfrac{b^n}{n} \quad (\lvert b\rvert \leq 1)$	$\dfrac{2}{\pi}\arctan\dfrac{b\sin x}{1-b\cos x}$
19.	$\dfrac{1-(-1)^n}{n}b^n \quad (\lvert b\rvert \leq 1)$	$\dfrac{2}{\pi}\arctan\dfrac{2b\sin x}{1-b^2}$

Finite Cosine Transforms

	$f_c(n)$	$F(x)$
1.	$f_c(n)=\int_0^{\pi} F(x)\cos nx\,dx \quad (n=0,\ 1,\ 2,\ \ldots)$	$F(x)$
2.	$(-1)^n f_c(n)$	$F(\pi - x)$
3.	0 when $n=1,\ 2,\ \ldots;\quad f_c(0)=\pi$	1
4.	$\frac{2}{n}\sin\frac{n\pi}{2};\quad f_c(0)=0$	$\begin{cases}1 & \text{when } 0<x<\pi/2\\ -1 & \text{when } \pi/2<x<\pi\end{cases}$
5.	$-\frac{1-(-1)^n}{n^2};\quad f_c(0)=\frac{\pi^2}{2}$	x
6.	$\frac{(-1)^n}{n^2};\quad f_c(0)=\frac{\pi^2}{6}$	$\frac{x^2}{2\pi}$
7.	$\frac{1}{n^2};\quad f_c(0)=0$	$\frac{(\pi-x)^2}{2\pi}-\frac{\pi}{6}$
8.	$3\pi^2\frac{(-1)^n}{n^2}-6\frac{1-(-1)^n}{n^4};\quad f_c(0)=\frac{\pi^4}{4}$	x^3
9.	$\frac{(-1)^n e^c\pi-1}{n^2+c^2}$	$\frac{1}{c}e^{cx}$
10.	$\frac{1}{n^2+c^2}$	$\frac{\cosh c(\pi-x)}{c\sinh c\pi}$
11.	$\frac{k}{n^2-k^2}\left[(-1)^n\cos\pi k-1\right] \quad (k\neq 0,\ 1,\ 2,\ \ldots)$	$\sin kx$
12.	$\frac{(-1)^{n+m}-1}{n^2-m^2};\quad f_c(m)=0 \quad (m=1,\ 2,\ \ldots)$	$\frac{1}{m}\sin mx$
13.	$\frac{1}{n^2-k^2} \quad (k\neq 0,\ 1,\ 2,\ \ldots)$	$-\frac{\cos k(\pi-x)}{k\sin k\pi}$
14.	0 when $n=1,\ 2,\ \ldots;\quad f_c(m)=\frac{\pi}{2} \quad (m=1,\ 2,\ \ldots)$	$\cos mx$

Fourier Sine Transforms

	F(x)	$f_s(\alpha)$
1.	$\begin{cases}1 & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\sqrt{\frac{2}{\pi}}\left[\frac{1-\cos\alpha}{\alpha}\right]$
2.	$x^{p-1} \quad (0<p<1)$	$\sqrt{\frac{2}{\pi}}\frac{\Gamma(p)}{\alpha^p}\sin\frac{p\pi}{2}$
3.	$\begin{cases}\sin x & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\frac{1}{\sqrt{2\pi}}\left[\frac{\sin[a(1-\alpha)]}{1-\alpha}-\frac{\sin[a(1+\alpha)]}{1+\alpha}\right]$
4.	e^{-x}	$\sqrt{\frac{2}{\pi}}\left[\frac{\alpha}{1+\alpha^2}\right]$
5.	$xe^{-x^2/2}$	$\alpha e^{-\alpha^2/2}$

F(x)	$f_s(\alpha)$
6. $\cos\frac{x^2}{2}$	$\sqrt{2}\left[\sin\frac{\alpha^2}{2}C\left(\frac{\alpha^2}{2}\right)-\cos\frac{\alpha^2}{2}S\left(\frac{\alpha^2}{2}\right)\right]^*$
7. $\sin\frac{x^2}{2}$	$\sqrt{2}\left[\cos\frac{\alpha^2}{2}C\left(\frac{\alpha^2}{2}\right)+\sin\frac{\alpha^2}{2}S\left(\frac{\alpha^2}{2}\right)\right]^*$

* $C(y)$ and $S(y)$ are the Fresnel integrals

$$C(y)=\frac{1}{\sqrt{2\pi}}\int_0^y\frac{1}{\sqrt{t}}\cos t\,dt$$

$$S(y)=\frac{1}{\sqrt{2\pi}}\int_0^y\frac{1}{\sqrt{t}}\sin t\,dt$$

Fourier Cosine Transforms

F(x)	$f_c(\alpha)$
1. $\begin{cases}1 & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\sqrt{\frac{2}{\pi}}\frac{\sin a\alpha}{\alpha}$
2. $x^{p-1}\quad(0<p<1)$	$\sqrt{\frac{2}{\pi}}\frac{\Gamma(p)}{\alpha^p}\cos\frac{p\pi}{2}$
3. $\begin{cases}\cos x & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\frac{1}{\sqrt{2\pi}}\left[\frac{\sin[a(1-\alpha)]}{1-\alpha}+\frac{\sin[a(1+\alpha)]}{1+\alpha}\right]$
4. e^{-x}	$\sqrt{\frac{2}{\pi}}\left(\frac{1}{1+\alpha^2}\right)$
5. $e^{-x^2/2}$	$e^{-\alpha^2/2}$
6. $\cos\frac{x^2}{2}$	$\cos\left(\frac{\alpha^2}{2}-\frac{\pi}{4}\right)$
7. $\sin\frac{x^2}{2}$	$\cos\left(\frac{\alpha^2}{2}-\frac{\pi}{4}\right)$

Fourier Transforms

	$F(x)$	$f(\alpha)$
1.	$\dfrac{\sin ax}{x}$	$\begin{cases}\sqrt{\dfrac{\pi}{2}} & \lvert\alpha\rvert < a \\ 0 & \lvert\alpha\rvert > a\end{cases}$
2.	$\begin{cases}e^{iwx} & (p < x < q) \\ 0 & (x < p, x > q)\end{cases}$	$\dfrac{i}{\sqrt{2\pi}}\dfrac{e^{ip(w+\alpha)} - e^{iq(w+\alpha)}}{(w+\alpha)}$
3.	$\begin{cases}e^{-cx+iwx} & (x > 0) \\ 0 & (x < 0)\end{cases}$ $(c > 0)$	$\dfrac{i}{\sqrt{2\pi}(w+\alpha+ic)}$
4.	e^{-px^2} $R(p) > 0$	$\dfrac{1}{\sqrt{2p}}e^{-\alpha^2/4p}$
5.	$\cos px^2$	$\dfrac{1}{\sqrt{2p}}\cos\left[\dfrac{\alpha^2}{4p} - \dfrac{\pi}{4}\right]$
6.	$\sin px^2$	$\dfrac{1}{\sqrt{2p}}\cos\left[\dfrac{\alpha^2}{4p} + \dfrac{\pi}{4}\right]$
7.	$\lvert x\rvert^{-p}$ $(0 < p < 1)$	$\sqrt{\dfrac{2}{\pi}}\dfrac{\Gamma(1-p)\sin\dfrac{p\pi}{2}}{\lvert\alpha\rvert^{(1-p)}}$
8.	$\dfrac{e^{-a\lvert x\rvert}}{\sqrt{\lvert x\rvert}}$	$\dfrac{\sqrt{\sqrt{(a^2+\alpha^2)}+a}}{\sqrt{a^2+\alpha^2}}$
9.	$\dfrac{\cosh ax}{\cosh \pi x}$ $(-\pi < a < \pi)$	$\sqrt{\dfrac{2}{\pi}}\dfrac{\cos\dfrac{a}{2}\cosh\dfrac{\alpha}{2}}{\cosh\alpha + \cos a}$
10.	$\dfrac{\sinh ax}{\sinh \pi x}$ $(-\pi < a < \pi)$	$\dfrac{1}{\sqrt{2\pi}}\dfrac{\sin a}{\cosh\alpha + \cos a}$
11.	$\dfrac{1}{\sqrt{a^2-x^2}}$ $(\lvert x\rvert < a)$ 0 $(\lvert x\rvert > a)$	$\sqrt{\dfrac{\pi}{2}}J_0(a\alpha)$
12.	$\dfrac{\sin\left[b\sqrt{a^2+x^2}\right]}{\sqrt{a^2+x^2}}$	$\begin{cases}0 & (\lvert\alpha\rvert > b) \\ \sqrt{\dfrac{\pi}{2}}J_0\left(\sqrt{b^2-\alpha^2}\right) & (\lvert\alpha\rvert < b)\end{cases}$
13.	$\begin{cases}P_n(x) & (\lvert x\rvert < 1) \\ 0 & (\lvert x\rvert > 1)\end{cases}$	$\dfrac{i^n}{\sqrt{\alpha}}j_{n+1/2}(\alpha)$
14.	$\begin{cases}\dfrac{\cos\left[b\sqrt{a^2-x^2}\right]}{\sqrt{a^2-x^2}} & (\lvert x\rvert < a) \\ 0 & (\lvert x\rvert > a)\end{cases}$	$\sqrt{\dfrac{\pi}{2}}J_0\left(a\sqrt{a^2+b^2}\right)$
15.	$\begin{cases}\dfrac{\cosh\left[b\sqrt{a^2-x^2}\right]}{\sqrt{a^2-x^2}} & (\lvert x\rvert < a) \\ 0 & (\lvert x\rvert > a)\end{cases}$	$\sqrt{\dfrac{\pi}{2}}J_0\left(a\sqrt{a^2-b^2}\right)$

The following functions appear among the entries of the tables on transforms.

Function	Definition	Name
Ei(x)	$\int_{-\infty}^{x} \frac{e^v}{v}\,dv$; or sometimes defined as $-\mathrm{Ei}(-x) = \int_{x}^{\infty} \frac{e^{-v}}{v}\,dv$	Exponential integral function
Si(x)	$\int_{0}^{x} \frac{\sin v}{v}\,dv$	Sine integral function
Ci(x)	$\int_{\infty}^{x} \frac{\cos v}{v}\,dv$; or sometimes defined as negative of this integral	Cosine integral function
erf(x)	$\frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-v^2}\,dv$	Error function
erfc(x)	$1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-v^2}\,dv$	Complementary function to error function
$L_n(x)$	$\frac{e^x}{n!} \frac{d^n}{dx^n}\left(x^n e^{-x}\right)$, $n = 0, 1, \ldots$	Laguerre polynomial of degree n

Bessel Functions

Bessel Functions of the First Kind, $J_n(x)$ (Also Called Simply *Bessel Functions*) (Figure 19.1.13)

Domain: [$x > 0$]
Recurrence relation:

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x), \quad n = 0, 1, 2, \ldots$$

Symmetry: $J_{-n}(x) = (-1)^n J_n(x)$

0. $J_0(20x)$
1. $J_1(20x)$
2. $J_2(20x)$
3. $J_3(20x)$
4. $J_4(20x)$
5. $J_5(20x)$

Bessel Functions of the Second Kind, $Y_n(x)$ (Also Called *Neumann Functions* or *Weber Functions*) (Figure 19.1.14)

Domain: [$x > 0$]
Recurrence relation:

$$Y_{n+1}(x) = \frac{2n}{x} Y_n(x) - Y_{n-1}(x), \quad n = 0, 1, 2, \ldots$$

Symmetry: $Y_{-n}(x) = (-1)^n Y_n(x)$

0. $Y_0(20x)$
1. $Y_1(20x)$
2. $Y_2(20x)$
3. $Y_3(20x)$
4. $Y_4(20x)$
5. $Y_5(20x)$

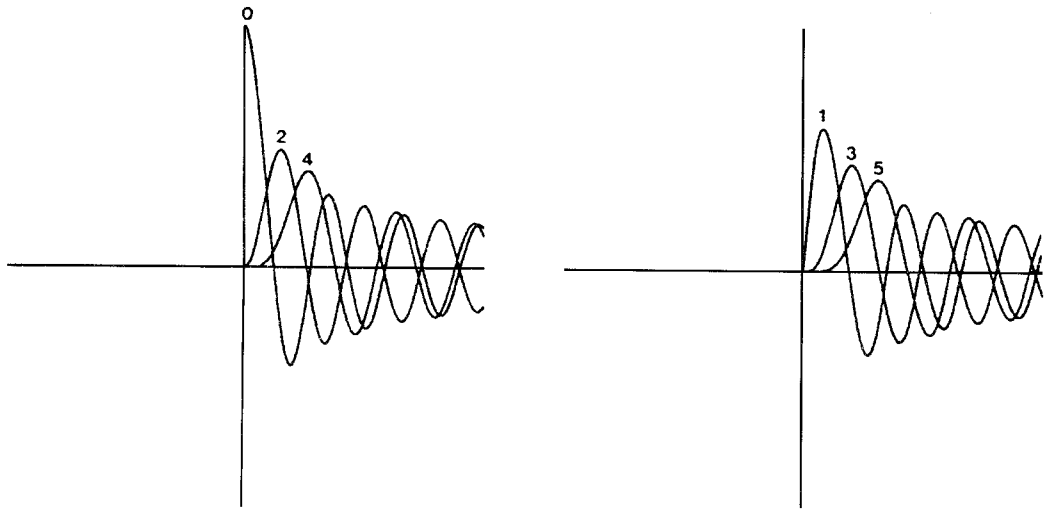

FIGURE 19.1.13 Bessel functions of the first kind.

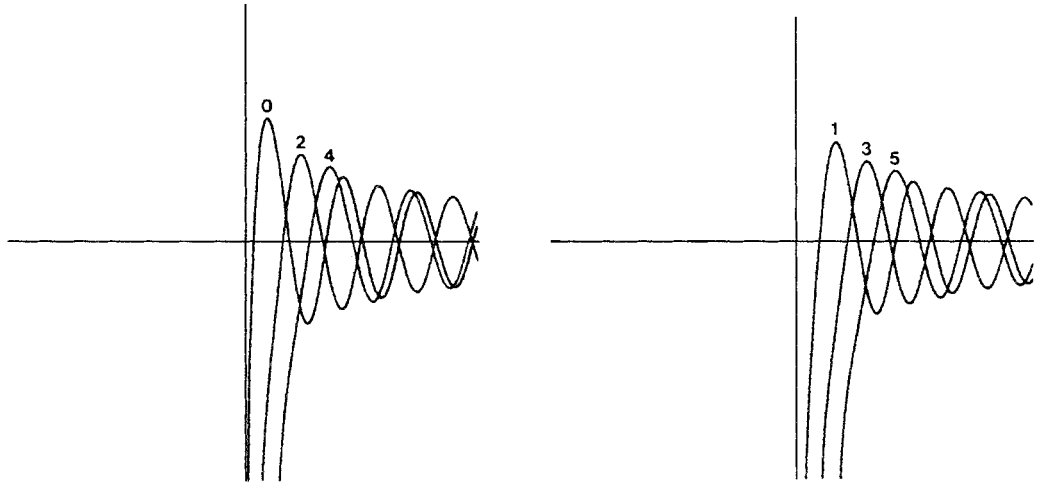

FIGURE 19.1.14 Bessel functions of the second kind.

Legendre Functions

Associated Legendre Functions of the First Kind, $P_n^m(x)$ (Figure 19.1.15)

Domain: $[-1 < x < 1]$
Recurrence relations:

$$P_{n+1}^m(x) = \frac{(2n+1)xP_n^m - (n+m)P_{n-1}^m(x)}{n-m+1}, \quad n = 1, 2, 3, \ldots$$

$$P_n^{m+1}(x) = \left(x^2 - 1\right)^{-1/2}\left[(n-m)xP_n^m(x) - (n+m)P_{n-1}^m(x)\right], \quad m = 0, 1, 2, \ldots$$

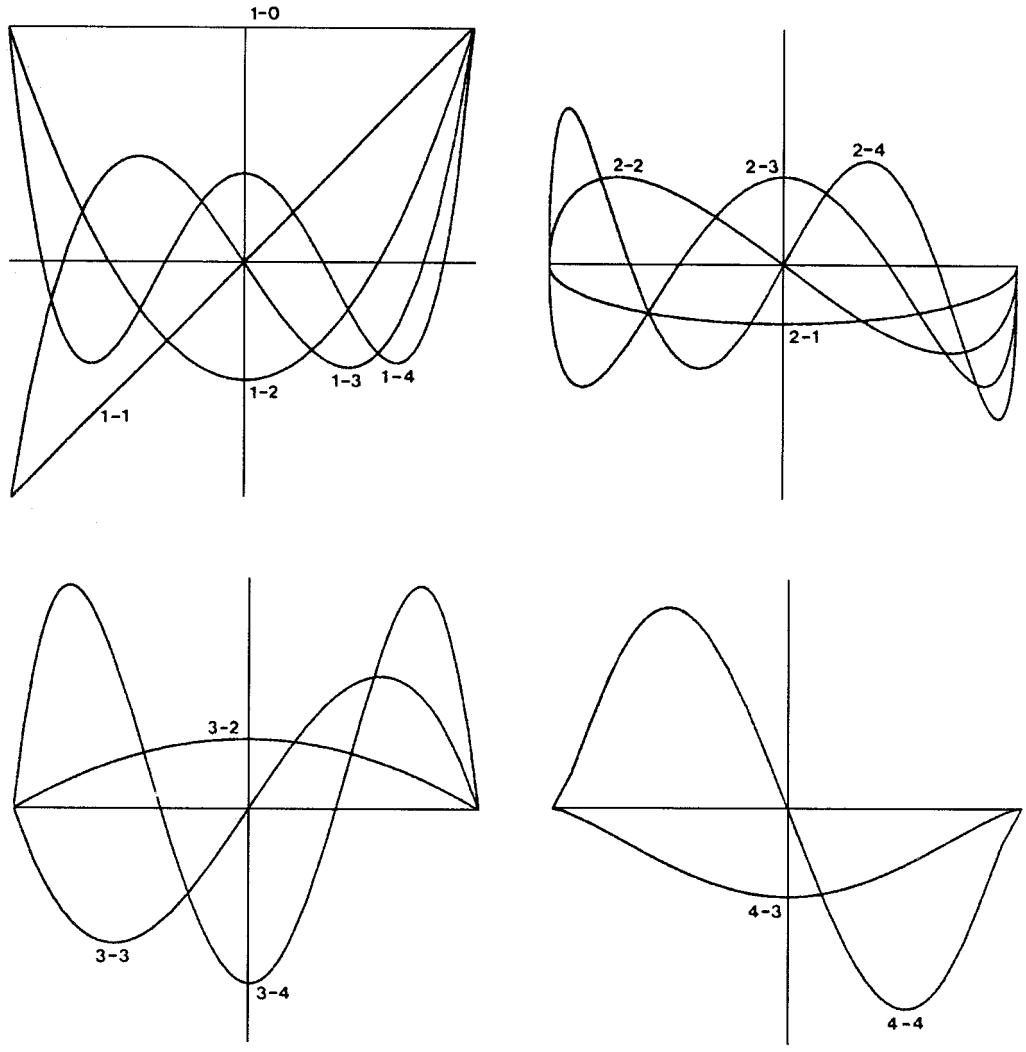

FIGURE 19.1.15 Legendre functions of the first kind.

with

$$P_0^0 = 1 \qquad P_1^0 = x$$

Special case: P_n^0 = Legendre polynomials

1-0. $P_0^0(x)$
1-1. $P_1^0(x)$ 2-1. 0.25 $P_1^1(x)$
1-2. $P_2^0(x)$ 2-2. 0.25 $P_2^1(x)$ 3-2. 0.10 $P_2^2(x)$
1-3. $P_3^0(x)$ 2-3. 0.25 $P_3^1(x)$ 3-3. 0.10 $P_3^2(x)$ 4-3. 0.025 $P_3^3(x)$
1-4. $P_4^0(x)$ 2-4. 0.25 $P_4^1(x)$ 3-4. 0.10 $P_4^2(x)$ 4-4. 0.025 $P_4^3(x)$

Table of Differential Equations

Equation	Solution
1. $y' = \dfrac{dy}{dx} = f(x)$	$y = \int f(x)\,dx + c$
2. $y' + p(x)y = q(x)$	$y = \exp[-\int p(x)\,dx]\{c + \int \exp[\int p(x)\,dx]q(x)dx\}$
3. $y' + p(x)y = q(x)y^{\alpha}$ $\alpha \neq 0,\ \alpha \neq 1$	Set $z = y^{1-\alpha} \rightarrow z' + (1-\alpha)p(x)z = (1-\alpha)q(x)$ and use 2
4. $y' = f(x)g(y)$	Integrate $\dfrac{dy}{g(y)} = f(x)\,dx$ (separable)
5. $\dfrac{dy}{dx} = f(x/y)$	Set $y = xu \rightarrow u + x\dfrac{du}{dx} = f(u)$
	$\int \dfrac{1}{f(u)-u}\,du = \ln\lvert x\rvert + c$
6. $y' = f\left(\dfrac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$	Set $x = X + \alpha, y = Y + \beta$
	Choose $\begin{cases} a_1\alpha + b_1\beta = -c_2 \\ a_2\alpha + b_2\beta = -c_2 \end{cases} \rightarrow Y' = f\left(\dfrac{a_1X + b_1Y}{a_2X + b_2Y}\right)$
	If $a_1b_2 - a_2b_1 \neq 0$, set $Y = Xu \rightarrow$ separable form
	$u + Xu' = f\left(\dfrac{a_1 + b_1u}{a_2 + b_2u}\right)$
	If $a_1b_2 - a_2b_1 = 0$, set $u = a_1x + b_1y \rightarrow$
	$\dfrac{du}{dx} = a_1 + b_1 f\left(\dfrac{u + c_1}{ku + c_2}\right)$ since
	$a_2x + b_2y = k(a,x + a_2y)$
7. $y'' + a^2y = 0$	$y = c_1 \cos ax + c_2 \sin ax$
8. $y'' - a^2y = 0$	$y = c_1e^{ax} + c_2e^{-ax}$
9. $y'' + ay' + by = 0$	Set $y = e^{-(a/2)x}u \rightarrow u'' + \left(b - \dfrac{a^2}{4}\right)u = 0$
10. $y'' + a(x)y' + b(x)y = 0$	Set $y = e^{-(1/2)\int a(x)\,dx} \rightarrow u'' + \left[b(x) - \dfrac{a^2}{4} - \dfrac{a'}{2}\right]u = 0$
11. $x^2y'' + xy' + (x^2 - a^2)\,y = 0$ $a \geq 0$ (Bessel)	i. If a is not an integer $y = c_1J_a(x) + c_2J_{-a}(x)$ (Bessel functions of first kind) ii. If a is an integer (say, n) $y = c_1J_n(x) + c_2Y_n(x)$ (Y_n is Bessel function of second kind)
12. $(1 - x^2)y'' - 2xy' + a(a + 1)y = 0$ a is real (Legendre)	$y(x) = c_1p_a(x) + c_1q_a(x)$ (Legendre functions)
13. $y' + ay^2 = bx^n$ (integrable Riccati) a, b, n real	Set $u' = ayu \rightarrow u'' - abx^nu = 0$ and use 14
14. $y'' - ax^{-1}y' + b^2x^{\mu}y = 0$	$y = x^p[c_1J_v(kx^q) + c_2J_{-v}(kx^q)$ where $p = (a + 1)/2$, $\nu = (a + 1)/(\mu + 2)$, $k = 2b/(\mu + 2)$, $q = (\mu + 2)/2$
15. Item 13 shows that the Riccati equation is linearized by raising the order of the equation. The *Riccati chain*, which is linearizable by raising the order, is $u' = uy,\quad u'' = u[y^1 + y^2],\quad u''' = u[y'' + 3yy' + y^3],$ $u^{(iv)} = u[y''' + 4yy'' + 6y^2y' + 3(y')^2 + y^4],\ldots$ To use this consider the second-order equation $y'' + 3yy' + y^3 = f(x)$. The Riccati transformation $u' = yu$ transforms this equation to the linear for $u''' = uf(x)$!	

References

Kanke, E. 1956. *Differentialgleichungen Lösungsmethoden und Lösungen*, Vol. I. Akad. Verlagsges., Leipzig.

Murphy, G. M. 1960. *Ordinary Differential Equations and Their Solutions*, Van Nostrand, New York.

Zwillger, D. 1992. *Handbook of Differential Equations*, 2nd ed. Academic Press, San Diego.

19.2 Linear Algebra and Matrices

George Cain

Basic Definitions

A *Matrix* **A** is a rectangular array of numbers (real or complex)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

The *size* of the matrix is said to be $n \times m$. The $1 \times m$ matrices $[a_{i1} \cdots a_{im}]$ are called rows of A, and the $n \times 1$ matrices

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}$$

are called *columns* of **A**. An $n \times m$ matrix thus consists of n rows and m columns; a_{ij} denotes the *element*, or *entry*, of **A** in the ith row and jth column. A matrix consisting of just one row is called a *row vector*, whereas a matrix of just one column is called a *column vector.* The elements of a vector are frequently called *components* of the vector. When the size of the matrix is clear from the context, we sometimes write $\mathbf{A} = (a_{ij})$.

A matrix with the same number of rows as columns is a *square* matrix, and the number of rows and columns is the *order* of the matrix. The diagonal of an $n \times n$ square matrix **A** from a_{11} to a_{nn} is called the *main*, or *principal*, *diagonal.* The word *diagonal* with no modifier usually means the main diagonal. The *transpose* of a matrix **A** is the matrix that results from interchanging the rows and columns of **A**. It is usually denoted by $\mathbf{A}^{\mathrm{T}}$. A matrix **A** such that $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ is said to be *symmetric.* The *conjugate transpose* of **A** is the matrix that results from replacing each element of $\mathbf{A}^{\mathrm{T}}$ by its complex conjugate, and is usually denoted by $\mathbf{A}^{\mathrm{H}}$. A matrix such that $\mathbf{A} = \mathbf{A}^{\mathrm{H}}$ is said to be *Hermitian*.

A square matrix $\mathbf{A} = (a_{ij})$ is *lower triangular* if $a_{ij} = 0$ for $j > i$ and is *upper triangular* if $a_{ij} = 0$ for $j < i$. A matrix that is both upper and lower triangular is a *diagonal* matrix. The $n \times n$ *identity matrix* is the $n \times n$ diagonal matrix in which each element of the main diagonal is 1. It is traditionally denoted $\mathbf{I}_n$, or simply **I** when the order is clear from the context.

Algebra of Matrices

The sum and difference of two matrices **A** and **B** are defined whenever **A** and **B** have the same size. In that case $\mathbf{C} = \mathbf{A} \pm \mathbf{B}$ is defined by $\mathbf{C} = (c_{ij}) = (a_{ij} \pm b_{ij})$. The product $t\mathbf{A}$ of a scalar t (real or complex number) and a matrix **A** is defined by $t\mathbf{A} = (ta_{ij})$. If **A** is an $n \times m$ matrix and **B** is an $m \times p$ matrix, the product $\mathbf{C} = \mathbf{AB}$ is defined to be the $n \times p$ matrix $\mathbf{C} = (c_{ij})$ given by $c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$. Note that the product of an $n \times m$ matrix and an $m \times p$ matrix is an $n \times p$ matrix, and the product is defined only when the number of columns of the first factor is the same as the number of rows of the second factor. Matrix multiplication is, in general, associative: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$. It also distributes over addition (and subtraction):

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \quad \text{and} \quad (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$

It is, however, not in general true that $\mathbf{AB} = \mathbf{BA}$, even in case both products are defined. It is clear that $(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}$ and $(\mathbf{A} + \mathbf{B})^{\mathrm{H}} = \mathbf{A}^{\mathrm{H}} + \mathbf{B}^{\mathrm{H}}$. It is also true, but not so obvious perhaps, that $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ and $(\mathbf{AB})^{\mathrm{H}} = \mathbf{B}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}$.

The $n \times n$ identity matrix **I** has the property that $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$ for every $n \times n$ matrix **A**. If **A** is square, and if there is a matrix **B** such at $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, then **B** is called the *inverse* of **A** and is denoted $\mathbf{A}^{-1}$. This terminology and notation are justified by the fact that a matrix can have at most one inverse. A matrix having an inverse is said to be *invertible,* or *nonsingular,* while a matrix not having an inverse is said to be *noninvertible,* or *singular.* The product of two invertible matrices is invertible and, in fact, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. The sum of two invertible matrices is, obviously, not necessarily invertible.

Systems of Equations

The system of n linear equations in m unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2m}x_m &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nm}x_m &= b_n \end{aligned}$$

may be written $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} = (a_{ij})$, $\mathrm{x} = [\mathrm{x}_1\ \mathrm{x}_2\ \cdots\ \mathrm{x}_m]^{\mathrm{T}}$, and $\mathrm{b} = [\mathrm{b}_1\ \mathrm{b}_2\ \cdots\ \mathrm{b}_n]^{\mathrm{T}}$. Thus A is an $n \times m$ matrix, and x and b are column vectors of the appropriate sizes.

The matrix **A** is called the *coefficient matrix* of the system. Let us first suppose the coefficient matrix is square; that is, there are an equal number of equations and unknowns. If **A** is upper triangular, it is quite easy to find all solutions of the system. The ith equation will contain only the unknowns x_i, x_{i+1}, …, x_n, and one simply solves the equations in reverse order: the last equation is solved for x_n; the result is substituted into the $(n - 1)$st equation, which is then solved for x_{n-1}; these values of x_n and x_{n-1} are substituted in the $(n - 2)$th equation, which is solved for x_{n-2}, and so on. This procedure is known as *back substitution.*

The strategy for solving an arbitrary system is to find an upper-triangular system equivalent with it and solve this upper-triangular system using back substitution. First suppose the element $a_{11} \neq 0$. We may rearrange the equations to ensure this, unless, of course the first column of **A** is all 0s. In this case proceed to the next step, to be described later. For each $i \geq 2$ let $m_{i1} = a_{i1}/a_{11}$. Now replace the ith equation by the result of multiplying the first equation by m_{i1} and subtracting the new equation from the ith equation. Thus,

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \ldots + a_{im}x_m = b_i$$

is replaced by

$$0 \cdot x_1 + \left(a_{i2} + m_{i1}a_{12}\right)x_2 + \left(a_{i3} + m_{i1}a_{13}\right)x_3 + \ldots + \left(a_{im} + m_{i1}a_{1m}\right)x_m = b_i + m_{i1}b_1$$

After this is done for all $i = 2, 3, \ldots, n$, there results the equivalent system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1 \\
0 \cdot x_1 + a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2 \\
0 \cdot x_1 + a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
0 \cdot x_1 + a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nn}x_n &= b'_n
\end{aligned}$$

in which all entries in the first column below a_{11} are 0. (Note that if all entries in the first column were 0 to begin with, then $a_{11} = 0$ also.) This procedure is now repeated for the $(n-1) \times (n-1)$ system

$$\begin{aligned}
a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2 \\
a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nn}x_n &= b'_n
\end{aligned}$$

to obtain an equivalent system in which all entries of the coefficient matrix below a'_{22} are 0. Continuing, we obtain an upper-triangular system Ux = c equivalent with the original system. This procedure is known as *Gaussian elimination*. The number m_{ij} are known as the *multipliers*.

Essentially the same procedure may be used in case the coefficient matrix is not square. If the coefficient matrix is not square, we may make it square by appending either rows or columns of 0s as needed. Appending rows of 0s and appending 0s to make b have the appropriate size equivalent to appending equations 0 = 0 to the system. Clearly the new system has precisely the same solutions as the original system. Appending columns of 0s and adjusting the size of x appropriately yields a new system with additional unknowns, each appearing only with coefficient 0, thus not affecting the solutions of the original system. In either case we may assume the coefficient matrix is square, and apply the Gauss elimination procedure.

Suppose the matrix **A** is invertible. Then if there were no row interchanges in carrying out the above Gauss elimination procedure, we have the *LU factorization* of the matrix **A**:

$$\mathbf{A} = \mathbf{LU}$$

where U is the upper-triangular matrix produced by elimination and L is the lower-triangular matrix given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ m_{21} & 1 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \\ m_{n1} & m_{n2} & \ldots & & 1 \end{bmatrix}$$

A *permutation* P_{ij} matrix is an $n \times n$ matrix such that P_{ij} A is the matrix that results from exchanging row i and j of the matrix A. The matrix P_{ij} is the matrix that results from exchanging rows i and j of the identity matrix. A product P of such matrices P_{ij} is called a *permutation* matrix. If row interchanges are required in the Gauss elimination procedure, then we have the factorization

$$\mathbf{PA} = \mathbf{LU}$$

where P is the permutation matrix giving the required row exchanges.

Vector Spaces

The collection of all column vectors with n real components is *Euclidean n-space*, and is denoted R^n. The collection of column vectors with n complex components is denoted C^n. We shall use *vector space* to mean either R^n or C^n. In discussing the space R^n, the word *scalar* will mean a real number, and in discussing the space C^n, it will mean a complex number. A subset S of a vector space is a *subspace* such that if **u** and **v** are vectors in S, and if c is any scalar, then **u** + **v** and c**u** are in S. We shall sometimes use the word *space* to mean a subspace. If $B = \{v_1, v_2, \ldots, v_k\}$ is a collection of vectors in a vector space, then the set S consisting of all vectors $c_1v_1 + c_2v_2 + \cdots + c_mv_m$ for all scalars $c_1, c_2, \ldots, c_m$ is a subspace, called the *span* of *B*. A collection $\{v_1, v_2, \ldots, v_m\}$ of vectors $c_1v_1 + c_2v_2 + \cdots + c_mv_m$ is a *linear combination* of *B*. If S is a subspace and $B = \{v_1, v_2, \ldots, v_m\}$ is a subset of S such that S is the span of *B*, then *B* is said to *span* S.

A collection $\{v_1, v_2, \ldots, v_m\}$ of n-vectors is *linearly dependent* if there exist scalars $c_1, c_2, \ldots, c_m$, not all zero, such that $c_1v_1 + c_2v_2 + \cdots + c_mv_m = 0$. A collection of vectors that is not linearly dependent is said to be *linearly independent*. The modifier *linearly* is frequently omitted, and we speak simply of dependent and independent collections. A linearly independent collection of vectors in a space S that spans S is a *basis* of S. Every basis of a space S contains the same number of vectors; this number is the *dimension* of S. The dimension of the space consisting of only the zero vector is 0. The collection $B = \{e_1, e_2, \ldots, e_n\}$, where $e_1 = [1,0,0,\ldots, 0]^T$, $e_2 = [0,1,0,\ldots, 0]^T$, and so forth ($e_i$ has 1 as its ith component and zero for all other components) is a basis for the spaces R^n and C^n. This is the *standard basis* for these spaces. The dimension of these spaces is thus n. In a space S of dimension n, no collection of fewer than n vectors can span S, and no collection of more than n vectors in S can be independent.

Rank and Nullity

The *column space* of an $n \times m$ matrix **A** is the subspace of R^n or C^n spanned by the columns of **A**. The *row space* is the subspace of R^m or C^m spanned by the rows or **A**. Note that for any vector $x = [x_1\, x_2, \cdots x_m]^T$,

$$\mathbf{Ax} = x_1\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + x_2\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \ldots + x_m\begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix}$$

so that the column space is the collection of all vectors, **Ax**, and thus the system **Ax** = **b** has a solution if and only *if* **b** is a member of the column space of **A**.

The dimension of the column space is the *rank* of **A**. The row space has the same dimension as the column space. The set of all solutions of the system **Ax** = 0 is a subspace called the *null space* of **A**, and the dimension of this null space is the *nullity* of **A**. A fundamental result in matrix theory is the fact that, for an $n \times m$ matrix **A**.

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = m$$

The difference of any two solutions of the linear system $\mathbf{Ax} = \mathbf{b}$ is a member of the null space of **A**. Thus this system has at most one solution if and only if the nullity of **A** is zero. If the system is square (that is, if **A** is $n \times n$), then there will be a solution for every right-hand side **b** if and only if the collection of columns of **A** is linearly independent, which is the same as saying the rank of **A** is n. In this case the nullity must be zero. Thus, for any **b**, the square system $\mathbf{Ax} = \mathbf{b}$ has exactly one solution if and only if rank $\mathbf{A} = n$. In other words the $n \times n$ matrix **A** is invertible if and only if rank $\mathbf{A} = n$.

Orthogonality and Length

The *inner product* of two vectors x and y is the scalar $x^H y$. The *length*, or *norm*, $\|x\|$, of the vector **x** is given by $\|x\| = \sqrt{\mathbf{x}^H \mathbf{x}}$. A *unit vector* is a vector of norm 1. Two vectors x and y are *orthogonal* if $x^H y = 0$. A collection of vectors $\{v_1, v_2, \ldots, v_m\}$ in a space S is said to be an *orthonormal* collection if $\mathbf{v}_i^H \mathbf{v}_j = 0$ for $i \neq j$ and $\mathbf{v}_i^H \mathbf{v}_i = 1$. An orthonormal collection is necessarily linearly independent. If S is a subspace (of R^n or C^n) spanned by the orthonormal collection $\{v_1, v_2, \ldots, v_m\}$, then the *projection* of a vector x onto S is the vector

$$\text{proj}(\mathbf{x}; \mathbf{S}) = (\mathbf{x}^H \mathbf{v}_1)\mathbf{v}_1 + (\mathbf{x}^H \mathbf{v}_2)\mathbf{v}_2 + \ldots + (\mathbf{x}^H \mathbf{v}_m)\mathbf{v}_m$$

The projection of x onto S minimizes the function $f(y) = \|x - y\|^2$ for $y \in S$. In other words the projection of x onto S is the vector in S that is "closest" to x.

If **b** is a vector and **A** is an $n \times m$ matrix, then a vector x minimizes $\|b - Ax\|^2$ if only if it is a solution of $A^H Ax = A^H b$. This system of equations is called the *system of normal equations* for the least-squares problem of minimizing $\|b - Ax\|^2$.

If **A** is an $n \times m$ matrix, and rank $\mathbf{A} = k$, then there is a $n \times k$ matrix Q whose columns form an orthonormal basis for the column space of **A** and a $k \times m$ upper-triangular matrix **R** of rank k such that

$$\mathbf{A} = \mathbf{QR}$$

This is called the *QR factorization* of **A**. It now follows that x minimizes $\|b - Ax\|^2$ if and only if it is a solution of the upper-triangular system $Rx = Q^H b$.

If $\{w_1, w_2, \ldots, w_m\}$ is a basis for a space S, the following procedure produces an orthonormal basis $\{v_1, v_2, \ldots, v_m\}$ for S.

Set $\mathbf{v}_1 = \mathbf{w}_1 / \|\mathbf{w}_1\|$.
Let $\tilde{\mathbf{v}}_2 = \mathbf{w}_2 - \text{proj}(\mathbf{w}_2; \mathbf{S}_1)$, where $\mathbf{S}_1$ is the span of $\{\mathbf{v}_1\}$; set $\mathbf{v}_2 = \tilde{\mathbf{v}}_2 / \|\tilde{\mathbf{v}}_2\|$.
Next, let $\tilde{\mathbf{v}}_3 = \mathbf{w}_3 - \text{proj}(\mathbf{w}_3; \mathbf{S}_2)$, where $\mathbf{S}_2$ is the span of $\{\mathbf{v}_1, v_2\}$; set $\mathbf{v}_3 = \tilde{\mathbf{v}}_3 / \|\tilde{\mathbf{v}}_3\|$.

And, so on: $\tilde{\mathbf{v}}_i = w_i - \text{proj}(w_i; S_{i-1})$, where S_{i-1} is the span of $\{v_1, v_2, \ldots, v_{i-1}\}$; set $v_i = \tilde{\mathbf{v}}_i / \|\tilde{\mathbf{v}}_i\|$. This the *Gram-Schmidt procedure.*

If the collection of columns of a square matrix is an orthonormal collection, the matrix is called a *unitary matrix.* In case the matrix is a real matrix, it is usually called an *orthogonal matrix.* A unitary matrix U is invertible, and $\mathbf{U}^{-1} = \mathbf{U}^H$. (In the real case an orthogonal matrix **Q** is invertible, and $\mathbf{Q}^{-1} = \mathbf{Q}^T$.)

Determinants

The *determinant* of a square matrix is defined inductively. First, suppose the determinant $\det \mathbf{A}$ has been defined for all square matrices of order $< n$. Then

$$\det \mathbf{A} = a_{11}\mathbf{C}_{11} + a_{12}\mathbf{C}_{12} + \ldots + a_{1n}\mathbf{C}_{1n}$$

where the numbers C_{ij} are *cofactors* of the matrix **A**:

$$\mathbf{C}_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij}$$

where M_{ij} is the $(n-1) \times (n-1)$ matrix obtained by deleting the ith row and jth column of **A**. Now det**A** is defined to be the only entry of a matrix of order 1. Thus, for a matrix of order 2, we have

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

There are many interesting but not obvious properties of determinants. It is true that

$$\det \mathbf{A} = a_{i1}\mathbf{C}_{i1} + a_{i2}\mathbf{C}_{i2} + \ldots + a_{in}\mathbf{C}_{in}$$

for any $1 \le i \le n$. It is also true that det**A** = det$\mathbf{A}^{\mathrm{T}}$, so that we have

$$\det \mathbf{A} = a_{1j}\mathbf{C}_{1j} + a_{2j}\mathbf{C}_{2j} + \ldots + a_{nj}\mathbf{C}_{nj}$$

for any $1 \le j \le n$.

If **A** and **B** are matrices of the same order, then det**AB** = (det**A**)(det**B**), and the determinant of any identity matrix is 1. Perhaps the most important property of the determinant is the fact that a matrix in invertible if and only if its determinant is not zero.

Eigenvalues and Eigenvectors

If **A** is a square matrix, and **Av** = λv for a scalar λ and a nonzero v, then λ is an *eigenvalue* of **A** and v is an *eigenvector* of **A** that *corresponds* to λ. Any nonzero linear combination of eigenvectors corresponding to the same eigenvalue λ is also an eigenvector corresponding to λ. The collection of all eigenvectors corresponding to a given eigenvalue λ is thus a subspace, called an *eigenspace* of **A**. A collection of eigenvectors corresponding to different eigenvalues is necessarily linear-independent. It follows that a matrix of order n can have at most n distinct eigenvectors. In fact, the eigenvalues of **A** are the roots of the nth degree polynomial equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

called the *characteristic equation* of **A**. (Eigenvalues and eigenvectors are frequently called *characteristic values* and *characteristic vectors.*)

If the nth order matrix **A** has an independent collection of n eigenvectors, then **A** is said to have a *full set* of eigenvectors. In this case there is a set of eigenvectors of **A** that is a basis for R^n or, in the complex case, C^n. In case there are n distinct eigenvalues of **A**, then, of course, **A** has a full set of eigenvectors. If there are fewer than n distinct eigenvalues, then **A** may or may not have a full set of eigenvectors. If there is a full set of eigenvectors, then

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \quad \text{or} \quad \mathbf{A} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1}$$

where **D** is a diagonal matrix with the eigenvalues of **A** on the diagonal, and **S** is a matrix whose columns are the full set of eigenvectors. If **A** is symmetric, there are n real distinct eigenvalues of **A** and the corresponding eigenvectors are orthogonal. There is thus an orthonormal collection of eigenvectors that span R^n, and we have

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}} \quad \text{and} \quad \mathbf{D} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$$

where **Q** is a real orthogonal matrix and **D** is diagonal. For the complex case, if **A** is Hermitian, we have

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{H} \quad \text{and} \quad \mathbf{D} = \mathbf{U}^{H}\mathbf{A}\mathbf{U}$$

where **U** is a unitary matrix and **D** is a *real* diagonal matrix. (A Hermitian matrix also has *n* distinct real eigenvalues.)

References

Daniel, J. W. and Nobel, B. 1988. *Applied Linear Algebra.* Prentice Hall, Englewood Cliffs, NJ.
Strang, G. 1993. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA.

19.3 Vector Algebra and Calculus

George Cain

Basic Definitions

A vector is a directed line segment, with two vectors being equal if they have the same length and the same direction. More precisely, a *vector* is an equivalence class of directed line segments, where two directed segments are equivalent if they have the same length and the same direction. The *length* of a vector is the common length of its directed segments, and the *angle between* vectors is the angle between any of their segments. The length of a vector u is denoted $|\mathbf{u}|$. There is defined a distinguished vector having zero length, which is usually denoted **0**. It is frequently useful to visualize a directed segment as an arrow; we then speak of the nose and the tail of the segment. The *sum* $\mathbf{u} + \mathbf{v}$ of two vectors **u** and **v** is defined by taking directed segments from **u** and **v** and placing the tail of the segment representing **v** at the nose of the segment representing **u** and defining $\mathbf{u} + \mathbf{v}$ to be the vector determined by the segment from the tail of the **u** representative to the nose of the **v** representative. It is easy to see that $\mathbf{u} + \mathbf{v}$ is well defined and that $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$. Subtraction is the inverse operation of addition. Thus the *difference* $\mathbf{u} - \mathbf{v}$ of two vectors is defined to be the vector that when added to **v** gives **u**. In other words, if we take a segment from **u** and a segment from **v** and place their tails together, the difference is the segment from the nose of v to the nose of **u**. The zero vector behaves as one might expect; $\mathbf{u} + \mathbf{0} = \mathbf{u}$, and $\mathbf{u} - \mathbf{u} = \mathbf{0}$. Addition is associative: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.

To distinguish them from vectors, the real numbers are called *scalars*. The product $t\mathbf{u}$ of a scalar t and a vector **u** is defined to be the vector having length $|t|\,|\mathbf{u}|$ and direction the same as **u** if $t > 0$, the opposite direction if $t < 0$, If $t = 0$, then $t\mathbf{u}$ is defined to be the zero vector. Note that $t(\mathbf{u} + \mathbf{v}) = t\mathbf{u} + t\mathbf{v}$, and $(t + s)\mathbf{u} = t\mathbf{u} + s\mathbf{u}$. From this it follows that $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-1)\mathbf{v}$.

The *scalar product* $\mathbf{u} \cdot \mathbf{v}$ of two vectors is $|\mathbf{u}||\mathbf{v}| \cos \theta$, where θ is the angle between **u** and **v**. The scalar product is frequently called the *dot product.* The scalar product distributes over addition:

$$\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$$

and it is clear that $(t\mathbf{u}) \cdot \mathbf{v} = t(\mathbf{u} \cdot \mathbf{v})$. The *vector product* $\mathbf{u} \times \mathbf{v}$ of two vectors is defined to be the vector perpendicular to both **u** and **v** and having length $|\mathbf{u}||\mathbf{v}| \sin \theta$, where θ is the angle between **u** and **v**. The direction of $\mathbf{u} \times \mathbf{v}$ is the direction a right-hand threaded bolt advances if the vector **u** is rotated to **v**. The vector is frequently called the *cross product.* The vector product is both associative and distributive, but not commutative: $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$.

Coordinate Systems

Suppose we have a right-handed Cartesian coordinate system in space. For each vector, **u**, we associate a point in space by placing the tail of a representative of **u** at the origin and associating with **u** the point at the nose of the segment. Conversely, associated with each point in space is the vector determined by the directed segment from the origin to that point. There is thus a one-to-one correspondence between the points in space and all vectors. The origin corresponds to the zero vector. The coordinates of the point associated with a vector **u** are called *coordinates* of **u**. One frequently refers to the vector **u** and writes **u** = (*x, y, z*), which is, strictly speaking, incorrect, because the left side of this equation is a vector and the right side gives the coordinates of a point in space. What is meant is that (*x, y, z*) are the coordinates of the point associated with **u** under the correspondence described. In terms of coordinates, for $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, we have

$$\mathbf{u} + \mathbf{v} = \left(u_1 + v_1, u_2 + v_2, u_3 + v_3\right)$$

$$t\mathbf{u} = \left(tu_1, tu_2, tu_3\right)$$

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3$$

$$\mathbf{u} \times \mathbf{v} = \left(u_2v_3 - v_2u_3, u_3v_1 - v_3u_1, u_1v_2 - v_1u_2\right)$$

The *coordinate vectors* **i**, **j**, and **k** are the unit vectors i = (1, 0, 0), j = (0, 1, 0), and k = (0, 0, 1). Any vector $\mathbf{u} = (u_1, u_2, u_3)$ is thus a linear combination of these coordinate vectors: $u = u_1i + u_2j + u_3k$. A convenient form for the vector product is the formal determinant

$$\mathbf{u} \times \mathbf{v} = \det\begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_2 \end{bmatrix}$$

Vector Functions

A *vector function* **F** *of one variable* is a rule that associates a vector **F**(*t*) with each real number *t* in some set, called the *domain* of **F**. The expression $\lim_{t \to t_0} \mathbf{F}(t) =$ a means that for any $\varepsilon > 0$, there is a $\delta > 0$ such that $|\mathbf{F}(t) - \mathbf{a}| < \varepsilon$ whenever $0 < |t - t_0| < \delta$. If $\mathbf{F}(t) = [x(t), y(t), z(t)]$ and $\mathbf{a} = (a_1, a_2, a_3)$, then $\lim_{t \to t_0} \mathbf{F}(t) = \mathbf{a}$ if and only if

$$\lim_{t \to t_0} x(t) = a_1$$

$$\lim_{t \to t_0} y(t) = a_2$$

$$\lim_{t \to t_0} z(t) = a_3$$

A vector function **F** is *continuous* at t_0 if $\lim_{t \to t_0} \mathbf{F}(t) = \mathbf{F}(t_0)$. The vector function **F** is continuous at t_0 if and only if each of the coordinates $x(t)$, $y(t)$, and $z(t)$ is continuous at t_0.

The function **F** is *differentiable* at t_0 if the limit

$$\lim_{h \to 0} \frac{1}{h}\left[\mathbf{F}(t + h) - \mathbf{F}(t)\right]$$

exists. This limit is called the *derivative* of **F** at t_0 and is usually written $\mathbf{F}'(t_0)$, or $(d\mathbf{F}/dt)(t_0)$. The vector function **F** is differentiable at t_0 if and only if each of its coordinate functions is differentiable at t_0. Moreover, $(d\mathbf{F}/dt)(t_0) = [(dx/dt)(t_0), (dy/dt)(t_0), (dz/dt)(t_0)]$. The usual rules for derivatives of real valued functions all hold for vector functions. Thus if **F** and **G** are vector functions and *s* is a scalar function, then

$$\frac{d}{dt}(\mathbf{F}+\mathbf{G}) = \frac{d\mathbf{F}}{dt} + \frac{d\mathbf{G}}{dt}$$

$$\frac{d}{dt}(s\mathbf{F}) = s\frac{d\mathbf{F}}{dt} + \frac{ds}{dt}\mathbf{F}$$

$$\frac{d}{dt}(\mathbf{F}\cdot\mathbf{G}) = \mathbf{F}\cdot\frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt}\cdot\mathbf{G}$$

$$\frac{d}{dt}(\mathbf{F}\times\mathbf{G}) = \mathbf{F}\times\frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt}\times\mathbf{G}$$

If **R** is a vector function defined for *t* in some interval, then, as *t* varies, with the tail of **R** at the origin, the nose traces out some object *C* in space. For nice functions **R**, the object *C* is a *curve*. If $\mathbf{R}(t) = [x(t), y(t), z(t)]$, then the equations

$$x = x(t)$$

$$y = y(t)$$

$$z = z(t)$$

are called *parametric equations* of *C*. At points where **R** is differentiable, the derivative $d\mathbf{R}/dt$ is a vector *tangent* to the curve. The unit vector $\mathbf{T} = (d\mathbf{R}/dt)/|d\mathbf{R}/dt|$ is called the *unit tangent vector*. If **R** is differentiable and if the length of the arc of curve described by **R** between $\mathbf{R}(a)$ and $\mathbf{R}(t)$ is given by $s(t)$, then

$$\frac{ds}{dt} = \left|\frac{d\mathbf{R}}{dt}\right|$$

Thus the length *L* of the arc from $\mathbf{R}(t_0)$ to $\mathbf{R}(t_1)$ is

$$L = \int_{t_0}^{t_1}\frac{ds}{dt}dt = \int_{t_0}^{t_1}\left|\frac{d\mathbf{R}}{dt}\right|dt$$

The vector $d\mathbf{T}/ds = (d\mathbf{T}/dt)/(ds/dt)$ is perpendicular to the unit tangent **T**, and the number $\kappa = |d\mathbf{T}/ds|$ is the *curvature* of *C*. The unit vector $\mathbf{N} = (1/\kappa)(d\mathbf{T}/ds)$ is the *principal normal*. The vector $\mathbf{B} = \mathbf{T}\times\mathbf{N}$ is the *binormal*, and $d\mathbf{B}/ds = -\tau N$. The number τ is the *torsion*. Note that *C* is a plane curve if and only if τ is zero for all *t*.

A *vector function* **F** *of two variables* is a rule that assigns a vector $\mathbf{F}(s, t)$ in some subset of the plane, called the *domain* of **F**. If $\mathbf{R}(s, t)$ is defined for all (s, t) in some region *D* of the plane, then as the point (s, t) varies over *D*, with its rail at the origin, the nose of $\mathbf{R}(s, t)$ traces out an object in space. For a nice function **R**, this object is a *surface*, *S*. The partial derivatives $(\partial\mathbf{R}/\partial s)(s, t)$ and $(\partial\mathbf{R}/\partial t)(s, t)$ are tangent to the surface at $\mathbf{R}(s, t)$, and the vector $(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$ is thus *normal* to the surface. Of course, $(\partial\mathbf{R}/\partial t)\times(\partial\mathbf{R}/\partial s) = -(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$ is also normal to the surface and points in the direction opposite that of $(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$. By electing one of these normal, we are choosing an *orientation* of the surface.

A surface can be oriented only if it has two sides, and the process of orientation consists of choosing which side is "positive" and which is "negative."

Gradient, Curl, and Divergence

If $f(x, y, z)$ is a scalar field defined in some region D, the *gradient* of **f** is the vector function

$$\text{grad } f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

If $\mathbf{F}(x, y, z) = F_1(x, y, z)\mathbf{i} + F_2(x, y, z)\mathbf{j} + F_3(x, y, z)\mathbf{k}$ is a vector field defined in some region D, then the *divergence* of **F** is the scalar function

$$\text{div } \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

The curl is the vector function

$$\text{curl } \mathbf{F} = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{k}$$

In terms of the vector operator *del*, $\nabla = \mathbf{i}(\partial/\partial x) + \mathbf{j}(\partial/\partial y) + \mathbf{k}(\partial/\partial z)$, we can write

$$\text{grad } f = \nabla f$$

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}$$

$$\text{curl } \mathbf{F} = \nabla \times \mathbf{F}$$

The *Laplacian operator* is div (grad) $= \nabla \cdot \nabla = \nabla^2 = (\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$.

Integration

Suppose C is a curve from the point (x_0, y_0, z_0) to the point (x_1, y_1, z_1) and is described by the vector function $\mathbf{R}(t)$ for $t_0 \le t \le t_1$. If f f is a scalar function (sometimes called a *scalar field*) defined on C, then the integral of f over C is

$$\int_C f(x,y,z)\, ds = \int_{t_0}^{t_1} f\left[\mathbf{R}(t)\right]\left|\frac{d\mathbf{R}}{dt}\right| dt$$

If **F** is a vector function (sometimes called a *vector field*) defined on C, then the integral of **F** over C is

$$\int_C \mathbf{F}(x,y,z)\cdot d\mathbf{R} = \int_{t_0}^{t_1} \mathbf{F}\left[\mathbf{R}(t)\right]\frac{d\mathbf{R}}{dt}\, dt$$

These integrals are called *line integrals*.

In case there is a scalar function f such that $\mathbf{F} = \text{grad } f$, then the line integral

$$\int_C \mathbf{F}(x,y,z)\cdot d\mathbf{R} = f\left[\mathbf{R}(t_1)\right] - f\left[\mathbf{R}(t_0)\right]$$

The value of the integral thus depends only on the end points of the curve C and not on the curve C itself. The integral is said to be *path-independent.* The function f is called a *potential function* for the vector field **F**, and **F** is said to be a *conservative field.* A vector field **F** with domain D is conservative if and only if the integral of **F** around every closed curve in D is zero. If the domain D is simply connected (that is, every closed curve in D can be continuously deformed in D to a point), then **F** is conservative if and only if curl **F** = 0 in D.

Suppose S is a surface described by $\mathbf{R}(s, t)$ for (s, t) in a region D of the plane. If f is a scalar function defined on D, then the integral of f over S is given by

$$\iint_S f(x,y,z)\, dS = \iint_D f\big[\mathbf{R}(s,t)\big] \left| \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right| ds\, dt$$

If **F** is a vector function defined on S, and if an orientation for S is chosen, then the integral **F** over S, sometimes called the flux of **F** through S, is

$$\iint_S \mathbf{F}(x,y,z) \cdot d\mathbf{S} = \iint_D \mathbf{F}\big[\mathbf{R}(s,t)\big] \left| \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right| ds\, dt$$

Integral Thorems

Suppose **F** is a vector field with a closed domain D bounded by the surface S oriented so that the normal points out from D. Then the *divergence theorem* states that

$$\iiint_D \text{div } \mathbf{F}\, dV = \iint_S \mathbf{F} \cdot \mathbf{dS}$$

If S is an orientable surface bounded by a closed curve C, the orientation of the closed curve C is chosen to be consistent with the orientation of the surface S. Then we have *Stoke's theorem:*

$$\iint_S (\text{curl } \mathbf{F}) \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot \mathbf{ds}$$

References

Davis, H. F. and Snider, A. D. 1991. *Introduction to Vector Analysis,* 6th ed., Wm. C. Brown, Dubuque, IA.
Wylie, C. R. 1975. *Advanced Engineering Mathematics,* 4th ed., McGraw-Hill, New York.

Further Information

More advanced topics leading into the theory and applications of tensors may be found in J. G. Simmonds, *A Brief on Tensor Analysis* (1982, Springer-Verlag, New York).

19.4 Difference Equations

William F. Ames

Difference equations are equations involving *discrete variables*. They appear as natural descriptions of natural phenomena and in the study of discretization methods for differential equations, which have continuous variables.

Let $y_n = y(nh)$, where n is an integer and h is a real number. (One can think of measurements taken at equal intervals, h, $2h$, $3h$, …, and y_n describes these). A typical equation is that describing the famous Fibonacci sequence — $y_{n+2} - y_{n+1} - y_n = 0$. Another example is the equation $y_{n+2} - 2zy_{n+1} + y_n = 0$, $z \in C$, which describes the Chebyshev polynomials.

First-Order Equations

The general first-order equation $y_{n+1} = f(y_n)$, $y_0 = y(0)$ is easily solved, for as many terms as are needed, by *iteration*. Then $y_1 = f(y_0)$; $y_2 = f(y_1)$,…. An example is the logistic equation $y_{n+1} = ay_n(1 - y_n) = f(y_n)$. The logistic equation has two fixed (critical or equilibrium) points where $y_{n+1} = y_n$. They are 0 and $\bar{y} = (a - 1)/a$. This has physical meaning only for $a > 1$. For $1 < a < 3$ the equilibrium $\bar{y}$ is asymptotically stable, and for $a > 3$ there are two points y_1 and y_2, called a *cycle of period two*, in which $y_2 = f(y_1)$ and $y_1 = f(y_2)$. This study leads into chaos, which is outside our interest. By iteration, with $y_0 = {}^1/_2$, we have $y_1 = (a/2)(1/2) = a/2^2$, $y_2 = a(a/2^2)(1 - a/2^2) = (a^2/2^2)(1 - a/2^2)$, ….

With a constant, the equation $y_{n+1} = ay_n$ is solved by making the assumption $y_n = A\lambda^n$ and finding λ so that the equation holds. Thus $A\lambda^{n+1} = aA\lambda^n$, and hence $\lambda = 0$ or $\lambda = a$ and A is arbitrary. Discarding the trivial solution 0 we find $y_n = Aa^{n+1}$ is the desired solution. By using a method called the *variation of constants*, the equation $y_{n+1} - ay_n = g_n$ has the solution $y_n = y_0a^n + \Sigma_{j=0}^{n-1} g_j a^{n-j-1}$, with y_0 arbitrary.

In various applications we find the first-order equation of *Riccati type* $y_ny_{n-1} + ay_n + by_{n-1} + c = 0$ where a, b, and c are real constants. This equation can be transformed to a linear second-order equation by setting $y_n = z_n/z_{n-1} - a$ to obtain $z_{n+1} + (b + a)z_n + (c - ab)z_{n-1} = 0$, which is solvable as described in the next section.

Second-Order Equations

The second-order linear equation with constant coefficients $y_{n+2} + ay_{n+1} + by_n = f_n$ is solved by first solving the homogeneous equation (with right-hand side zero) and adding to that solution any solution of the inhomogeneous equation. The *homogeneous equation* $y_{n+2} + ay_{n+1}$ $by_n = 0$ is solved by assuming $y_n = \lambda^n$, whereupon $\lambda^{n+2} + a\lambda^{n+1} + b\lambda^n = 0$ or $\lambda = 0$ (rejected) or $\lambda^2 + a\lambda + b = 0$. The roots of this quadratic are $\lambda_1 = {}^1/_2\,(-a + \sqrt{a^2 - 4b})$, $\lambda_2 = -{}^1/_2\,(a + \sqrt{a^2 - 4b})$ and the solution of the homogeneous equation is $y_n = c_1\lambda_1^n + c_2\lambda_2^n$. As an example consider the Fibonacci equation $y_{n+2} - y_{n+1} - y_n = 0$. The roots of $\lambda^2 - \lambda - 1 = 0$ are $\lambda_1 = {}^1/_2\,(1 + \sqrt{5})$, $\lambda_2 = {}^1/_2\,(1 - \sqrt{5})$, and the solution $y_n = c_1[(1 + \sqrt{5})/2]^n + c_2[(1 - \sqrt{5})/2]^n$ is known as the *Fibonacci sequence.*

Many of the orthogonal polynomials of differential equations and numerical analysis satisfy a second-order difference equation (recurrence relation) involving a discrete variable, say n, and a continuous variable, say z. One such is the *Chebshev equation* $y_{n+2} - 2zy_{n+1} + y_n = 0$ with the initial conditions $y_0 = 1$, $y_1 = z$ (*first-kind* Chebyshev polynomials) and $y_{n-1} = 0$, $y_0 = 1$ (second-kind Chebyshev polynomials). They are denoted $T_n(z)$ and $V_n(z)$, respectively. By iteration we find

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_2(z) = 2z^2 - 1,$$

$$T_3(z) = 4z^3 - 3z, \quad T_4(z) = 8z^4 - 8z^2 + 1$$

$$V_0(z) = 0, \quad V_1(z) = 1, \quad V_2(z) = 2z,$$

$$V_3(z) = 4z^2 - 1, \quad V_4(z) = 8z^3 - 4z$$

and the general solution is $y_n(z) = c_1T_n(z) + c_2V_{n-1}(z)$,

Linear Equations with Constant Coefficients

The genral kth-order linear equation with constant coefficients is $\Sigma_{i=0}^{k} p_i y_{n+k-i} = g_n$, $p_0 = 1$. The solution to the corresponding homogeneous equation (obtained by setting $g_n = 0$) is as follows. (a) $y_n = \Sigma_{i=1}^{k} c_i \lambda_i^n$ if the λ_i are the distinct roots of the characteristic polynomial $p(\lambda) = \Sigma_{i=0}^{k} p_i \lambda^{k-i} = 0$. (b) if m_s is the multiplicity of the root λ_s, then the functions $y_{n,s} = u_s(n)\lambda_s^n$, where $u_s(n)$ are polynomials in n whose degree does not exceed $m_s - 1$, are solutions of the equation. Then the general solution of the homogeneous equation is $y_n = \Sigma_{i=1}^{d} a_i u_i(n)\lambda_i^n = \Sigma_{i=1}^{d} a_i \Sigma_{j=0}^{m_i-1} c_j n^j \lambda_i^n$. To this solution one adds any particular solution to obtain the general solution of the general equation.

Example 19.4.1. A model equation for the price p_n of a product, at the nth time, is $p_n + b/a(1 + \rho)p_{n-1} - (b/a)\rho p_{n-2} + (s_0 - d_0)/a = 0$. The equilibrium price is obtained by setting $p_n = p_{n-1} = p_{n-2} = p_e$, and one finds $p_e = (d_0 - s_0)/(a + b)$. The homogeneous equation has the characteristic polynomial $\lambda^2 + (b/a)(1 + \rho)\lambda - (b/a)\rho = 0$. With λ_1 and λ_2 as the roots the general solution of the full equation is $p_n = c_1\lambda_1^n + c_2\lambda_2^n + p_e$, since p_e is a solution of the full equation. This is one method for finding the solution of the nonhomogeneous equation.

Generating Function (z Transform)

An elegant way of solving linear difference equations with constant coefficients, among other applications, is by use of *generating functions* or, as an alternative, the z transform. The generating function of a sequence $\{y_n\}$, $n = 0, 1, 2, \ldots$, is the function $f(x)$ given by the formal series $f(x) = \Sigma_{n=0}^{\infty} y_n x^n$. The z transform of the same sequence is $z(x) = \Sigma_{n=0}^{\infty} y_n x^{-n}$. Clearly, $z(x) = f(1/x)$. A table of some important sequences is given in Table 19.4.1.

Table 19.4.1 Important Sequences

y_n	$f(x)$	Convergence Domain
1	$(1 - x)^{-1}$	$\|x\| < 1$
n	$x(1 - x)^{-2}$	$\|x\| < 1$
n^m	$xp_m(x)(1 - x)^{-n-1}$*	$\|x\| < 1$
k^n	$(1 - kx)^{-1}$	$\|x\| < k^{-1}$
e^{an}	$(1 - e^a x)^{-1}$	$\|x\| < e^{-a}$
$k^n \cos an$	$\dfrac{1 - kx\cos a}{1 - 2kx\cos a + k^2x^2}$	$\|x\| < k^{-1}$
$k^n \sin an$	$\dfrac{kx\sin a}{1 - 2kx\cos a + k^2x^2}$	$\|x\| < k^{-1}$
$\binom{n}{m}$	$x^m(1 - x)^{-m-1}$	$\|x\| < 1$
$\binom{k}{n}$	$(1 + x)^k$	$\|x\| < 1$

* The term $p_m(z)$ is a polynomial of degree m satisfying $p_{m+1}(z) = (mz + 1) \cdot p_m(z) + z(1 - z)\ p_m'(x)$, $p_1 = 1$.

To solve the linear difference equation $\Sigma_{i=0}^{k} p_i y_{n+k-i} = 0$, $p_0 = 1$ we associate with it the two formal series $P = p_0 + p_1x + \cdots + p_kx^k$ and $Y = y_0 + y_1x + y_2x^{2+} \cdots$. If $p(x)$ is the characteristic polynomial then $P(x) = x^k p(1/x) = \bar{p}(x)$. The *product* of the two series is $Q = YP = q_0 + q_1x + \cdots + q_{k-1}x^{k-1} + q_kx^k + \cdots$ where $q_n = \Sigma_{i=0}^{n} p_i y_{n-i}$. Because $p_{k+1} = p_{k+2} = \cdots = 0$, it is obvious that $q_{k+1} = q_{k+2} = \cdots = 0$ — that is, Q is a polynomial (formal series with finite number of terms). Then $Y = P^{-1}Q = q(x)/\bar{p}(x) = q(x)/x^k p(1/x)$, where p is the characteristic polynomial and $q(x) = \Sigma_{i=0}^{k} q_i x^i$. The roots of $\bar{p}(x)$ are x_i^{-1} where the x_i are the roots of $p(x)$.

Theorem 1. If the roots of $p(x)$ are less than one in absolute value, then $Y(x)$ converges for $|x| < 1$.

Thorem 2. If $p(x)$ has no roots greater than one in absolute value and those on the unit circle are simple roots, then the coefficients y_n of Y are bounded. Now $q_k = g_0$, $q_{n+k} = g_n$, and $Q(x) = Q_1(x) + x^k Q_2(x)$. Hence $\Sigma_{i=1}^{\infty} y_i x^i = [Q_1(x) + x^k Q_2(x)]/[\bar{p}(x)]$.

Example 19.4.2. Consider the equation $y_{n+1} + y_n = -(n + 1)$, $y_0 = 1$. Here $Q_1 = 1$, $Q_2 = -\Sigma_{n=0}^{\infty}(n + 1)x^n = -1/(1 - x)^2$.

$$G(x) = \frac{1 - x/(1-x)^2}{1 + x} = \frac{5}{4}\frac{1}{1+x} - \frac{1}{4}\frac{1}{1-x} - \frac{1}{2}\frac{x}{(1-x)^2}$$

Using the table term by term, we find $\Sigma_{n=0}^{\infty} y_n x^n = \Sigma_{n=0}^{\infty} [{}^5/_4(-1)^n - {}^1/_4 - {}^1/_2\, n]x^n$, so $y_n = {}^5/_4(-1)^n - {}^1/_4 - {}^1/_2\, n$.

References

Fort, T. 1948. *Finite Differences and Difference Equations in the Real Domain.* Oxford University Press, London.

Jordan, C. 1950. *Calculus of Finite Differences,* Chelsea, New York.

Jury, E. I. 1964. *Theory and Applications of the Z Transform Method.* John Wiley & Sons, New York.

Lakshmikantham, V. and Trigrante, D. 1988. *Theory of Difference Equations.* Academic Press, Boston, MA.

Levy, H. and Lessman, F. 1961. *Finite Difference Equations.* Macmillan, New York.

Miller, K. S. 1968. *Linear Difference Equations,* Benjamin, New York.

Wilf, W. S. 1994. *Generating Functionology,* 2nd ed. Academic Press, Boston, MA.

19.5 Differential Equations

William F. Ames

Any equation involving derivatives is called a *differential equation.* If there is only one independent variable the equation is termed a *total differential equation* or an *ordinary differential equation*. If there is more than one independent variable the equation is called a *partial differential equation.* If the highest-order derivative is the *n*th then the equation is said to be *n*th order. If there is no function of the dependent variable and its derivatives other than the linear one, the equation is said to be *linear.* Otherwise, it is *nonlinear.* Thus $(d^3y/dx^3) + a(dy/dx) + by = 0$ is a *linear* third-order ordinary (total) differential equation. If we replace by with by^3, the equation becomes nonlinear. An example of a second-order linear partial differential equation is the famous wave equation $(\partial^2u/\partial x^2) - a^2(\partial^2u/\partial t^2) = f(x)$. There are two independent variables x and t and $a^2 > 0$ (of course). If we replace $f(x)$ by $f(u)$ (say u^3 or $\sin u$) the equation is nonlinear. Another example of a nonlinear third-order partial differential equation is $u_t + uu_x = au_{xxx}$. This chapter uses the common subscript notation to indicate the partial derivatives.

Now we briefly indicate some methods of solution and the solution of some commonly occurring equations.

Ordinary Differential Equations

First-Order Equations

The *general* first-order equation is $f(x, y, y') = 0$. Equation capable of being written in either of the forms $y' = f(x)g(y)$ or $f(x)g(y)y' + F(x)G(y) = 0$ are *separable* equations. Their solution is obtained by using $y' = dy/dx$ and writing the equations in differential form as $dy/g(y) = f(x)dx$ or $g(y)[dy/G(y)] = -F(x)[dx/f(x)]$ and integrating. An example is the famous *logistic* equation of inhibited growth $(dy/dt) = ay(1 - y)$. The integral of $dy/y(1 - y) = adt$ is $y = 1/[1 + (y_0^{-1} - 1)e^{-at}]$ for $t \geq 0$ and $y(0) = y_0$ (the initial state called the *initial condition*).

Equations may not have unique solutions. An example is $y' = 2y^{1/2}$ with the initial condition $y(0) = 0$. One solution by separation is $y = x^2$. But there are an *infinity* of others — namely, $y_a(x) = 0$ for $-\infty < x \leq a$, and $(x - a)^2$ for $a \leq x < \infty$.

If the equation $P(x, y)dy + Q(x, y)dy = 0$ is reducible to

$$\frac{dy}{dx} = f\left(\frac{y}{x}\right) \quad \text{or} \quad \frac{dy}{dx} = f\left(\frac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$$

the equation is called *homogenous* (nearly homogeneous). The first form reduces to the separable equation $u + x(du/dx) = f(u)$ with the substitution $y/x = u$. The nearly homogeneous equation is handled by setting $x = X + a$, $y = Y + \beta$, and choosing α and β so that $a_1\alpha + b_1\beta + c_1 = 0$ and $a_2\alpha + b_2\beta + c_2 = 0$. If $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \neq 0$ this is always possible; the equation becomes $dY/dX = [a_1 + b_1(Y/X)]/[a_2 + b_2(Y/X)]$ and the substitution $Y = Xu$ gives a separable equation. If $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = 0$ then $a_2x + b_2y = k(a_1x + b_1y)$ and the equation becomes $du/dx = a_1 + b_1(u + c_1)/(ku + c_2)$, with $u = a_1x + b_1y$. Lastly, any equation of the form $dy/dx = f(ax + by + c)$ transforms into the separable equation $du/dx = a + bf(u)$ using the change of variable $u = ax + by + c$.

The general first-order linear equation is expressible in the form $y' + f(x)y = g(x)$. It has the *general solution* (a solution with an arbitrary constant c)

$$y(x) = \exp\left[-\int f(x)\,dx\right]\left\{c + \int \exp\left[f(x)\right]g(x)\,dx\right\}$$

Two noteworthy examples of first-order equations are as follows:

1. An often-occurring nonlinear equation is the *Bernoulli equation*, $y' + p(x)y = g(x)y^{\alpha}$, with α real, $\alpha \neq 0$, $\alpha \neq 1$. The transformation $z = y^{1-\alpha}$ converts the equation to the linear first-order equation $z' + (1 - \alpha)p(x)z = (1 - \alpha)q(x)$.
2. The famous *Riccati equation*, $y' = p(x)y^2 + q(x)y + r(x)$, cannot in general be solved by integration. But some useful transformations are helpful. The substitution $y = y_1 + u$ leads to the equation $u' - (2py_1 + q)u = pu^2$, which is a Bernoulli equation for u. The substitution $y = y_1 + v^{-1}$ leads to the equation $v' + (2py_1 + q)v + p = 0$, which is a linear first-order equation for v. Once either of these equations has been solved, the general solution of the Riccati equation is $y = y_1 + u$ or $y = y_1 + v^{-1}$.

Second-Order Equations

The simplest of the second-order equations is $y'' + ay' + by = 0$ (a, b real), with the initial conditions $y(x_0) = y_0$, $y'(x_0) = y_0'$ or the boundary conditions $y(x_0) = y_0$, $y(x_1) = y_1$. The general solution of the equation is given as follows.

1. $a^2 - 4b > 0$, $\lambda_1 = {}^1\!/_2\,(-a + \sqrt{a^2 - 4b})$, $\lambda_2 = {}^1\!/_2\,(-a - \sqrt{a^2 - 4b})$
   $y = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x)$
2. $a^2 - 4b = 0$, $\lambda_1 = \lambda_2 = -\frac{a}{2}$, $y = (c_1 + c_2 x)\exp(\lambda_1 x)$
3. $a^2 - 4b < 0$, $\lambda_1 = {}^1\!/_2\,(-a + i\sqrt{4b - a^2})$, $\lambda_2 = {}^1\!/_2\,(-a - i\sqrt{4b - a^2})$,
   $i^2 = -1$
   With $p = -a/2$ and $q = {}^1\!/_2\,\sqrt{4b - a^2}$,

$$y = c_1 \exp\left[(p + iq)x\right] + c_2 \exp\left[(p - iq)x\right] = \exp(px)\left[A\sin qx + B\cos qx\right]$$

   The initial conditions or boundary conditions are used to evaluate the arbitrary constants c_1 and c_2 (or A and B).
   Note that a linear problem with specified data may not have a solution. This is especially serious if numerical methods are employed without serious thought.

For example, consider $y'' + y = 0$ with the boundary condition $y(0) = 1$ and $y(\pi) = 1$. The general solution is $y = c_1 \sin x + c_2 \cos x$. The first condition $y(0) = 1$ gives $c_2 = 1$, and the second condition requires $y(\pi) = c_1 \sin \pi + \cos \pi$ or "1 = –1," which is a *contradiction.*

Example 19.5.1 — The Euler Strut. When a strut of uniform construction is subject to a compressive load P it exhibits no transverse displacement until P exceeds some critical value P_1. When this load is exceeded, buckling occurs and large deflections are produced as a result of small load changes. Let the rod of length ℓ be placed as shown in Figure 19.5.1.

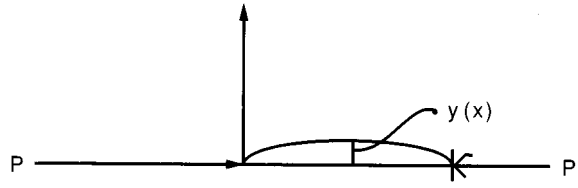

FIGURE 19.5.1

From the linear theory of elasticity (Timoshenko), the transverse displacement $y(x)$ satisfies the linear second-order equation $y'' + (Py/EI) = 0$, where E is the modulus of elasticity and I is the moment of inertia of the strut. The boundary conditions are $y(0) = 0$ and $y(a) = 0$. With $k^2 = P/EI$ the general solution is $y = c_1 \sin kx + c_2 \cos kx$. The condition $y(0) = 0$ gives $c_2 = 0$. The second condition gives $c_1 \sin ka = 0$. Since $c_1 = 0$ gives the trival solution $y = 0$ we must have $\sin ka = 0$. This occurs for $ka = n\pi$, $n = 0, 1, 2, \ldots$ (these are called *eigenvalues*). The first nontrivial solution occurs for $n = 1$ — that is, $k = \pi/a$ — whereupon $y_1 = c_1 \sin(\pi/a)$, with arbitrary c_1. Since $P = EIk^2$ the critical compressive load is $P_1 = EI\pi^2/a^2$. This is the buckling load. The weakness of the linear theory is its failure to model the situation when buckling occurs.

Example 19.5.2 — Some Solvable Nonlinear Equations. Many physical phenomena are modeled using nonlinear second-order equations. Some general cases are given here.

1. $y'' = f(y)$, *first integral* $(y')^2 = 2\int f(y)\,dy + c$.
2. $f(x, y', y'') = 0$. Set $p = y'$ and obtain a first-order equation $f(x, p, dp/dx) = 0$. Use first-order methods.
3. $f(y, y', y'') = 0$. Set $p = y'$ and then $y'' = p(dp/dy)$ so that a first-order equation $f[y, p, p(dp/dy) = 0$ for p as a function of y is obtained.
4. The *Riccati transformation* $du/dx = yu$ leads to the Riccati chain of equations, which linearize by raising the order. Thus,

Equation in y	Equation in u
1. $y' + y^2 = f(x)$	$u'' = f(x)u$
2. $y'' + 3yy' + y^3 = f(x)$	$u''' = f(x)u$
3. $y''' + 6y^2y' + 3(y')^2 + 4yy'' = f(x)$	$u^{(iv)} = f(x)u$

This method can be generalized to $u' = a(x)yu$ or $u' = a(x)f(u)y$.

Second-Order Inhomogeneous Equations

The general solution of $a_0(x)y'' + a_1(x)y' + a_2(x)y = f(x)$ is $y = y_H(x) + y_p(x)$ where $y_H(x)$ is the general solution of the homogeneous equation (with the right-hand side zero) and y_p is the particular integral of the equation. Construction of particular integrals can sometimes be done by the *method of undetermined coefficients*. See Table 19.5.1. This applies only to the linear constant coefficient case in which the function $f(x)$ is a linear combination of a polynomial, exponentials, sines and cosines, and some products of these functions. This method has as its base the observation that repeated differentiation of such functions gives rise to similar functions.

Table 19.5.1 Method of Undetermined Coefficients — Equation $L(y) = f(x)$ (Constant Coefficients)

Terms in $f(x)$	Terms To Be Included in $y_p(x)$
1. Polynomial of degree n	(i) If $L(y)$ contains y, try $y_p = a_0x^n + a_1x^{n-1} + \cdots + a_n$, (ii) If $L(y)$ does not contain y and lowest-order derivative is $y^{(r)}$, try $y_p = a_0x^{n+r} + \cdots + a_nx^r$.
2. $\sin qx$, $\cos qx$	(i) $\sin qx$ and/or $\cos qx$ are not in y_H; $y_p = B \sin qx + C \cos qx$. (ii) y_H contains terms of form $x^r \sin qx$ and/or $x^r \cos qx$ for $r = 0, 1, \ldots, m$; include in y_p terms of the form $a_0x^{m+1} \sin qx + a_1x^{m+1} \cos qx$.
3. e^{ax}	(i) y_H does not contain e^{ax}; include Ae^{ax} in y_p. (ii) y_H contains $e^{ax}, xe^{ax}, \ldots, x^ne^{ax}$; include in y_p terms of the form $Ax^{n+1}e^{ax}$,
4. $e^{px} \sin qx$, $e^{px} \cos qx$	(i) y_H does not contain these terms; in y_p include $Ae^{px} \sin qx + Be^{px} \cos qx$. (ii) y_H contains $x^r e^{px} \sin qx$ and/or $x^r e^{px} \cos qx$; $r = 0,1, \ldots, m$ include in y_p. $Ax^{m+1}e^{px} \sin qx + Bx^{m+1} e^{px} \cos qx$.

Example 19.5.3. Consider the equation $y'' + 3y' + 2y = \sin 2x$. The characteristic equation of the homogeneous equation $\lambda^2 + 3\lambda + 2 = 0$ has the two roots $\lambda_1 = -1$ and $\lambda_2 = -2$. Consequently, $y_H = c_1e^{-x} + c_2e^{-2x}$. Since $\sin 2x$ is not linearly dependent on the exponentials and since $\sin 2x$ repeats after two

differentiations, we assume a particular solution with undetermined coefficients of the form $y_p(x) = B \sin 2x + C \cos 2x$. Substituting into the original equation gives $-(2B + 6C) \sin 2x + (6B - 2C) \cos 2x = \sin 2x$. Consequently, $-(2B + 6C) = 1$ and $6B - 2C = 0$ to satisfy the equation. These two equations in two unknowns have the solution $B = -1/20$ and $C = -3/20$. Hence $y_p = -1/20 (\sin 2x + 3 \cos 2x)$ and $y = c_1e^{-x} + c_2e^{-2x} - 1/20 (\sin 2x + 3 \cos 2x)$.

A general method for finding $y_p(x)$ called *variation of parameters* uses as its starting point $y_H(x)$. This method applies to *all* linear differential equations irrespective of whether they have constant coefficients. But it assumes $y_H(x)$ is known. We illustrate the idea for $a(x)y'' + b(x)y' + c(x)y = f(x)$. If the solution of the homogeneous equation is $y_H(x) = c_1\phi_1(x) + c_2\phi_2(x)$, then vary the parameters c_1 and c_2 to seek $y_p(x)$ as $y_p(x) = u_1(x)\phi_1(x) + u_2(x)\phi_2(x)$. Then $y_P' = u_1\,\phi_1' + u_2\,\phi_2' + u_1'\,\phi_1 + u_2'\,\phi_2$ and choose $u_1'\,\phi_1 + u_2'\,\phi_2 = 0$. Calculating y_P'' and setting in the original equation gives $a(x)\,u_1'\phi_1' + a(x)\,u_2'\phi_2' = f$. Solving the last two equations for u_1' and u_2' gives $u_1' = -\phi_2 f/wa$, $u_2' = \phi_1 f/wa$, where $w = \phi_1\,\phi_2' - \phi_1'\,\phi_2 \neq 0$. Integrating the general solution gives $y = c_1\phi_1(x) + c_2\phi_2(x) - \{\int[\phi_2 f(x)]/wa\}\phi_1(x) + [\int(\phi_1 f/wa)dx]\phi_2(x)$.

Example 19.5.4. Consider the equations $y'' - 4y = \sin x/(1 + x^2)$ and $y_H = c_1e^{-2x} + c_2e^{-2x}$. With $\phi_1 = e^{2x}$, and $\phi_2 = e^{-2x}$, $w = 4$, so the general solution is

$$y = c_1e^{2x} + c_2e^{-2x} - \frac{e^{-2x}}{4}\int \frac{e^{2x}\sin x}{1+x^2}\,dx + \frac{e^{2x}}{4}\int \frac{e^{-2x}\sin x}{1+x^2}\,dx$$

The method of variation of parameters can be generalized as described in the references.

Higher-order systems of linear equations with constant coefficients are treated in a similar manner. Details can be found in the references.

Series Solution

The solution of differential equations can only be obtained in closed form in special cases. For all others, series or approximate or numerical solutions are necessary. In the simplest case, for an initial value problem, the solution can be developed as a Taylor series expansion about the point where the initial data are specified. The method fails in the *singular case* — that is, a point where the coefficient of the highest-order derivative is zero. The general method of approach is called the *Frobenius method.*

To understand the nonsingular case consider the equation $y'' + xy = x^2$ with $y(2) = 1$ and $y'(2) = 2$ (an initial value problem). We seek a series solution of the form $y(x) = a_0 + a_1(x - 2) + a_2(x - 2)^2 + \cdots$. To proceed, set $1 = y(2) = a_0$, which evaluates a_0. Next $y'(x) = a_1 + 2a_2(x - 2) + \cdots$, so $2 = y'(2) = a_1$ or $a_1 = 2$. Next $y''(x) = 2a_2 + 6a_3(x - 2) + \cdots$. and from the equation, $y'' = x^2 - xy$, so $y''(2) = 4 - 2y(2) = 4 - 2 = 2$. Hence $2 = 2a_2$ or $a_2 = 1$. Thus, to third-order $y(x) = 1 + 2(x - 2) + (x - 2)^2 + R_2(x)$, where the remainder $R_2(x)$ $[(x - 2)^3/3]y'''(\xi)$, where $2 < \xi < x$ can be bounded for each x by finding the maximum of $y'''(x) = 2x - y - xy'$. The third term of the series follows by evaluating $y'''(2) = 4 - 1 - 2 \cdot 2 = -1$, so $6a_3 = -1$ or $a_3 = -1/6$.

By now the nonsingular process should be familiar. The algorithm for constructing a series solution about a nonsingular (ordinary) point x_0 of the equation $P(x)y'' + Q(x)y' + R(x)y = f(x)$ (note that $P(x_0) \neq 0$) is as follows:

1. Substitute into the differential equation the expressions

$$y(x) = \sum_{n=0}^{\infty} a_n\left(x - x_0\right)^n, \qquad y'(x) = \sum_{n=1}^{\infty} na_n\left(x - x_0\right)^{n-1}, \qquad y''(x) = \sum_{n=2}^{\infty} n(n-1)a_n\left(x - x_0\right)^{n-2}$$

2. Expand $P(x)$, $Q(x)$, $R(x)$, and $f(x)$ about the point x_0 in a power series in $(x - x_0)$ and substitute these series into the equation.
3. Gather all terms involving the same power of $(x - x_0)$ to arrive at an identity of the form $\Sigma_{n=0}^{\infty}$ $A_n(x - x_0)^n \equiv 0$.

4. Equate to zero each coefficient A_n of step 3.
5. Use the expressions of step 4 to determine a_2, a_3, … in terms of a_0, a_1 (we need two arbitrary constants) to arrive at the general solution.
6. With the given initial conditions, determine a_0 and a_1.

If the equation has a regular singular point — that is, a point x_0 at which $P(x)$ vanishes and a series expansion is sought about that point — a solution is sought of the form $y(x) = (x - x_0)^r \sum_{n=0}^{\infty} a_n(x - x_0)^n$, $a_0 \neq 0$ and the index r and coefficients a_n must be determined from the equation by an algorithm analogous to that already described. The description of this Frobenius method is left for the references.

Partial Differential Equations

The study of partial differential equations is of continuing interest in applications. It is a vast subject, so the focus in this chapter will be on the most commonly occurring equations in the engineering literature — the second-order equations in two variables. Most of these are of the three basic types: elliptic, hyperbolic, and parabolic.

Elliptic equations are often called *potential equations* since they occur in potential problems where the potential may be temperature, voltage, and so forth. They also give rise to the steady solutions of parabolic equations. They require boundary conditions for the complete determination of their solution.

Hyperbolic equations are often called *wave equations* since they arise in the propagation of waves. For the development of their solutions, initial and boundary conditions are required. In principle they are solvable by the method of characteristics.

Parabolic equations are usually called *diffusion equations* because they occur in the transfer (diffusion) of heat and chemicals. These equations require initial conditions (for example, the initial temperature) and boundary conditions for the determination of their solutions.

Partial differential equations (PDEs) of the second order in two independent variables (x, y) are of the form $a(x, y)u_{xx} + b(x, y)u_{xy} + c(x, y)u_{yy} = E(x, y, u, u_x, u_y)$. If $E = E(x, y)$ the equation is linear; if E depends also on u, u_x, and u_y, it is said to be *quasilinear,* and if E depends only on x, y, and u, it is *semilinear.* Such equations are classified as follows: If $b^2 - 4ac$ is less than, equal to, or greater than zero at some point (x, y), then the equation is elliptic, parabolic, or hyperbolic, respectively, at that point. A PDE of this form can be transformed into canonical (standard) forms by use of new variables. These standard forms are most useful in analysis and numerical computations.

For hyperbolic equations the standard form is $u_{\xi\eta} = \phi(u, u_\eta, u_\xi, \eta, \xi)$, where $\xi_x/\xi_y = (-b + \sqrt{b^2 - 4ac})/2a$, and $\eta_x/\eta_y = (-b - \sqrt{b^2 - 4ac})/2a$. . The right-hand sides of these equations determine the so-called characteristics $(dy/dx)|_+ = (-b + \sqrt{b^2 - 4ac})/2a$, $(dy/dx)|_- = (-b - \sqrt{b^2 - 4ac})/2a$.

Example 19.5.5. Consider the equation $y^2u_{xx} - x^2u_{yy} = 0$, $\xi_x/\xi_y = -x/y$, $\eta_x/\eta_y = x/y$, so $\xi = y^2 - x^2$ and $\eta = y^2 + x^2$. In these new variables the equation becomes $u_{\xi\eta} = (\xi u_\eta - \eta u_\xi)/2(\xi^2 - \eta^2)$.

For parabolic equations the standard form is $u_{\xi\xi} = \phi(u, u_\eta, u_\xi, \eta, \xi)$ or $u_{\eta\eta} = \phi(u, u_\eta, u_\xi, \xi, \eta)$, depending upon how the variables are defined. In this case $\xi_x/\xi_y = -b/2a$ if $a \neq 0$, and $\xi_x/\xi_y = -b/2c$ if $c \neq 0$. Only ξ must be determined (there is only one characteristic) and η can be chosen as any function that is linearly independent of ξ.

Example 19.5.6. Consider the equation $y^2u_{xx} - 2xyu_{xy} + x^2u_{yy} + u_y = 0$. Clearly, $b^2 - 4ac = 0$. Neither a nor c is zero so either path can be chosen. With $\xi_x/\xi_y = -b/2a = x/y$, there results $\xi = x^2 + y^2$. With $\eta = x$, the equation becomes $u_{\eta\eta} = [2(\xi + \eta)u_\xi + u_\eta]/(\xi - \eta^2)$.

For *elliptic equations* the standard form is $u_{\alpha\alpha} + u_{\beta\beta} = \phi(u, u_\alpha, u_\beta, \alpha, \beta)$, where ξ and η are determined by solving the ξ and η equations of the hyperbolic system (they are complex) and taking $\alpha = (\eta + \xi)/2$, $\beta = (\eta - \xi)/2i (i^2 = -1)$. Since ξ and η are complex conjugates, both α and β are real.

Example 19.5.7. Consider the equation $y^2u_{xx} + x^2u_{yy} = 0$. Clearly, $b^2 - 4ac < 0$, so the equation is elliptic. Then $\xi_x/\xi_y = -ix/y$, $\eta_x/\eta_y = ix/y$, so $\alpha = (\eta + \xi)/2 = y^2$ and $\beta = (\eta - \xi)/2i = x^2$. The standard form is $u_{\alpha\alpha} + u_{\beta\beta} = -(u_\alpha/2\alpha + u_\beta/2\beta)$.

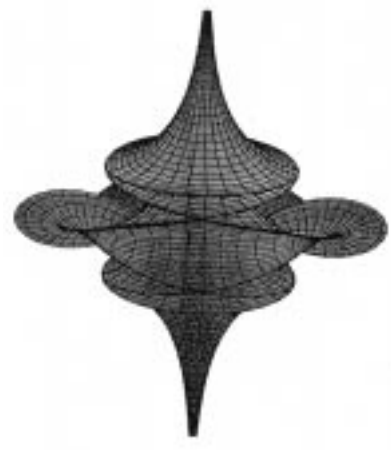

Figure 19.5.2

Figure 19.5.3

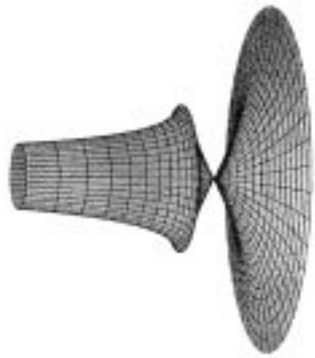

Figure 19.5.4

Figure 19.5.5

FIGURE 19.5.2 to 19.5.5 The mathematical equations used to generate these three-dimensional figures are worth a thousand words. The figures shown illustrate some of the nonlinear ideas of engineering, applied physics, and chemistry. Figure 19.5.2 represents a breather soliton surface for the sine-Gordon equation $w_{uv} = \sin w$ generated by a Backlund transformation. A single-soliton surface for the sine-Gordon equation $w_{uv} = \sin w$ is illustrated in Figure 19.5.3. Figure 19.5.4 represents a single-soliton surface for the Tzitzecia-Dodd-Bullough equation associated with an integrable anisentropic gas dynamics system. Figure 19.5.5 represents a single-soliton Bianchi surface. The solutions to the equations were developed by W. K. Schief and C. Rogers at the Center for Dynamical Systems and Nonlinear Studies at the Georgia Institute of Technology and the University of New South Wales in Sydney, Australia. All of these three-dimensional projections were generated using the MAPLE software package. (Figures courtesy of Schief and Rogers).

Methods of Solution

Separation of Variables. Perhaps the most elementary method for solving linear PDEs with homogeneous boundary conditions is the method of *separation of variables.* To illustrate, consider $u_t - u_{xx} = 0$, $u(x, 0) = f(x)$ (the initial condition) and $u(0, t) = u(1, t) = 0$ for $t > 0$ (the boundary conditions). A solution is assumed in "separated form" $u(x, t) = X(x)T(t)$. Upon substituting into the equation we find $\dot{T}/T = X''/X$ (where $\dot{T} = dT/dt$ and $X'' = d^2X/dx^2$). Since $T = T(t)$ and $X = X(x)$, the ratio must be constant, and for finiteness in t the constant must be negative, say $-\lambda^2$. The solutions of the separated equations $X'' + \lambda^2 X = 0$ with the boundary conditions $X(0) = 0$, $X(1) = 0$, and $\dot{T} = -\lambda^2 T$ are $X = A \sin \lambda x + B \cos \lambda x$ and $T = Ce^{-\lambda^2 t}$, where A, B, and C are arbitrary constants. To satisfy the boundary condition $X(0) = 0$, $B = 0$. An infinite number of values of λ (eigenvalues), say $\lambda_n = n\pi (n = 1, 2, 3, \ldots)$, permit all the eigenfunctions $X_n = b_n \sin \lambda_n x$ to satisfy the other boundary condition $X(1) = 0$. The solution of the

equation and boundary conditions (not the initial condition) is, by superposition, $u(x, t) = \Sigma_{n=1}^{\infty} b_n e^{-n^2\pi^2 t} \cdot \sin n\pi x$ (a Fourier sine series), where the b_n are arbitrary. These values are obtained from the initial condition using the orthogonality properties of the trigonometric function (e.g., $\int_{-\pi}^{\pi} \sin mx \sin n\,x\,d\,x$ is 0 for $m \neq n$ and is π for $m = n \neq 0$) to be $b_n = 2\int_0^1 f(r) \sin n\,\pi\,r\,d\,r$. Then the solution of the problem is $u(x, t) = \Sigma_{n=1}^{\infty}[2\int_0^1 f(r) \sin n\,\pi\,r\,d\,r]\, e^{-n^2\pi^2 t} \sin n\,\pi\,x$, which is a Fourier sine series.

If $f(x)$ is a piecewise smooth or a piecewise continuous function defined for $a \le x \le b$, then its Fourier series within $a \le x \le b$ as its fundamental interval (it is extended periodically ouside that interval) is

$$f(x) \sim \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left[2n\pi x/(b-a)\right] + b_n \sin\left[2n\pi x/(b-a)\right]$$

where

$$a_n = \left[\frac{2}{(b-a)}\right]\int_a^b f(x)\cos\left[2n\pi x/(b-a)\right] dx, \quad n = 0,\ 1,\ \ldots$$

$$b_n = \left[\frac{2}{(b-a)}\right]\int_a^b f(x)\sin\left[2n\pi x/(b-a)\right] dx, \quad n = 1,\ 2,\ \ldots$$

The Fourier sine series has $a_n \equiv 0$, and the Fourier cosine series has $b_n \equiv 0$. The symbol ~ means that the series converges to $f(x)$ at points of continuity, and at the (allowable) points of finite discontinuity the series converges to the *average value* of the discontinuous values.

Caution: This method *only* applies to linear equations with homogeneous boundary conditions. Linear equations with variable coefficients use other orthogonal functions, such as the Besel functions, Laguerre functions, Chebyshev functions, and so forth.

Some inhomogeneous boundary value problems can be transformed into homogeneous ones. Consider the problem $u_t - u_{xx} = 0$, $0 \le x \le 1$, $0 \le t < \infty$ with initial condition $u(x, 0) = f(x)$, and boundary conditions $u(0, t) = g(t)$, $u(1, t) = h(t)$. To homogenize the boundary conditions set $u(x, t) = w(x, t) + x[h(t) - g(t)] + g(t)$ and then solve $w_t - w_x = [\dot{g}(t) - \dot{h}(t)]x - \dot{g}(t)$ with the initial condition $w(x, 0) = f(x) - x[h(0) - g(0)] + g(0)$ and $w(0, t) = w(1, t) = 0$.

Operational Methods. A number of integral transforms are useful for solving a variety of linear problems. To apply the Laplace transform to the problem $u_t - u_{xx} = \delta(x)\,\delta(t)$, $-\infty < x < \infty$, $0 \le t$ with the initial condition $u(x, 0^-) = 0$, where δ is the Dirac delta function, we multiply by e^{-st} and integrate with respect to t from 0 to ∞. With the Laplace transform of $u(x, t)$ denoted by $U(x, s)$ — that is, $U(x, s) = \int_0^{\infty} e^{-st}\,u(x, t)\,dt$ — we have $sU - U_{xx} = \delta(x)$, which has the solution

$$U(x,s) = A(s)e^{-x\sqrt{s}} + B(s)e^{x\sqrt{s}} \qquad \text{for } x > 0$$

$$U(x,s) = C(s)e^{-x\sqrt{s}} + D(s)e^{x\sqrt{s}} \qquad \text{for } x < 0$$

Clearly, $B(s) = C(s) = 0$ for bounded solutions as $|x| \to \infty$. Then, from the boundary condition, $U(0^+, s) - U(0^-, s) = 0$ and integration of $sU - U_{xx} = \delta(x)$ from 0^- to 0^+ gives $U_x(0^+, s) - U(0^-, s) = -1$, so $A = D = 1/2\sqrt{s}$. Hence, $U(x, s) = (1/2\sqrt{s})e^{-\sqrt{s}|x|}$ and the inverse is $u(x, t) = (1/2\pi i)\int_{\Gamma} e^{st}U(x, s)\,ds$, where Γ is a Bromwich path, a vertical line taken to the right of all singularities of U on the sphere.

Similarity (Invariance). This very useful approach is related to dimensional analysis; both have their foundations in group theory. The three important transformations that play a basic role in Newtonian mechanics are translation, scaling, and rotations. Using two independent variables x and t and one dependent variable $u = u(x, t)$, the *translation group* is $\bar{x} = x + \alpha a$, $\bar{t} = t + \beta a$, $\bar{u} = u + \gamma a$; the *scaling*

group is $\bar{x} = a^{\alpha}x$, $\bar{t} = a^{\beta}t$, and $\bar{u} = a^{\gamma}u$; the *rotation group* is $\bar{x} = x\cos a + t\sin a$, $\bar{t} = t\cos a - x\sin a$, $\bar{u} = u$, with a nonnegative real number a. Important in which follows are the *invariants* of these groups. For the translation group there are two $\eta = x - \lambda t$, $\lambda = \alpha/\beta$, $f(\eta) = u - \varepsilon t$, $\varepsilon = \gamma/\beta$ or $f(\eta) = u - \theta x$, $\theta = \gamma/\alpha$; for the scaling group the invariants are $\eta = x/t^{\alpha/\beta}$ (or $t/x^{\beta/\alpha}$) and $f(\eta) = u/t^{\gamma/\beta}$ (or $u/x^{\gamma/\alpha}$); for the rotation group the invariants are $\eta = x^2 + t^2$ and $u = f(\eta) = f(x^2 + t^2)$.

If a PDE and its data (initial and boundary conditions) are left invariant by a transformation group, then similar (invariant) solutions are sought using the invariants. For example, if an equation is left invariant under scaling, then solutions are sought of the form $u(x, t) = t^{\gamma/\beta} f(\eta)$, $\eta = xt^{-\alpha/\beta}$ or $u(x, t) = x^{\gamma/\alpha} f(tx^{-\beta/\alpha})$; invariance under translation gives solutions of the form $u(x, t) = f(x - \lambda t)$; and invariance under rotation gives rise to solutions of the form $u(x, t) = f(x^2 + t^2)$.

Examples of invariance include the following:

1. The equation $u_{xx} + u_{yy} = 0$ is invariant under rotation, so we search for solutions of the form $u = f(x^2 + y^2)$. Substitution gives the ODE $f' + \eta f'' = 0$ or $(\eta f')' = 0$. The solution is $u(x, t) = c \ln \eta = c \ln(x^2 + t^2)$, which is the (so-called) fundamental solution of Laplace's equation.
2. The nonlinear diffusion equation $u_t = (u^n u_x)_x$ $(n > 0)$, $0 \le x$, $0 \le t$, $u(0, t) = ct^n$ is invariant under scaling with the similar form $u(x, t) = t^n f(\eta)$, $\eta = xt^{-(n+1)/2}$. Substituting into the PDE gives the equation $(f^n f')' + ((n + 1)/2)\eta f' - nf = 0$, with $f(0) = c$ and $f(\infty) = 0$. Note that the equation is an ODE.
3. The wave equation $u_{xx} - u_{tt} = 0$ is invariant under translation. Hence, solutions exist of the form $u = f(x - \lambda t)$. Substitution gives $f''(1 - \lambda^2) = 0$. Hence, $\lambda = \pm 1$ or f is linear. Rejecting the trivial linear solution we see that $u = f(x - t) + g(x + t)$, which is the general (d'Alembert) solution of the wave equation; the quantities $x - t = \alpha$, $x + t = \beta$ are the characteristics of the next section.

The construction of all transformations that leave a PDE invariant is a solved problem left for the references.

The study of "solitons" (solitary traveling waves with special properties) has benefited from symmetry considerations. For example, the nonlinear third-order (Korteweg-de Vries) equation $u_t + uu_x - au_{xxx} = 0$ is invariant under translation. Solutions are sought of the form $u = f(x - \lambda t)$, and f satisfies the ODE, in $\eta = x - \lambda t$, $-\lambda f' + ff' - af''' = 0$.

Characteristics. Using the characteristics the solution of the hyperbolic problem $u_{tt} - u_{xx} = p(x, t)$, $-\infty < x < \infty$, $0 \le t$, $u(x, 0) = f(x)$, $u_t(x, 0) = h(x)$ is

$$u(x,t) = \tfrac{1}{2}\int_0^t d\tau \int_{x-(t-\tau)}^{x+(t-\tau)} p(\xi,\tau)\, d\xi + \tfrac{1}{2}\int_{x-t}^{x+t} h(\xi)\, d\xi + \tfrac{1}{2}\big[f(x+t) + f(x-t)\big]$$

The solution of $u_{tt} - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u_t(x, 0) = h(x)$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = {}^1/_2 \int_{-x+t}^{x+t} h(\xi)\, d\xi$.

The solution of $u_{tt} - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u_t(x, 0) = 0$, $u(0, t) = g(t)$, $t > 0$ is

$$u(x,t) = \begin{cases} 0 & \text{if } t < x \\ g(t-x) & \text{if } t > x \end{cases}$$

From time to time, lower-order derivatives appear in the PDE in use. To remove these from the equation $u_{tt} - u_{xx} + au_x + bu_t + cu = 0$, where a, b, and c are constants, set $\xi = x + t$, $\mu = t - x$, whereupon $u(x, t) = u[(\xi - \mu)/2, (\xi + \mu)/2] = U(\xi, \mu)$, where $U_{\xi\mu} + [(b + a)/4]\, U_{\xi} + [(b - a)/4]\, U_{\mu} + (c/4)U = 0$. The transformation $U(\xi, \mu) = W(\xi, \mu) \exp[-(b - a)\xi/4 - (b + a)\mu/4]$ reduces to satisfying $W_{\xi\mu} + \lambda W = 0$, where $\lambda = (a^2 - b^2 + 4c)/16$. If $\lambda \ne 0$, we lose the simple d'Alembert solution. But the equation for W is still easier to handle.

In linear problems discontinuities propagate along characteristics. In nonlinear problems the situation is usually different. The characteristics are often used as new coordinates in the numerical method of characteristics.

Green's Function. Consider the diffusion problem $u_t - u_{xx} = \delta(t)\delta(x - \xi)$, $0 \le x < \infty$, $\xi > 0$, $u(0, t) = 0$, $u(x, 0) = 0$ [$u(\infty, t) = u(\infty, 0) = 0$], a problem that results from a unit source somewhere in the domain subject to a homogeneous (zero) boundary condition. The solution is called a *Green's function of the first kind.* For this problem there is $G_1(x, \xi, t) = F(x - \xi, t) - F(x + \xi, t)$, where $F(x, t) = e^{-x^2/4t}/\sqrt{4\pi t}$ is the *fundamental* (invariant) *solution*. More generally, the solution of $u_t - u_{xx} = \delta(x - \xi)\,\delta(t - \tau)$, $\xi > 0$, $\tau > 0$, with the same conditions as before, is the Green's function of the first kind.

$$G_1(x,\xi,t-\tau) = \frac{1}{\sqrt{4\pi(t-\tau)}}\left[e^{-(x-\xi)^2/4(t-\tau)} - e^{-(x+\xi)^2/4(t-\tau)}\right]$$

for the semi-infinite interval.

The solution of $u_t - u_{xx} = p(x, t)$, $0 \le x < \infty$, $0 \le t < \infty$, with $u(x, 0) = 0$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = \int_0^t d\tau \int_0^\infty p(\xi, \tau)G_1(x, \xi, t - \tau)]\, d\xi$, which is a superposition. Note that the Green's function and the desired solution must both satisfy a zero boundary condition at the origin for this solution to make sense.

The solution of $u_t - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = f(x)$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = \int_0^\infty f(\xi)G_1(x, \xi, t)\, d\xi$.

The solution of $u_t - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u(0, t) = g(t)$, $t > 0$ (nonhomogeneous) is obtained by transforming to a new problem that has a homogeneous boundary condition. Thus, with $w(x, t) = u(x, t) - g(t)$ the equation for w becomes $w_t - w_{xx} = -\dot{g}(t) - g(0)\,\delta(t)$ and $w(x, 0) = 0$, $w(0, t) = 0$. Using G_1 above, we finally obtain $u(x, t) = (x/\sqrt{4\pi})\int_0^t g(t - \tau)e^{-x^2/4t}/\tau^{3/2}\, d\tau$.

The Green's function approach can also be employed for elliptic and hyperbolic problems.

Equations in Other Spatial Variables. The sperically symmetric wave equation $u_{rr} + 2u_r/r - u_{tt} = 0$ has the general solution $u(\tau, t) = [f(t - r) + g(t + r)]/r$.

The Poisson-Euler-Darboux equation, arising in gas dynamics,

$$u_{rs} + N\left(u_r + u_s\right)/(r+s) = 0$$

where N is a positive integer ≥ 1, has the general solution

$$u(r,s) = k + \frac{\partial^{N-1}}{\partial r^{N-1}}\left[\frac{f(r)}{(r+s)^N}\right] + \frac{\partial^{N-1}}{\partial s^{N-1}}\left[\frac{g(s)}{(r+s)^N}\right]$$

Here, k is an arbitrary constant and f and g are arbitrary functions whose form is determined from the problem initial and boundary conditions.

Conversion to Other Orthogonal Coordinate Systems. Let (x^1, x^2, x^3) be rectangular (Cartesian) coordinates and (u^1, u^2, u^3) be any orthogonal coordinate system related to the rectangular coordinates by $x^i = x^i(u^1, u^2, u^3)$, $i = 1, 2, 3$. With $(ds)^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 = g_{11}(du^1)^2 + g_{22}(du^2)^2 + g_{33}(du^3)^2$, where $g_{ii} = (\partial x^1/\partial u^i)^2 + (\partial x^2/\partial u^i)^2 + (\partial x^3/\partial u^i)^2$. In terms of these "metric" coefficients the basic operations of applied mathematics are expressible. Thus (with $g = g_{11}g_{22}g_{33}$)

$$dA = \left(g_{11}g_{22}\right)^{1/2} du^1\, du^2; \qquad dV = \left(g_{11}g_{22}g_{33}\right)^{1/2} du^1\, du^2\, du^3$$

$$\text{grad}\,\phi = \frac{\vec{a}_1}{(g_{11})^{1/2}}\frac{\partial\phi}{\partial u^1} + \frac{\vec{a}_2}{(g_{22})^{1/2}}\frac{\partial\phi}{\partial u^2} + \frac{\vec{a}_3}{(g_{33})^{1/2}}\frac{\partial\phi}{\partial u^3}$$

($\vec{a}_i$ are unit vectors in direction i);

$$\text{div}\,\vec{E} = g^{-1/2}\left\{\frac{\partial}{\partial u^1}\left[(g_{22}g_{33})^{1/2}E_1\right] + \frac{\partial}{\partial u^2}\left[(g_{11}g_{33})^{1/2}E_2\right] + \frac{\partial}{\partial u^3}\left[(g_{11}g_{22})^{1/2}E_3\right]\right\}$$

[here $\vec{E} = (E_1, E_2, E_3)$];

$$\begin{aligned}\text{curl}\,\vec{E} = g^{-1/2}\Bigg\{&\vec{a}_1(g_{11})^{1/2}\left(\frac{\partial}{\partial u^2}\left[(g_{33})^{1/2}E_3\right] - \frac{\partial}{\partial u^3}\left[(g_{22})^{1/2}E_2\right]\right)\\ &+ \vec{a}_2(g_{22})^{1/2}\left(\frac{\partial}{\partial u^3}\left[(g_{11})^{1/2}E_1\right] - \frac{\partial}{\partial u^1}\left[(g_{33})^{1/2}E_3\right]\right)\\ &+ \vec{a}_3(g_{33})^{1/2}\left(\frac{\partial}{\partial u^1}\left[(g_{22})^{1/2}E_2\right] - \frac{\partial}{\partial u^2}\left[(g_{11})^{1/2}E_1\right]\right)\Bigg\}\end{aligned}$$

$$\text{div}\,\text{grad}\,\psi = \nabla^2\psi = \text{Laplacian of } \psi = g^{-1/2}\sum_{i=1}^{3}\frac{\partial}{\partial u^i}\left[\frac{g^{1/2}}{g_{ii}}\frac{\partial\psi}{\partial u^i}\right]$$

Table 19.5.2 shows some coordinate systems.

Table 19.5.2 Some Coordinate Systems

Coordinate System	Metric Coefficients	
Circular Cylindrical		
$x = r\cos\theta$	$u^1 = r$	$g_{11} = 1$
$y = r\sin\theta$	$u^2 = \theta$	$g_{22} = r^2$
$z = z$	$u^3 = z$	$g_{33} = 1$
Spherical		
$x = r\sin\psi\cos\theta$	$u^1 = r$	$g_{11} = 1$
$y = r\sin\psi\sin\theta$	$u^2 = \psi$	$g_{22} = r^2$
$z = r\cos\psi$	$u^3 = \theta$	$g_{33} = r^2\sin^2\psi$
Parabolic Coordinates		
$x = \mu\nu\cos\theta$	$u^1 = \mu$	$g_{11} = \mu^2 + \nu^2$
$y = \mu\nu\sin\theta$	$u^2 = \nu$	$g_{22} = \mu^2 + \nu^2$
$z = 1/2\,(\mu^2 - \nu^2)$	$u^3 = \theta$	$g_{33} = \mu^2\nu^2$

Other metric coefficients and so forth can be found in Moon and Spencer [1961].

References

Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 1.* Academic Press, Boston, MA.

Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 2.* Academic Press, Boston, MA.

Brauer, F. and Nohel, J. A. 1986. *Introduction to Differential Equations with Applications,* Harper & Row, New York.
Jeffrey, A. 1990. *Linear Algebra and Ordinary Differential Equations,* Blackwell Scientific, Boston, MA.
Kevorkian, J. 1990. *Partial Differential Equations.* Wadsworth and Brooks/Cole, Belmont, CA.
Moon, P. and Spencer, D. E. 1961. *Field Theory Handbook,* Springer, Berlin.
Rogers, C. and Ames, W. F. 1989. *Nonlinear Boundary Value Problems in Science and Engineering.* Academic Press, Boston, MA.
Whitham, G. B. 1974. *Linear and Nonlinear Waves.* John Wiley & Sons, New York.
Zauderer, E. 1983. *Partial Differential Equations of Applied Mathematics.* John Wiley & Sons, New York.
Zwillinger, D. 1992. *Handbook of Differential Equations.* Academic Press, Boston, MA.

Further Information

A collection of solutions for linear and nonlinear problems is found in E. Kamke, *Differential-gleichungen-Lösungsmethoden und Lösungen,* Akad. Verlagsges, Leipzig, 1956. Also see G. M. Murphy, *Ordinary Differential Equations and Their Solutions,* Van Nostrand, Princeton, NJ, 1960 and D. Zwillinger, *Handbook of Differential Equations,* Academic Press, Boston, MA, 1992. For nonlinear problems see

Ames, W. F. 1968. *Ordinary Differential Equations in Transport Phenomena.* Academic Press, Boston, MA.
Cunningham, W. J. 1958. *Introduction to Nonlinear Analysis.* McGraw-Hill, New York.
Jordan, D. N. and Smith, P. 1977. *Nonlinear Ordinary Differential Equations.* Clarendon Press, Oxford, UK.
McLachlan, N. W. 1955. *Ordinary Non-Linear Differential Equations in Engineering and Physical Sciences,* 2nd ed. Oxford University Press, London.
Zwillinger, D. 1992.

19.6 Integral Equations

William F. Ames

Classification and Notation

Any equation in which the unknown function $u(x)$ appears under the integral sign is called an *integral equation*. If $f(x)$, $K(x, t)$, a, and b are known then the integral equation for u, $\int_a^b K(x, t)\, u(t)\, dt = f(x)$ is called a *linear integral equation of the first kind of Fredholm type*. $K(x, t)$ is called the *kernel function* of the equation. If b is replaced by x (the independent variable) the equation is an equation of *Volterra type of the first kind*.

An equation of the form $u(x) = f(x) + \lambda \int_a^b K(x, t)u(t)\, dt$ is said to be a linear integral equation of *Fredholm type of the second kind*. If b is replaced by x it is of *Volterra type*. If $f(x)$ is not present the equation is homogeneous.

The equation $\phi(x)\, u(x) = f(x) + \lambda \int_a^{b \text{ or } x} K(x, t)u(t)\, dt$ is the *third kind equation* of Fredholm or Volterra type. If the unknown function u appears in the equation in any way other than to the first power then the integral equation is said to be *nonlinear*. Thus, $u(x) = f(x) + \int_a^b K(x, t) \sin u(t)\, dt$ is nonlinear. An integral equation is said to be *singular* when either or both of the limits of integration are infinite or if $K(x, t)$ becomes infinite at one or more points of the integration interval.

Example 19.6.1. Consider the singular equations $u(x) = x + \int_0^\infty \sin(x\, t)\, u(t)\, dt$ and $f(x) = \int_0^x [u(t)/(x - t)^2]\, dt$.

Relation to Differential Equations

The *Leibnitz rule* $(d/dx) \int_{a(x)}^{b(x)} F(x, t)\, dt \int_{a(x)}^{b(x)} (\partial F/\partial x)\, dt + F[x, b(x)](db/dx) - F[x, a(x)] \times (da/dx)$ is useful for differentiation of an integral involving a parameter (x in this case). With this, one can establish the relation

$$I_n(x) = \int_a^x (x - t)^{n-1} f(t)\, dt = (n-1)! \underbrace{\int_a^x \cdots \int_a^x}_{n \text{ times}} f(x)\, \underbrace{dx \ldots dx}_{n \text{ times}}$$

This result will be used to establish the relation of the second-order initial value problem to a Volterra integral equation.

The second-order differential equation $y''(x) + A(x)y'(x) + B(x)y = f(x), y(a) = y_0, y'(a) = y_0'$ is equivalent to the integral equations

$$y(x) = -\int_a^x \left\{A(t) + (x - t)\left[B(t) - A'(t)\right]\right\} y(t)\, dt + \int_a^x (x - t) f(t)\, dt + \left[A(a)\, y_0 + y_0'\right](x - a) + y_0$$

which is of the type $(x)y = \int_a^x K(x, t)y(t)\, dt + F(x)$ where $K(x, t) = (t - x)[B(t) - A'(t)] - A(t)$ and $F(x)$ includes the rest of the terms. Thus, this initial value problem is equivalent to a Volterra integral equation of the second kind.

Example 19.6.2. Consider the equation $y'' + x^2y' + xy = x$, $y(0) = 1$, $y'(0) = 0$. Here $A(x) = x^2$, $B(x) = x$, $f(x) = x$, $a = 0$, $y_0 = 1$, $y_0' = 0$. The integral equation is $y(x) = \int_0^x t(x - 2t)y(t)\, dt + (x^3/6) + 1$.

The expression for $I_n(x)$ can also be useful in converting boundary value problems to integral equations. For example, the problem $y''(x) + \lambda y = 0$, $y(0) = 0$, $y(a) = 0$ is equivalent to the Fredholm equation $y(x) = \lambda \int_0^a K(x, t)y(t)\, dt$, where $K(x, t) = (t/a)(a - x)$ when $t < x$ and $K(x, t) = (x/a)(a - t)$ when $t > x$.

In both cases the differential equation can be recovered from the integral equation by using the Leibnitz rule.

Nonlinear differential equations can also be transformed into integral equations. In fact this is one method used to establish properties of the equation and to develop approximate and numerical solutions. For example, the "forced pendulum" equation $y''(x) + a^2 \sin y(x) = f(x)$, $y(0) = y(1) = 0$ transforms into the nonlinear Fredholm equation.

$$y(x) = \int_0^1 K(x,t)\left[a^2 \sin y(t) - f(t)\right] dt$$

with $K(x, t) = x(1 - t)$ for $0 < x < t$ and $K(x, t) = t(1 - x)$ for $t < x < 1$.

Methods of Solution

Only the simplest integral equations can be solved exactly. Usually approximate or numerical methods are employed. The advantage here is that integration is a "smoothing operation," whereas differentiation is a "roughening operation." A few exact and approximate methods are given in the following sections. The numerical methods are found under 19.12.

Convolution Equations

The special convolution equation $y(x) = f(x) + \lambda \int_0^x K(x - t)y(t)\, dt$ is a special case of the Volterra equation of the second kind. $K(x - t)$ is said to be a *convolution kernel*. The integral part is the convolution integral discussed under 19.8. The solution can be accomplished by transforming with the Laplace transform: $L[y(x)] = L[f(x)] + \lambda L[y(x)]L[K(x)]$ or $y(x) = L^{-1}\{L[f(x)]/(1 - \lambda L[K(x)])\}$.

Abel Equation

The Volterra equation $f(x) = \int_0^x y(t)/(x - t)^\alpha\, dt$, $0 < \alpha < 1$ is the (singular) Abel equation. Its solution is $y(x) = (\sin \alpha\pi/\pi)(d/dx) \int_0^x F(t)/(x - t)^{1-\alpha} dt$.

Approximate Method (Picard's Method)

This method is one of successive approximations that is described for the equation $y(x) = f(x) + \lambda \int_a^x K(x, t)y(t)\, dt$. Beginning with an initial guess $y_0(t)$ (often the value at the initial point a) generate the next approximation with $y_1(x) = f(x) + \int_a^x K(x, t)y_0(t)\, dt$ and continue with the general iteration

$$y_n(x) = f(x) + \lambda \int_a^x K(x,t)y_{n-1}(t)\, dt$$

Then, by iterating, one studies the convergence of this process, as is described in the literature.

Example 19.6.3. Let $y(x) = 1 + \int_0^x xt[y(t)]^2\, dt$, $y(0) = 1$, With $y_0(t) = 1$ we find $y_1(x) = 1 + \int_0^x xt\, dt = 1 + (x^3/2)$ and $y_2(x) = 1 + \int_0^x xt[1 + (t^3/2)^2 dt$, and so forth.

References

Jerri, A. J. 1985. *Introduction to Integral Equations with Applications*, Marcel Dekker, New York.
Tricomi, F. G. 1958. *Integral Equations.* Wiley-Interscience, New York.
Yosida, K. 1960. *Lectures on Differential and Integral Equations.* Wiley-Interscience, New York.

19.7 Approximation Methods

William F. Ames

The term *approximation methods* usually refers to an analytical process that generates a symbolic approximation rather than a numerical one. Thus, $1 + x + x^2/2$ is an approximation of e^x for small x. This chapter introduces some techniques for approximating the solution of various operator equations.

Perturbation

Regular Perturbation

This procedure is applicable to *some* equations in which a small parameter, ε, appears. Use this procedure with care; the procedure involves expansion of the dependent variables and data in a power series in the small parameter. The following example illustrates the procedure.

Example 19.7.1. Consider the equation $y'' + \varepsilon y' + y = 0$, $y(0) = 1$, $y'(0) = 0$. Write $y(x; \varepsilon) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \cdots$, and the initial conditions (data) become

$$y_0(0) + \varepsilon y_1(0) + \varepsilon^2 y_2(0) + \cdots = 1$$

$$y_0'(0) + \varepsilon y_1'(0) + \varepsilon^2 y_2'(0) + \cdots = 0$$

Equating like powers of ε in all three equations yields the sequence of equations

$$O(\varepsilon^0): y_0'' + y_0 = 0, \qquad y_0(0) = 1, \qquad y_0'(0) = 0$$

$$O(\varepsilon^1): y_1'' + y_1 = -y_0', \qquad y_1(0) = 0, \qquad y_1'(0) = 0$$

$$\vdots$$

The solution for y_0 is $y_0 = \cos x$ and using this for y_1 we find $y_1(x) = 1/2 (\sin x - x \cos x)$. So $y(x; \varepsilon) = \cos x + \varepsilon(\sin x - x \cos x)/2 + O(\varepsilon^2)$. Appearance of the term $x \cos x$ indicates a *secular term* that becomes arbitrarily large as $x \to \infty$. Hence, this approximation is valid only for $x \ll 1/\varepsilon$ and for small ε. If an approximation is desired over a larger range of x then the method of multiple scales is required.

Singular Perturbation

The *method of multiple scales* is a singular method that is *sometimes* useful if the regular perturbation method fails. In this case the assumption is made that the solution depends on *two* (or more) different length (or time) scales. By trying various possibilities, one can determine those scales. The scales are treated as dependent variables when transforming the given ordinary differential equation into a partial differential equation, but then the scales are treated as independent variables when solving the equations.

Example 19.7.2. Consider the equation $\varepsilon y'' + y' = 2$, $y(0) = 0$, $y(1) = 1$. This is singular since (with $\varepsilon = 0$) the resulting first-order equation cannot satisfy both boundary conditions. For the problem the proper length scales are $u = x$ and $v = x/\varepsilon$. The second scale can be ascertained by substituting $\varepsilon^n x$ for x and requiring $\varepsilon y''$ and y' to be of the same order in the transformed equation. Then

$$\frac{d}{dx} = \frac{\partial}{\partial u}\frac{du}{dx} + \frac{\partial}{\partial v}\frac{dv}{dx} = \frac{\partial}{\partial u} + \frac{1}{\varepsilon}\frac{\partial}{\partial v}$$

and the equation becomes

$$\varepsilon\left(\frac{\partial}{\partial u}+\frac{1}{\varepsilon}\frac{\partial}{\partial v}\right)^2 y+\left(\frac{\partial}{\partial u}+\frac{1}{\varepsilon}\frac{\partial}{\partial v}\right)y=2$$

With $y(x;\varepsilon)=y_0(u,v)+\varepsilon y_1(u,v)+\varepsilon^2 y_2(u,v)+\cdots$ we have terms

$$O\left(\varepsilon^{-1}\right):\frac{\partial^2 y_0}{\partial v^2}+\frac{\partial y_0}{\partial v}=0 \qquad \left(\text{actually ODEs with parameter } u\right)$$

$$O\left(\varepsilon^{0}\right):\frac{\partial^2 y_1}{\partial v^2}+\frac{\partial y_1}{\partial v}=2-2\frac{\partial^2 y_0}{\partial u\,\partial v}-\frac{\partial y_0}{\partial u}$$

$$O\left(\varepsilon^{1}\right):\frac{\partial^2 y_2}{\partial v^2}+\frac{\partial y_2}{\partial v}=-2\frac{\partial^2 y_1}{\partial u\,\partial v}-\frac{\partial y_1}{\partial u}-\frac{\partial^2 y_0}{\partial u^2}$$

$$\vdots$$

Then $y_0(u, v) = A(u) + B(u)e^{-v}$ and so the second equation becomes $\partial^2 y_1/\partial v^2 + \partial y_1/\partial v = 2 - A'(u) + B'(u)e^{-v}$, with the solution $y_1(u, v) = [2 - A'(u)]v + vB'(u)e^{-v} + D(u) + E(u)e^{-v}$. Here A, B, D and E are still arbitrary. Now the solvability condition — "higher order terms must vanish no slower (as $\varepsilon \to 0$) than the previous term" (Kevorkian and Cole, 1981) — is used. For y_1 to vanish no slower than y_0 we must have $2 - A'(u) = 0$ and $B'(u) = 0$. If this were not true the terms in y_1 would be larger than those in y_0 ($v \gg 1$). Thus $y_0(u, v) = (2u + A_0) + B_0e^{-v}$, or in the original variables $y(x; \varepsilon) \approx (2x + A_0) + B_0e^{-x/\varepsilon}$ and matching to both boundary conditions gives $y(x; \varepsilon) \approx 2x - (1 - e^{-x/\varepsilon})$.

Boundary Layer Method

The boundary layer method is applicable to regions in which the solution is *rapidly varying.* See the references at the end of the chapter for detailed discussion.

Iterative Methods

Taylor Series

If it is known that the solution of a differential equation has a power series in the independent variable (t), then we may proceed from the initial data (the easiest problem) to compute the Taylor series by differentiation.

Example 19.7.3. Consider the equation $(d^2x/dt) = -x - x^2$, $x(0) = 1$, $x'(0) = 1$. From the differential equation, $x''(0) = -2$, and, since $x''' = -x' - 2xx'$, $x'''(0) = -1 - 2 = -3$, so the four term approximation for $x(t) \approx 1 + t - (2t^2/2!) - (3t^3/3!) = 1 + t - t^2 - t^3/2$. An estimate for the error at $t = t_1$, (see a discussion of series methods in any calculus text) is not greater than $|d^4x/dt^4|_{\max}[(t_1)^4/4!]$, $0 \le t \le t_1$.

Picard's Method

If the vector differential equation $x' = f(t, x)$, $x(0)$ given, is to be approximated by Picard iteration, we begin with an initial guess $x_0 = x(0)$ and calculate iteratively $x_i' = f(t, x_{i-1})$.

Example 19.7.4. Consider the equation $x' = x + y^2$, $y' = y - x^3$, $x(0) = 1$, $y(0) = 2$. With $x_0 = 1$, $y_0 = 2$, $x_1' = 5$, $y_1' = 1$, so $x_1 = 5t + 1$, $y_1 = t + 2$, since $x_i(0) = 1$, $y_i(0) = 2$ for $i \ge 0$. To continue, use $x_{i+1}' = x_i + y_i^2$, $y_{i+1}' = y_i - x_i^3$. A modification is the utilization of the first calculated term immediately in the second equation. Thus, the calculated value of $x_1 = 5t + 1$, when used in the second equation, gives $y_1' = y_0 - (5t + 1)^3 = 2 - (125t^3 + 75t^2 + 15t + 1)$, so $y_1 = 2t - (125t^4/4) - 25t^3 - (15t^2/2) - t + 2$. Continue with the iteration $x_{i+1}' = x_i + y_i^3$, $y_{i+1}' = y_i - (x_{i+1})^3$.

Another variation would be $x_{i+1}' = x_{i+1} + (y_i)^3$, $y_{i+1}' = y_{i+1} - (x_{i+1})^3$.

References

Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Science and Engineering, Volume I.* Academic Press, Boston, MA.

Ames, W. F. 1968. *Nonlinear Ordinary Differential Equations in Transport Processes.* Academic Press, Boston, MA.

Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Science and Engineering, Volume II.* Academic Press, Boston, MA.

Kevorkian, J. and Cole, J. D. 1981. *Perturbation Methods in Applied Mathematics,* Springer, New York.

Miklin, S. G. and Smolitskiy, K. L. 1967. *Approximate Methods for Solutions of Differential and Integral Equations.* Elsevier, New York.

Nayfeh, A. H. 1973. *Perturbation Methods.* John Wiley & Sons, New York.

Zwillinger, D. 1992. *Handbook of Differential Equations,* 2nd ed. Academic Press, Boston, MA.

19.8 Integral Transforms

William F. Ames

All of the integral transforms are special cases of the equation $g(s) = \int_a^b K(s, t)f(t)d\,t$, in which $g(s)$ is said to be the *transform* of $f(t)$, and $K(s, t)$ is called the *kernel* of the transform. Table 19.8.1 shows the more important kernels and the corresponding intervals (a, b).

Details for the first three transforms listed in Table 19.8.1 are given here. The details for the other are found in the literature.

Laplace Transform

The Laplace transform of $f(t)$ is $g(s) = \int_0^\infty e^{-st} f(t)\,dt$. It may be thought of as transforming one class of functions into another. The advantage in the operation is that under certain circumstances it replaces complicated functions by simpler ones. The notation $L[f(t)] = g(s)$ is called the *direct transform* and $L^{-1}[g(s)] = f(t)$ is called the *inverse transform.* Both the direct and inverse transforms are tabulated for many often-occurring functions. In general $L^{-1}[g(s)] = (1/2\,\pi i)\int_{\alpha-i\infty}^{\alpha+i\infty} e^{st}g(s)\,ds$, and to evaluate this integral requires a knowledge of complex variables, the theory of residues, and contour integration.

Properties of the Laplace Transform

Let $L[f(t)] = g(s)$, $L^{-1}[g(s)] = f(t)$.

1. The Laplace transform may be applied to a function $f(t)$ if $f(t)$ is continuous or piecewise continuous; if $t^n|f(t)|$ is finite for all t, $t \to 0$, $n < 1$; and if $e^{-at}|f(t)|$ is finite as $t \to \infty$ for some value of a, $a > 0$.
2. L and L^{-1} are unique.
3. $L[af(t) + bh(t)] = aL[f(t)] + bL[h(t)]$ (linearity).
4. $L[e^{at}f(t)] = g(s - a)$ (shift theorem).
5. $L[(-t)^k f(t)] = d^k g/ds^k$; k a positive integer.

Example 19.8.1. $L[\sin a\,t] = \int_0^\infty e^{-st} \sin a\,t\,d\,t = a/(s^2 + a^2)$, $s > 0$. By property 5,

$$\int_0^\infty e^{-st} t \sin at\,dt = L[t\sin at] = \frac{2as}{s^2 + a^2}$$

Table 19.8.1 Kernels and Intervals of Various Integral Transforms

Name of Transform	(a, b)	$K(s, t)$
Laplace	$(0, \infty)$	e^{-st}
Fourier	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}} e^{-ist}$
Fourier cosine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \cos st$
Fourier sine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \sin st$
Mellin	$(0, \infty)$	t^{s-1}
Hankel	$(0, \infty)$	$tJ_\nu(st)$, $\nu \geq -\frac{1}{2}$

6.
$$L[f'(t)] = sL[f(t)] - f(0)$$
$$L[f''(t)] = s^2 L[f(t)] - sf(0) - f'(0)$$
$$\vdots$$
$$L\left[f^{(n)}(t)\right] = s^n L[f(t)] - s^{n-1} f(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0)$$

In this property it is apparent that the initial data are automatically brought into the computation.

Example 19.8.2. Solve $y'' + y = e^t$, $y(0) = 1$, $y'(0) = 1$. Now $L[y''] = s^2L[y] - sy(0) - y'(0) = s^2L[y] - s - 1$. Thus, using the linear property of the transform (property 3), $s^2L[y] + L[y] - s - 1 = L[e^t] = 1/(s - 1)$. Therefore, $L[y] = s^2/[(s - 1)(s^2 + 1)]$.

With the notations $\Gamma(n + 1) = \int_0^\infty x^n e^{-x}\, dx$ (gamma function) and $J_n(t)$ the Bessel function of the first kind of order n, a short table of Laplace transforms is given in Table 19.8.2.

7. $$L\left[\int_a^t f(t)\, dt\right] = \tfrac{1}{s} L[f(t)] + \tfrac{1}{s}\int_a^0 f(t)\, dt.$$

Example 19.8.3. Find $f(t)$ if $L[f(t)] = (1/s^2)[1/(s^2 - a^2)]$. $L[1/a \sinh a\ t] = 1/(s^2 - a^2)$. Therefore, $f(t) = \int_0^t [\int_0^t \frac{1}{a} \sinh a\ t\ d\ t]d\ t = 1/a^2[(\sinh a\ t)/a - t]$.

$$L\left[\frac{f(t)}{t}\right] = \int_s^\infty g(s)\, ds; \qquad L\left[\frac{f(t)}{t^k}\right] = \underbrace{\int_s^\infty \cdots \int_s^\infty}_{k \text{ integrals}} g(s)(ds)^k$$

Example 19.8.4. $L[(\sin a\ t)/t] = \int_s^\infty L[\sin a\ t]d\ s = \int_s^\infty [a\ d\ s/(s^2 + a^2)] = \cot^{-1}(s/a)$.

9. The *unit step function* $u(t - a) = 0$ for $t < a$ and 1 for $t > a$. $L[u(t - a) = e^{-as}/s$.
10. The *unit impulse function* is $\delta(a) = u'(t - a) = 1$ at $t = a$ and 0 elsewhere. $L[u'(t - a) = e^{-as}$.
11. $L^{-1}[e^{-as}g(s)] = f(t - a)u(t - a)$ (second shift theorem).
12. If $f(t)$ is *periodic* of period b — that is, $f(t + b) = f(t)$ — then $L[f(t)] = [1/(1 - e^{-bs})] \times \int_0^b e^{-st} f(t)\, dt$.

Example 19.8.5. The equation $\partial^2 y/(\partial t \partial x) + \partial y/\partial t + \partial y/\partial x = 0$ with $(\partial y/\partial x)(0, x) = y(0, x) = 0$ and $y(t, 0) + (\partial y/\partial t)(t, 0) = \delta(0)$ (see property 10) is solved by using the Laplace transform of y with respect to t. With $g(s, x) = \int_0^\infty e^{-st} y(t, x)\, dt$, the transformed equation becomes

Table 19.8.2 Some Laplace Transforms

$f(t)$	$g(s)$	$f(t)$	$g(s)$
1	$\frac{1}{s}$	$e^{-at}(1 - a\,t)$	$\frac{s}{(s+a)^2}$
t^n, n is a + integer	$\frac{n!}{s^{n+1}}$	$\frac{t\sin at}{2a}$	$\frac{s}{\left(s^2+a^2\right)^2}$
t^n, $n \neq$ a + integer	$\frac{\Gamma(n+1)}{s^{n+1}}$	$\frac{1}{2a^2}\sin at\sinh at$	$\frac{s}{s^4+4a^4}$
$\cos a\,t$	$\frac{s}{s^2+a^2}$	$\cos a\,t\cosh a\,t$	$\frac{s^3}{s^4+4a^4}$
$\sin a\,t$	$\frac{a}{s^2+a^2}$	$\frac{1}{2a}(\sinh at+\sin at)$	$\frac{s^2}{s^4-a^4}$
$\cosh a\,t$	$\frac{s}{s^2-a^2}$	$\frac{1}{2}(\cosh at+\cos at)$	$\frac{s^3}{s^4-a^4}$
$\sinh a\,t$	$\frac{a}{s^2-a^2}$	$\frac{\sin at}{t}$	$\tan^{-1}\frac{a}{s}$
e^{-at}	$\frac{1}{s+a}$	$J_0(a\,t)$	$\frac{1}{\sqrt{s^2+a^2}}$
$e^{-bt}\cos a\,t$	$\frac{s+b}{(s+b)^2+a^2}$	$\frac{n}{a^n}\frac{J_n(at)}{t}$	$\frac{1}{\left(\sqrt{s^2+a^2}+s\right)^n}$
$e^{-bt}\sin a\,t$	$\frac{a}{(s+b)^2+a^2}$	$J_0\left(2\sqrt{at}\right)$	$\frac{1}{s}e^{-a/s}$

$$s\frac{\partial g}{\partial x}-\frac{\partial y}{\partial x}(0,x)+sg-y(0,x)+\frac{\partial g}{\partial x}=0$$

or

$$(s+1)\frac{\partial g}{\partial x}+sg=\frac{\partial y}{\partial x}(0,x)+y(0,x)=0$$

The second (boundary) condition gives $g(s, 0) + sg(s, 0) – y(0, 0) = 1$ or $g(s, 0) = 1/(1 + s)$. A solution of the preceding ordinary differential equation consistent with this condition is $g(s, x) = [1/(s + 1)]e^{-sx/(s+1)}$. Inversion of this transform gives $y(t, x) = e^{-(t+x)}I_0\,(2/\sqrt{tx})$, where I_0 is the zero-order Bessel function of an imaginary argument.

Convolution Integral

The *convolution integral* (*faltung*) of two functions $f(t)$, $r(t)$ is $x(t) = f(t)*r(t) = \int_0^t f(\tau)r(t-\tau)\,d\,\tau$.

Example 19.8.6. $t * \sin t = \int_0^t \tau\sin(t-\tau)\,d\,\tau = t - \sin t$.

13. $L[f(t)]L[h(t)] = L[f(t) * h(t)]$.

Fourier Transform

The *Fourier transform* is given by $F[f(t)] = (1/\sqrt{2\pi})\int_{-\infty}^{\infty} f(t)e^{-ist}\,d\,t = g(s)$ and its *inverse* by $F^{-1}[g(s)] = (1/\sqrt{2\pi})\int_{-\infty}^{\infty} g(s)e^{ist}\,d\,t = f(t)$. In brief, the condition for the Fourier transform to exist is that $\int_{-\infty}^{\infty}|f(t)|\,d\,t < \infty$, although certain functions may have a Fourier transform even if this is violated.

Example 19.8.7. The function $f(t) = 1$ for $-a \le t \le a$ and $= 0$ elsewhere has

$$F\big[f(t)\big] = \int_{-a}^{a} e^{-ist}\,dt = \int_{0}^{a} e^{ist}\,dt + \int_{0}^{a} e^{-ist}\,dt = 2\int_{0}^{a} \cos st\,dt = \frac{2\sin sa}{s}$$

Properties of the Fourier Transform

Let $F[f(t)] = g(s)$; $F^{-1}[g(s)] = f(t)$.

1. $F[f^{(n)}(t)] = (i\,s)^n\,F[f(t)]$
2. $F[af(t) + bh(t)] = aF[f(t)] + bF[h(t)]$
3. $F[f(-t)] = g(-s)$
4. $F[f(at)] = 1/a\;g(s/a), \quad a > 0$
5. $F[e^{-iwt}f(t)] = g(s + w)$
6. $F[f(t + t_1)] = e^{ist_1}\,g(s)$
7. $F[f(t)] = G(i\,s) + G(-i\,s)$ if $f(t) = f(-t)$ ($f(t)$ even)
   $F[f(t)] = G(i\,s) - G(-i\,s)$ if $f(t) = -f(-t)$ (f odd)

where $G(s) = L[f(t)]$. This result allows the use of the Laplace transform tables to obtain the Fourier transforms.

Example 19.8.8. Find $F[e^{-a|t|}]$ by property 7. The term $e^{-a|t|}$ is even. So $L[e^{-at}] = 1/(s + a)$. Therefore, $F[e^{-a|t|}] = 1/(i\,s + a) + 1/(-i\,s + a) = 2a/(s^2 + a^2)$.

Fourier Cosine Transform

The *Fourier cosine transform* is given by $F_c[f(t)] = g(s) = \sqrt{(2/\pi)}\int_0^\infty f(t)\cos s\,t\,d\,t$ and its *inverse* by $F_c^{-1}[g(s)] = f(t) = \sqrt{(2/\pi)}\int_0^\infty g(s)\cos s\,t\,d\,s$. The *Fourier sine transform* F_s is obtainable by replacing the cosine by the sine in the above integrals.

Example 19.8.9. $F_c[f(t)]$, $f(t) = 1$ for $0 < t < a$ and 0 for $a < t < \infty$. $F_c\,[f(t)] = \sqrt{(2/\pi)}\int_0^a \cos s\,t\,d\,t = \sqrt{(2/\pi)}\,(\sin a\,s)/s$.

Properties of the Fourier Cosine Transform

$F_c[f(t)] = g(s)$.

1. $F_c[af(t) + bh(t)] = aF_c[f(t)] + bF_c[h(t)]$
2. $F_c[f(at)] = (1/a)\;g\;(s/a)$
3. $F_c[f(at)\cos bt] = 1/2a\;[g\;((s + b)/a) + g((s - b)/a)]$, $a, b > 0$
4. $F_c[t^{2n}f(t)] = (-1)^n(d^{2n}g)/(d\,s^{2n})$
5. $F_c[t^{2n+1}f(t)] = (-1)^n(d^{2n+1})/(d\,s^{2n+1})\;F_s[f(t)]$

Table 19.8.3 presents some Fourier cosine transforms.

Example 19.8.10. The temperature θ in the semiinfinite rod $0 \le x < \infty$ is determined by the differential equation $\partial\theta/\partial t = k(\partial^2\theta/\partial x^2)$ and the condition $\theta = 0$ when $t = 0$, $x \ge 0$; $\partial\theta/\partial x = -\mu$ = constant when $x = 0$, $t > 0$. By using the Fourier cosine transform, a solution may be found as $\theta(x, t) = (2\mu/\pi)\int_0^\infty (\cos px/p)\,(1 - e^{-kp2t})\;\;d\,p$.

References

Churchill, R. V. 1958. *Operational Mathematics.* McGraw-Hill, New York.

Ditkin, B. A. and Proodnikav, A. P. 1965. *Handbook of Operational Mathematics* (in Russian). Nauka, Moscow.

Doetsch, G. 1950–1956. *Handbuch der Laplace Transformation,* vols. I-IV (in German). Birkhauser, Basel.

Table 19.8.3 Fourier Cosine Transforms

$f(t)$	$\dfrac{g(s)}{\sqrt{2/\pi}}$
$\left.\begin{matrix} t & 0<t<1 \\ 2-t & 1<t<2 \\ 0 & 2<t<\infty \end{matrix}\right\}$	$\dfrac{1}{s^2}[2\cos s - 1 - \cos 2s]$
$t^{-1/2}$	$\pi^{1/2}(s)^{-1/2}$
$\left.\begin{matrix} 0 & 0<t<a \\ (t-a)^{-1/2} & a<t<\infty \end{matrix}\right\}$	$\pi^{1/2}(s)^{-1/2}\,[\cos a\,s - \sin a\,s]$
$(t^2 + a^2)^{-1}$	$\frac{1}{2}\pi a^{-1}e^{-as}$
$e^{-at}, \quad a>0$	$\dfrac{a}{s^2+a^2}$
$e^{-at^2}, \quad a>0$	$\frac{1}{2}\pi^{1/2}a^{-1/2}e^{-s^2/4a}$
$\dfrac{\sin at}{t} \quad a>0$	$\begin{cases} \pi/2 & s<a \\ \pi/4 & s=a \\ 0 & s>a \end{cases}$

Nixon, F. E. 1960. *Handbook of Laplace Transforms.* Prentice-Hall, Englewood Cliffs, NJ.
Sneddon, I. 1951. *Fourier Transforms.* McGraw-Hill, New York.
Widder, D. 1946. *The Laplace Transform*, Princeton University Press, Princeton, NJ.

Further Information

The references citing G. Doetsch, *Handbuch der Laplace Transformation*, vols. I-IV, Birkhauser, Basel, 1950–1956 (in German) *and* B. A. Ditkin and A. P. Prodnikav, *Handbook of Operational Mathematics*, Moscow, 1965 (in Russian) are the most extensive tables known. The latter reference is 485 pages.

19.9 Calculus of Variations

William F. Ames

The basic problem in the *calculus of variations* is to determine a function such that a certain *functional*, often an integral involving that function and certain of its derivatives, takes on *maximum or minimum values*. As an example, find the function $y(x)$ such that $y(x_1) = y_1$, $y(x_2) = y_2$ and the integral (functional) $I = 2\pi \int_{x_1}^{x_2} y[1 + y')^2]^{1/2}\, dx$ is a minimum. A second example concerns the transverse deformation $u(x, t)$ of a beam. The energy functional $I = \int_{t_1}^{t_2} \int_0^L [1/2\ \rho\ (\partial u/\partial t)^2 - 1/2\ EI\ (\partial^2 u/\partial x^2)^2 + fu]\, dx\, dt$ is to be minimized.

The Euler Equation

The elementary part of the theory is concerned with a *necessary* condition (generally in the form of a differential equation with boundary conditions) that the required function must satisfy. To show mathematically that the function obtained actually maximizes (or minimizes) the integral is much more difficult than the corresponding problems of the differential calculus.

The *simplest case* is to determine a function $y(x)$ that makes the integral $I = \int_{x_1}^{x_2} F(x, y, y')\, dx$ stationary and that satisfies the prescribed end conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. Here we suppose F has continuous second partial derivatives with respect to x, y, and $y' = dy/dx$. If $y(x)$ is such a function, then it must satisfy the *Euler equation* $(d/dx)(\partial F/\partial y') - (\partial F/\partial y) = 0$, which is the required necessary condition. The indicated partial derivatives have been formed by treating x, y, and y' as independent variables. Expanding the equation, the equivalent form $F_{y'y'}y'' + F_{y'y}y' + (F_{y'x} - F_y) = 0$ is found. This is second order in y unless $F_{y'y'} = (\partial^2 F)/[(\partial y')^2] = 0$. An alternative form $1/y'[d/dx(F - (\partial F/\partial y')(dy/dx)) - (\partial F/\partial x)] = 0$ is useful. Clearly, if F does not involve x explicitly $[(\partial F/\partial x) = 0]$ a first integral of Euler's equation is $F - y'(\partial F/\partial y') = c$. If F does not involve y explicitly $[(\partial F/\partial y) = 0]$ a first integral is $(\partial F/\partial y') = c$.

The Euler equation for $I = 2\pi \int_{x_1}^{x_2} y[1 + (y')^2]^{1/2}\, dx$, $y(x_1) = y_1$, $y(x_2) = y_2$ is $(d/dx)[yy'/[1 + (y')^2]^{1/2}] - [1 + (y')^2]^{1/2} = 0$ or after reduction $yy'' - (y')^2 - 1 = 0$. The solution is $y = c_1 \cosh(x/c_1 + c_2)$, where c_1 and c_2 are integration constants. Thus the required minimal surface, if it exists, must be obtained by revolving a catenary. Can c_1 and c_2 be chosen so that the solution passes through the assigned points? The answer is found in the solution of a transcendental equation that has two, one, or no solutions, depending on the prescribed values of y_1 and y_2.

The Variation

If $F = F(x, y, y')$, with x independent and $y = y(x)$, then the *first variation* δF of F is defined to be $\delta F = (\partial F/\partial x)\, \delta y + (\partial F/\partial y)\, \delta y'$ and $\delta y' = \delta\,(dy/dx) = (d/dx)\,(\delta y)$ — that is, they commute. Note that the first variation, δF, of a functional is a first-order change from curve to curve, whereas the differential of a function is a first-order approximation to the change in that function along a *particular curve*. The laws of δ are as follows: $\delta(c_1F + c_2G) = c_1\delta F + c_2\delta G$; $\delta(FG) = F\delta G + G\delta F$; $\delta(F/G) = (G\delta F - F\delta G)/G^2$; if x is an independent variable, $\delta x = 0$; if $u = u(x, y)$; $(\partial/\partial x)(\delta u) = \delta(\partial u/\partial x)$, $(\partial/\partial y)\ (\delta u) = \delta(\partial u/\partial y)$.

A necessary condition that the integral $I = \int_{x_1}^{x_2} F(x, y, y')\, dx$ be stationary is that its (first) variation vanish — that is, $\delta I = \delta \int_{x_1}^{x_2} F(x, y, y')\, dx = 0$. Carrying out the variation and integrating by parts yields of $\delta I = \int_{x_1}^{x_2} [(\partial F/\partial y) - (d/dx)(\partial F/\partial y')]\, \delta y\, dx + [(\partial F/\partial y')\, \partial y]_{x_1}^{x_2} = 0$. The arbitrary nature of δy means the square bracket must vanish and the last term constitutes the *natural boundary conditions.*

Example. The *Euler equation* of $\int_{x_1}^{x_2} F(x, y, y', y'')\, dx$ is $(d^2/dx^2)(\partial F/\partial y'') - (d/dx)(\partial F/\partial y') + (\partial F/\partial y) = 0$, with natural boundary conditions $\{[(d/dx)(\partial F/\partial y'') - (\partial F/\partial y')]\, \delta y\}_{x_1}^{x_2} = 0$ and $(\partial F/\partial y'')\, \delta y'|_{x_1}^{x_2} = 0$. The Euler equation of $\int_{x_1}^{x_2} \int_{y_1}^{y_2} F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy})\, dx\, dy$ is $(\partial^2/\partial x^2)(\partial F/\partial u_{xx}) + (\partial^2/\partial x \partial y)(\partial F/u_{xy}) + (\partial^2/\partial y^2)(\partial F/\partial u_{yy}) - (\partial/\partial x)(\partial F/\partial u_x) - (\partial/\partial y)(\partial F/\partial u_y) + (\partial F/\partial u)$, and the natural boundary conditions are

$$\left[\left(\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_{xx}}\right)+\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_{xy}}\right)-\frac{\partial F}{\partial u_x}\right)\delta u\right]_{x_1}^{x_2}=0, \qquad \left[\left(\frac{\partial F}{\partial u_{xx}}\right)\delta u_x\right]_{x_1}^{x_2}=0$$

$$\left[\left(\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_{yy}}\right)+\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_{xy}}\right)-\frac{\partial F}{\partial u_y}\right)\delta u\right]_{y_1}^{y_2}=0, \qquad \left[\left(\frac{\partial F}{\partial u_{yy}}\right)\delta u_y\right]_{y_1}^{y_2}=0$$

In the more general case of $I = \iint_R F(x, y, u, v, u_x, u_y, v_x, v_y)\, dx\, dy$, the condition $\delta I = 0$ gives rise to the two Euler equations $(\partial/\partial x)(\partial F/\partial u_x) + (\partial/\partial y)(\partial F/\partial u_y) - (\partial F/\partial u) = 0$ and $(\partial/\partial x)(\partial F/\partial v_x) + (\partial/\partial y)(\partial F/\partial v_y) - (\partial F/\partial v) = 0$. These are two PDEs in u and v that are linear or quasi-linear in u and v. The Euler equation for $I = \iiint_R (u_x^2 + u_y^2 + u_z^2)\, dx\, dy\, dz$, from $\delta I = 0$, is Laplace's equation $u_{xx} + u_{yy} + u_{zz} = 0$.

Variational problems are easily derived from the differential equation and associated boundary conditions by multiplying by the variation and integrating the appropriate number of times. To illustrate, let $F(x)$, $\rho(x)$, $p(x)$, and w be the tension, the linear mass density, the natural load, and (constant) angular velocity of a rotating string of length L. The equation of motion is $(d/dx)[F\,(dy/dx)] + \rho w^2 y + p = 0$. To formulate a corresponding variational problem, multiply all terms by a variation δy and integrate over $(0, L)$ to obtain

$$\int_0^L \frac{d}{dx}\left(F\frac{dy}{dx}\right)\delta y\, dx + \int_0^L \rho w^2 y \delta y\, dx + \int_0^L p\,\delta y\, dx = 0$$

The second and third integrals are the variations of $1/2\ \rho w^2 y^2$ and py, respectively. To treat the first integral, integrate by parts to obtain

$$\left[F\frac{dy}{dx}\delta y\right]_0^L - \int_0^L F\frac{dy}{dx}\delta\frac{dy}{dx}\,dx = \left[F\frac{dy}{dx}\delta y\right]_0^L - \int_0^L \frac{1}{2}F\delta\left(\frac{dy}{dx}\right)^2 dx = 0$$

So the variation formulation is

$$\delta\int_0^L\left[\frac{1}{2}\rho w^2 y^2 + py - \frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right]dx + \left[F\frac{dy}{dx}\delta y\right]_0^L = 0$$

The last term represents the *natural boundary conditions*. The term $1/2\ \rho w^2 y^2$ is the kinetic energy per unit length, the term $-py$ is the potential energy per unit length due to the radial force $p(x)$, and the term $1/2\ F(dy/dx)^2$ is a first approximation to the potential energy per unit length due to the tension $F(x)$ in the string. Thus the integral is often called the *energy integral*.

Constraints

The variations in some cases cannot be arbitrarily assigned because of one or more auxiliary conditions that are usually called *constraints*. A typical case is the functional $\int_{x_1}^{x_2} F(x, u, v, u_x, v_x)\, dx$ with a constraint $\phi(u, v) = 0$ relating u and v. If the variations of u and v (δu and δv) vanish at the end points, then the variation of the integral becomes

$$\int_{x_1}^{x_2}\left\{\left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)\right]\delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right)\right]\delta v\right\}dx = 0$$

The variation of the constraint $\phi(u, v) = 0$, $\phi_u \delta u + \phi_v \delta v = 0$ means that the variations cannot both be assigned arbitrarily inside (x_1, x_2), so their coefficients need not vanish separately. Multiply $\phi_u \delta u + \phi_v \delta v = 0$ by a Lagrange multiplier λ (may be a function of x) and integrate to find $\int_{x_1}^{x_2} (\lambda\phi_u \delta u + \lambda\phi_v \delta v)\, dx = 0$. Adding this to the previous result yields

$$\int_{x_1}^{x_2} \left\{ \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x} \right) + \lambda\phi_u \right] \delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x} \right) + \lambda\phi_v \right] \delta v \right\} dx = 0$$

which must hold for any λ. Assign λ so the first square bracket vanishes. Then δv can be assigned to vanish inside (x_1, x_2) so the two systems

$$\frac{d}{dx}\left[\frac{\partial F}{\partial u_x} \right] - \frac{\partial F}{\partial u} - \lambda\phi_u = 0, \qquad \frac{d}{dx}\left[\frac{\partial F}{\partial v_x} \right] - \frac{\partial F}{\partial v} - \lambda\phi_v = 0$$

plus the constraint $\phi(u, v) = 0$ are three equations for u, v and λ.

References

Gelfand, I. M. and Fomin, S. V. 1963. *Calculus of Variations*. Prentice Hall, Englewood Cliffs, NJ.

Lanczos, C. 1949. *The Variational Principles of Mechanics.* Univ. of Toronto Press, Toronto.

Schechter, R. S. 1967. *The Variational Method in Engineering,* McGraw-Hill, New York.

Vujanovic, B. D. and Jones, S. E. 1989. *Variational Methods in Nonconservative Phenomena.* Academic Press, New York.

Weinstock, R. 1952. *Calculus of Variations, with Applications to Physics and Engineering.* McGraw-Hill, New York.

19.10 Optimization Methods

George Cain

Linear Programming

Let **A** be an $m \times n$ matrix, **b** a column vector with m components, and **c** a column vector with n components. Suppose $m < n$, and assume the rank of **A** is m. The standard linear programming problem is to find, among all nonnegative solutions of $\mathbf{Ax} = \mathbf{b}$, one that minimizes

$$\mathbf{c}^T\mathbf{x} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

This problem is called a *linear* program. Each solution of the system $\mathbf{Ax} = \mathbf{b}$ is called a *feasible* solution, and the *feasible set* is the collection of all *feasible solutions.* The function $\mathbf{c}^T\mathbf{x} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is the cost function, or the objective function. A solution to the linear program is called an *optimal feasible solution.*

Let **B** be an $m \times n$ submatrix of **A** made up of m linearly independent columns of **A**, and let **C** be the $m \times (n - m)$ matrix made up of the remaining columns of **A**. Let $\mathbf{x_B}$ be the vector consisting of the components of **x** corresponding to the columns of **A** that make up **B**, and let $\mathbf{x_C}$ be the vector of the remaining components of **x**, that is, the components of **x** that correspond to the columns of **C**. Then the equation $\mathbf{Ax} = \mathbf{b}$ may be written $\mathbf{Bx_B} + \mathbf{Cx_C} = \mathbf{b}$. A solution of $\mathbf{Bx_B} = \mathbf{b}$ together with $\mathbf{x_C} = \mathbf{0}$ gives a solution **x** of the system $\mathbf{Ax} = \mathbf{b}$. Such a solution is called a *basic solution*, and if it is, in addition, nonnegative, it is a *basic feasible solution.* If it is also optimal, it is an *optimal basic feasible solution.* The components of a basic solution are called *basic variables.*

The Fundamental Theorem of Linear Programming says that if there is a feasible solution, there is a basic feasible solution, and if there is an optimal feasible solution, there is an optimal basic feasible solution. The linear programming problem is thus reduced to searching among the set of basic solutions for an optimal solution. This set is, of course, finite, containing as many as $n!/[m!(n - m)!]$ points. In practice, this will be a very large number, making it imperative that one use some efficient search procedure in seeking an optimal solution. The most important of such procedures is the *simplex method,* details of which may be found in the references.

The problem of finding a solution of $\mathbf{Ax} \le \mathbf{b}$ that minimizes $\mathbf{c}^T\mathbf{x}$ can be reduced to the standard problem by appending to the vector **x** an additional m nonnegative components, called *slack variables*. The vector **x** is replaced by **z**, where $\mathbf{z}^T = [x_1,x_2\ldots x_n\ s_1,s_2\ldots x_m]$, and the matrix **A** is replaced by $\mathbf{B} = [\mathbf{A}\ \mathbf{I}]$, where **I** is the $m \times m$ identity matrix. The equation $\mathbf{Ax} = \mathbf{b}$ is thus replaced by $\mathbf{Bz} = \mathbf{Ax} + s = \mathbf{b}$, where $s^T = [s_1,s_2,\ldots,s_m]$. Similarly, if inequalities are reversed so that we have $\mathbf{Ax} \le \mathbf{b}$, we simply append $-s$ to the vector x. In this case, the additional variables are called *surplus variables.*

Associated with every linear programming problem is a corresponding dual problem. If the *primal* problem is to minimize $\mathbf{c}^T\mathbf{x}$ subject to $\mathbf{Ax} \ge \mathbf{b}$, and $\mathbf{x} \ge 0$, the corresponding *dual* problem is to maximize $\mathbf{y}^T\mathbf{b}$ subject to $\mathbf{t}^T\mathbf{A} \le \mathbf{c}^T$. If either the primal problem or the dual problem has an optimal solution, so also does the other. Moreover, if $\mathbf{x_p}$ is an optimal solution for the primal problem and $\mathbf{y_d}$ is an optimal solution for the corresponding dual problem $\mathbf{c}^T\mathbf{x_p} = \mathbf{y_d}^T\mathbf{b}$.

Unconstrained Nonlinear Programming

The problem of minimizing or maximizing a sufficiently smooth nonlinear function $f(x)$ of n variables, $\mathbf{x}^T = [x_1,x_2\ldots x_n]$, with no restrictions on x is essentially an ordinary problem in calculus. At a minimizer or maximizer x^*, it must be true that the gradient of f vanishes:

$$\nabla f\left(\mathbf{x}^*\right) = 0$$

Thus x^* will be in the set of all solutions of this system of n generally nonlinear equations. The solution of the system can be, of course, a nontrivial undertaking. There are many recipes for solving systems of nonlinear equations. A method specifically designed for minimizing f is the *method of steepest descent*. It is an old and honorable algorithm, and the one on which most other more complicated algorithms for unconstrained optimization are based. The method is based on the fact that at any point x, the direction of maximum decrease of f is in the direction of $-\nabla f(x)$. The algorithm searches in this direction for a minimum, recomputes $\nabla f(x)$ at this point, and continues iteratively. Explicitly:

1. Choose an initial point x_0.
2. Assume x_k has been computed; then compute $y_k = \nabla f(x_k)$, and let $t_k \geq 0$ be a local minimum of $g(t) = f(x_k - ty_k)$. Then $x_{k+1} = x_k - t_k y_k$.
3. Replace k by $k + 1$, and repeat step 2 until t_k is small enough.

Under reasonably general conditions, the sequence (x_k) converges to a minimum of f.

Constrained Nonlinear Programming

The problem of finding the maximum or minimum of a function $f(x)$ of n variables, subject to the constraints

$$\mathbf{a}(\mathbf{x}) = \begin{bmatrix} a_1(x_1, x_2, \ldots, x_n) \\ a_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ a_m(x_1, x_2, \ldots, x_n) \end{bmatrix} = \begin{bmatrix} b_1 \\ b \\ \vdots \\ b_m \end{bmatrix} = \mathbf{b}$$

is made into an unconstrained problem by introducing the new function $L(x)$:

$$L(\mathbf{x}) = f(\mathbf{x}) + \mathbf{z}^{\mathrm{T}}\mathbf{a}(\mathbf{x})$$

where $\mathbf{z}^{\mathrm{T}} = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ is the vector of *Lagrange multipliers*. Now the requirement that $\nabla L(x) = 0$, together with the constraints $a(x) = \mathbf{b}$, give a system of $n + m$ equations

$$\nabla f(\mathbf{x}) + \mathbf{z}^{\mathrm{T}}\nabla\mathbf{a}(\mathbf{x}) = 0$$

$$\mathbf{a}(\mathbf{x}) = \mathbf{b}$$

for the $n + m$ unknowns $x_1, x_2 \ldots, x_n, \lambda_1\lambda_2 \ldots, \lambda_m$ that must be satisfied by the minimizer (or maximizer) x.

The problem of inequality constraints is significantly more complicated in the nonlinear case than in the linear case. Consider the problem of minimizing $f(x)$ subject to m equality constraints $\mathbf{a}(\mathbf{x}) = \mathbf{b}$, and p inequality constraints $c(x) \leq \mathbf{d}$ [thus $\mathbf{a}(\mathbf{x})$ and $\mathbf{b}$ are vectors of m components, and $c(x)$ and $\mathbf{d}$ are vectors of p components.] A point x^* that satisfies the constraints is a *regular point* if the collection

$$\left\{\nabla a_1(\mathbf{x}^*), \nabla a_2(\mathbf{x}^*), \ldots, \nabla a_m(\mathbf{x}^*)\right\} \cup \left\{\nabla c_j(\mathbf{x}^*) : j \in J\right\}$$

where

$$J = \left\{j : c_j(\mathbf{x}^*) = d_j\right\}$$

is linearly independent. If x^* is a local minimum for the constrained problem and if it is a regular point, there is a vector $\boldsymbol{z}$ with m components and a vector $\mathbf{w} \geq \mathbf{0}$ with p components such that

$$\nabla f\left(\mathbf{x}^*\right) + \mathbf{z}^{\mathrm{T}} \nabla \mathbf{a}\left(\mathbf{x}^*\right) + \mathbf{w}^{\mathrm{T}} \nabla \mathbf{c}\left(\mathbf{x}^*\right) = 0$$

$$\mathbf{w}^{\mathrm{T}}\left(\mathbf{c}\left(\mathbf{x}^*\right) - \mathbf{d}\right) = 0$$

These are the *Kuhn-Tucker conditions*. Note that in order to solve these equations, one needs to know for which j it is true that $c_j(x^*) = 0$. (Such a constraint is said to be *active*.)

References

Luenberger, D. C. 1984. *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, MA.

Peressini, A. L. Sullivan, F. E., and Uhl, J. J., Jr. 1988. *The Mathematics of Nonlinear Programming*. Springer-Verlag, New York.

19.11 Engineering Statistics

Y. L. Tong

Introduction

In most engineering experiments, the outcomes (and hence the observed data) appear in a random and on deterministic fashion. For example, the operating time of a system before failure, the tensile strength of a certain type of material, and the number of defective items in a batch of items produced are all subject to random variations from one experiment to another. In engineering statistics, we apply the theory and methods of statistics to develop procedures for summarizing the data and making statistical inferences, thus obtaining useful information with the presence of randomness and uncertainty.

Elementary Probability

Random Variables and Probability Distributions

Intuitively speaking, a random variable (denoted by X, Y, Z, etc.) takes a numerical value that depends on the outcome of the experiment. Since the outcome of an experiment is subject to random variation, the resulting numerical value is also random. In order to provide a stochastic model for describing the probability distribution of a random variable X, we generally classify random variables into two groups: the discrete type and the continuous type. The discrete random variables are those which, technically speaking, take a finite number or a countably infinite number of possible numerical values. (In most engineering applications they take nonnegative integer values.) Continuous random variables involve outcome variables such as time, length or distance, area, and volume. We specify a function $f(x)$, called the probability density function (p.d.f.) of a random variable X, such that the random variable X takes a value in a set A (or real numbers) as given by

$$P[X \in A] = \begin{cases} \sum_{x \in A} f(x) & \text{for all sets } A \text{ if } X \text{ is discrete} \\ \int_A f(x)\,dx & \text{for all intervals } A \text{ if } X \text{ is continuous} \end{cases} \tag{9.11.1}$$

By letting A be the set of all values that are less than or equal to a fixed number t, i.e., $A = (-\infty, t)$, the probability function $P[X \le t]$, denoted by $F(t)$, is called the distribution function of X. We note that, by calculus, if X is a continuous random variable and if $F(x)$ is differentiable, then $f(x) = \frac{d}{dx} F(x)$.

Expectations

In many applications the "payoff" or "reward" of an experiment with a numerical outcome X is a specific function of X ($u(X)$, say). Since X is a random variable, $u(X)$ is also a random variable. We define the expected value of $u(X)$ by

$$Eu(X) = \begin{cases} \sum_x u(x) f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} u(x) f(x)\,dx & \text{if } X \text{ is continuous} \end{cases} \tag{9.11.12}$$

provided of course, that, the sum or the integral exists. In particular, if $u(x) = x$, the $EX \equiv \mu$ is called the mean of X (of the distribution) and $E(X - \mu)^2 \equiv \sigma^2$ is called the variance of X (of the distribution). The mean is a measurement of the central tendency, and the variance is a measurement of the dispersion of the distribution.

Some Commonly Used Distributions

Many well-known distributions are useful in engineering statistics. Among the discrete distributions, the hypergeometric and binomial distributions have applications in acceptance sampling problems and quality control, and the Poisson distribution is useful for studying queuing theory and other related problems. Among the continuous distributions, the uniform distribution concerns random numbers and can be applied in simulation studies, the exponential and gamma distributions are closely related to the Poisson distribution, and they, together with the Weibull distribution, have important applications in life testing and reliability studies. All of these distributions involve some unknown parameter(s), hence their means and variances also depend on the parameter(s). The reader is referred to textbooks in this area for details. For example, Hahn and Shapiro (1967, pp. 163–169 and pp. 120–134) contains a comprehensive listing of these and other distributions on their p.d.f.'s and the graphs, parameter(s), means, variances, with discussions and examples of their applications.

The Normal Distribution

Perhaps *the* most important distribution in statistics and probability is the normal distribution (also known as the Gaussian distribution). This distribution involves two parameters: μ and σ^2, and its p.d.f. is given by

$$f(x) = f\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{9.11.3}$$

for $-\infty < \mu < \infty$, $\sigma^2 > 0$, and $-\infty < \chi < \infty$. It can be shown analytically that, for a p.d.f. of this form, the values of μ and σ^2 are, respectively, that of the mean and the variance of the distribution. Further, the quantity, $\sigma = \sqrt{\sigma^2}$ is called the standard deviation of the distribution. We shall use the symbol $X \sim N(\mu, \sigma^2)$ to denote that X has a normal distribution with mean μ and variance σ^2. When plottting the p.d.f. $f(x; \mu, \sigma^2)$ given in Equation (19.11.3) we see that the resulting graph represents a bell-shaped curve symmetric about μ, as shown in Figure 19.11.1.

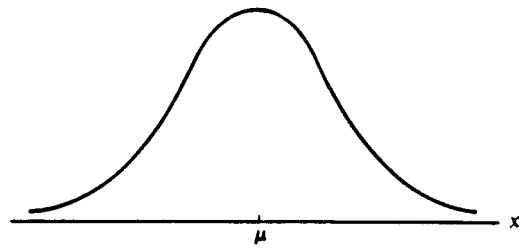

FIGURE 19.11.1 The normal curve with mean μ and variance σ^2.

If a random variable Z has an $N(0,1)$ distribution, then the p.d.f. of Z is given by (from Equation (19.11.3))

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty < z < \infty \tag{9.11.4}$$

The distribution function of Z,

$$\Phi(z) = \int_{-\infty}^{x} \phi(u)\, du \qquad -\infty < z < \infty \tag{9.11.5}$$

cannot be given in a closed form, hence it has been tabulated. The table of $\Phi(z)$ can be found in most textbooks in statistics and probability, including those listed in the references at the end of this section. (We note in passing that, by the symmetry property, $\Phi(z) + \Phi(-z) = 1$ holds for all z.)

Random Sample and Sampling Distributions

Random Sample and Related Statistics

As noted in Box et al., (1978), the design and analysis of engineering experiments usually involves the following steps:

1. The choice of a suitable stochastic model by assuming that the observations follow a certain distribution. The functional form of the distribution (or the p.d.f.) is assumed to be known, except the value(s) of the parameters(s).
2. Design of experiments and collection of data.
3. Summarization of data and computation of certain statistics.
4. Statistical inference (including the estimation of the parameters of the underlying distribution and the hypothesis-testing problems).

In order to make statistical inference concerning the parameter(s) of a distribution, it is essential to first study the sampling distributions. We say that $X_1, X_2, \ldots, X_n$ represent a random sample of size n if they are independent random variables and each of them has the same p.d.f., $f(x)$. (Due to space limitations, the notion of independence will not be carefully discussed here. Nevertheless, we say that $X_1, X_2, \ldots, X_n$ are independent if

$$P\left[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\right] = \prod_{i=1}^{n} P\left[X_i \in A_i\right] \tag{19.11.6}$$

holds for all sets $A_1, A_2, \ldots, A_n$.) Since the parameter(s) of the population is (are) unknown, the population mean μ and the population variance σ^2 are unknown. In most commonly used distributions μ and σ^2 can be estimated by the sample mean $\bar{X}$ and the sample variance S^2, respectively, which are given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] \tag{19.11.7}$$

(The second equality in the formula for S^2 can be verified algebraically.) Now, since $X_1, X_2, \ldots, X_n$ are random variables $\bar{X}$ and S^2 are also random variables. Each of them is called a statistic and has a probability distribution which also involves the unknown parameter(s). In probability theory there are two fundamental results concerning their distributional properties.

Theorem 1. (Weak Law of Large Numbers). As the sample size n becomes large, $\bar{X}$ converges to μ in probability and S^2 converges to σ^2 in probability. More precisely, for every fixed positive number $\varepsilon > 0$ we have

$$P\left[\left|\bar{X} - \mu\right| \le \varepsilon\right] \to 1, \qquad P\left[\left|S^2 - \sigma^2\right| \le \varepsilon\right] \to 1 \tag{19.11.8}$$

as $n \to \infty$.

Theorem 2. (Central Limit Theorem). As n becomes large, the distribution of the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma\ \sqrt{n}} = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \tag{19.11.9}$$

has approximately an $N(0,1)$ distribution. More precisely,

$$P[Z \le z] \to \Phi(z) \text{ for every fixed } z \text{ as } n \to \infty \tag{19.11.10}$$

Normal Distribution-Related Sampling Distributions

One-Sample Case

Additional results exist when the observations come from a normal population. If $X_1, X_2, \ldots, X_n$ represent a random sample of size n from an $N(\mu,\sigma^2)$ population, then the following sample distributions are useful:

Fact 3. For every fixed n the distribution of Z given in Equation (19.11.9) has *exactly* an N(0,1) distribution.

Fact 4. The distribution of the statistic $T = \sqrt{n}(\bar{X} - \mu)/S$, where $S = \sqrt{S^2}$ is the sample standard deviation, is called a *Student's t distribution* with $\nu = n - 1$ degrees of freedom, in symbols, $t(n - 1)$.

This distribution is useful for making inference on μ when σ^2 is unknown; a table of the percentiles can be found in most statistics textbooks.

Fact 5. The distribution of the statistic $W = (n - 1)S^2/\sigma^2$ is called a *chi-squared distribution* with $\nu = n - 1$ degrees of freedom, in symbols $\chi^2(\nu)$.
Such a distribution is useful in making inference on σ^2; a table of the percentiles can also be found in most statistics books.

Two-Sample Case

In certain applications we may be interested in the comparisons of two different treatments. Suppose that independent samples from treatments T_1 and T_2 are to be observed as shown in Table 19.11.1.

TABLE 19.11.1 Summarization of Data for a Two-Sample Problem

Treatment	Observations	Distribution	Sample Size	Sample Mean	Sample Variance
T_1	$X_{11}, X_{12}, \ldots, X_{1n_1}$	$N(\mu_1, \sigma_1^2)$	n_1	$\bar{X}_1$	S_1^2
T_2	$X_{21}, X_{22}, \ldots, X_{2n_2}$	$N(\mu_2, \sigma_2^2)$	n_2	$\bar{X}_2$	S_2^2

The difference of the population means $(\mu_1 - \mu_2)$ and the ratio of the population variances can be estimated, respectively, by $(\bar{X}_1 - \bar{X}_2)$ and S_1^2/S_2^2. The following facts summarize the distributions of these statistics:

Fact 6. Under the assumption of normality, $(\bar{X}_1 - \bar{X}_2)$ has an $N(\mu_1 - \mu_2, (\sigma_1^2/n_1) + (\sigma_2^2/n_2))$ distribution; or equivalently, for all n_1, n_2 the statistic.

$$Z = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^{1/2} \tag{19.11.11}$$

has an $N(0,1)$ distribution.

Fact 7. When $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, the common population variance is estimated by

$$S_p^2 = \left(n_1 + n_2 - 2\right)^{-1}\left[\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2\right] \tag{19.11.12}$$

and $(n_1 + n_2 - 2)S_p^2/\sigma^2$ has a $\chi^2(n_1 + n_2 - 2)$ distribution.

Fact 8. When $\sigma_1^2 = \sigma_2^2$, the statistic

$$T = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.13}$$

has a $t(n_1 + n_2 - 2)$ distribution, where $S_p = \sqrt{S_p^2}$.

Fact 9. The distribution of $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ is called an F distribution with degrees of freedom $(n_1 - 1, n_2 - 1)$, in symbols, $F(n_1 - 1, n_2 - 1)$,

The percentiles of this distribution have also been tabulated and can be found in statistics books.

In the following two examples we illustrate numerically how to find probabilities and percentiles using the existing tables for the normal, Student's *t*, chi-squared, and *F* distributions.

Example 10. Suppose that in an experiment four observations are taken, and that the population is assumed to have a normal distribution with mean μ and variance σ^2. Let $\bar{X}$ and S^2 be the sample mean and sample variance as given in Equation (19.11.7).

(a) If, based on certain similar experiments conducted in the past, we know that $\sigma^2 = 1.8^2 \times 10^{-6}$ ($\sigma = 1.8 \times 10^{-3}$), then from $\Phi(-1.645) = 0.05$ and $\Phi(1.96) = 0.975$ we have

$$P\left[-1.645 \le \frac{\bar{X} - \mu}{1.8 \times 10^{-3}\ \sqrt{4}} \le 1.96\right] = 0.975 - 0.05 = 0.925$$

or equivalently,

$$P\left[-1.645 \times 0.9 \times 10^{-3} \le \bar{X} - \mu \le 1.96 \times 0.9 \times 10^{-3}\right] = 0.925$$

(b) The statistic $T = 2(\bar{X} - \mu)/S$ has a Student's *t* distribution with 3 degrees of freedom (in symbols, $t(3)$). From the *t* table we have

$$P\left[-3.182 \le 2\left(\bar{X} - \mu\right)/S \le 3.182\right] = 0.95$$

which yields

$$P\left[-3.182 \times \frac{S}{2} \le \bar{X} - \mu \le 3.182 \times \frac{S}{2}\right] = 0.95$$

or equivalently,

$$P\left[\bar{X} - 3.182 \times \frac{S}{2} \le \mu \le \bar{X} + 3.182 \times \frac{S}{2}\right] = 0.95$$

This is, in fact, the basis for obtaining the confidence interval for μ given in Equation (19.11.17) when σ^2 is unknown.

(c) The statistic $3S^2/\sigma^2$ has a chi-squared distribution with 3 degrees of freedom (in symbols, $\chi^2(3)$). Thus from the chi-squared table we have $P[0.216 \le 3S^2/\sigma^2 \le 9.348] = 0.95$, which yields

$$P\left[\frac{3S^2}{9.348} \le \sigma^2 \le \frac{3S^2}{0.216}\right] = 0.95$$

and it forms the basis for obtaining a confidence interval for σ^2 as given in Equation (19.11.18).

Example 11. Suppose that in Table 19.11.1 (with two treatments) we have $n_1 = 4$ and $n_2 = 5$, and we let $\bar{X}_1, \bar{X}_2$ and S_1^2, S_2^2 denote the corresponding sample means and sample variances, respectively.

(a) Assume that $\sigma_1^2 = \sigma_2^2$ where the common variance is unknown and is estimated by S_p^2 given in Equation (19.11.12). Then the statistic

$$T = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2}$$

has a $t(7)$ distribution. Thus from the t table we have

$$P = \left[-2.998 \le \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2} \le 2.998\right] = 0.98$$

which is equivalent to saying that

$$P\left[-2.998 S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + 2.998 S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2}\right] = 0.98$$

(b) The statistic $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ has an $F(3,4)$ distribution. Thus from the F-table we have

$$P\left[\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{S_1^2}{S_2^2}\right) \le 6.59\right] = 0.95$$

or equivalently,

$$P\left[\frac{\sigma_2^2}{\sigma_1^2} \le 6.59 \frac{S_2^2}{S_1^2}\right] = 0.95$$

The distributions listed above (normal, Student's t, chi-squared, and F) form an integral part of the classical statistical inference theory, and they are developed under the assumption that the observations follow a normal distribution. When the distribution of the population is not normal and inference on the populations means is to be made, we conclude that (1) if the sample sizes n_1, n_2 are large, then the statistic Z in Equation (19.11.11) has an approximate $N(0,1)$ distribution and (2) in the small-sample case, the exact distribution of $\bar{X}$ (of $(\bar{X}_1 - \bar{X}_2)$)) depends on the population p.d.f. There are several analytical methods for obtaining it, and those methods can be found in statistics textbooks.

Confidence Intervals

A method for estimating the population parameters based on the sample mean(s) and sample variance(s) involves the confidence intervals for the parameters.

One-Sample Case

1. Confidence Interval for μ When σ^2 is Known. Consider the situation in which a random sample of size n is taken from an $N(\mu,\sigma^2)$ population and σ^2 is known. An interval, I_1, of the form $I_1 = (\bar{X} - d, \bar{X} + d)$ (with width $2d$) is to be constructed as a "confidence interval or μ." If we make the assertion that μ is in this interval (i.e., μ is bounded below by $\bar{X} - d$ and bounded above by $\bar{X} + d$), then sometimes this assertion is correct and sometimes it is wrong, depending on the value of $\bar{X}$ in a given experiment. If for a fixed α value we would like to have a confidence probability (called confidence coefficient) such that

$$P[\mu \in I_1] = P[\bar{X} - d < \mu < \bar{X} + d] = 1 - \alpha \tag{19.11.14}$$

then we need to choose the value of d to satisfy $d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, i.e.,

$$I_1 = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{19.11.15}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th percentile of the $N(0,1)$ distribution such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. To see this, we note that from the sampling distribution of $\bar{X}$ (Fact 3) we have

$$\begin{aligned} P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= P\left[\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right] \\ &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha \end{aligned} \tag{19.11.16}$$

We further note that, even when the original population is not normal, by Theorem 2 the confidence probability is approximately $(1 - \alpha)$ when the sample size is reasonably large.

2. Confidence Interval for μ When σ^2 is Unknown. Assume that the observations are from an $N(\mu,\sigma^2)$ population. When σ^2 is unknown, by Fact 4 and a similar argument we see that

$$I_2 = \left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right) \tag{19.11.17}$$

is a confidence interval for μ with confidence probability $1 - \alpha$, where $t_{\alpha/2}(n - 1)$ is the $(1 - \alpha/2)$th percentile of the $t(n - 1)$ distribution.

3. Confidence Interval for σ^2. If, under the same assumption of normality, a confidence interval for σ^2 is needed when μ is unknown, then

$$I_3 = \left((n-1)S^2/\chi^2_{1-\alpha/2}(n-1), (n-1)S^2/\chi^2_{\alpha/2}(n-1)\right) \tag{19.11.18}$$

has a confidence probability $1 - \alpha$, when $\chi^2_{1-\alpha/2}(n - 1)$ and $\chi^2_{\alpha/2}(n - 1)$ are the $(\alpha/2)$th and $(1 - \alpha/2)$th percentiles, respectively, of the $\chi^2(n - 1)$ distribution.

Two-Sample Case

1. Confidence Intervals for $\mu_1 - \mu_2$ When $\sigma_1^2 = \sigma_2^2$ are Known. Consider an experiment that involves the comparison of two treatments, T_1 and T_2, as indicated in Table 19.11.1. If a confidence interval for $\delta = \mu_1 - \mu_2$ is needed when σ_1^2 and σ_2^2 are unknown, then by Fact 6 and a similar argument, the confidence interval

$$I_4 = \left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right) \tag{19.11.19}$$

has a confidence probability $1 - \alpha$.

2. Confidence Interval for $\mu_1 - \mu_2$ when σ_1^2, σ_2^2 are Unknown but Equal. Under the additional assumption that $\sigma_1^2 = \sigma_2^2$, but the common variance is unknown, then by Fact 8 the confidence interval

$$I_5 = \left(\left(\bar{X}_1 - \bar{X}_2\right) - d, \left(\bar{X}_1 - \bar{X}_2\right) + d\right) \tag{19.11.20}$$

has a confidence probability $1 - \alpha$, where

$$d = t_{\alpha/2}\left(n_1 + n_2 - 2\right) S_p \left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.21}$$

3. Confidence Interval for σ_2^2/σ_1^2. A confidence interval for the ratio of the variances σ_2^2/σ_1^2 can be obtained from the F distribution (see Fact 9), and the confidence interval

$$I_6 = \left(F_{1-\alpha/2}\left(n_1 - 1, n2_1 - 1\right)\frac{S_2^2}{S_1^2}, F_{\alpha/2}\left(n_1 - 1, n2_1 - 1\right)\frac{S_2^2}{S_1^2}\right) \tag{19.11.22}$$

has a confidence probability $1 - \alpha$, where $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ and $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ are, respectively, the $(\alpha/2)$th and $(1 - \alpha/2)$th percentiles of the $F(n_1 - 1, n_2 - 1)$ distribution.

Testing Statistical Hypotheses

A statistical hypothesis concerns a statement or assertion about the true value of the parameter in a given distribution. In the two-hypothesis problems, we deal with a null hypothesis and an alternative hypothesis, denoted by H_0 and H_1, respectively. A decision is to be made, based on the data of the experiment, to either accept H_0 (hence reject H_1) or reject H_0 (hence accept H_1). In such a two-action problem, we may commit two types of errors: the type I error is to reject H_0 when it is true, and the type II error is to accept H_0 when it is false. As a standard practice, we do not reject H_0 unless there is significant evidence indicating that it may be false. (In doing so, the burden of proof that H_0 is false is on the experimenter.) Thus we usually choose a small fixed number, α (such as 0.05 or 0.01), such that the probability of committing a type I error is at most (or equal to) α. With such a given α, we can then determine the region in the data space for the rejection of H_0 (called the critical region).

One-Sample Case

Suppose that $X_1, X_2, \ldots, X_n$ represent a random sample of size n from an $N(\mu,\sigma^2)$ population, and $\bar{X}$ and S^2 are, respectively, the sample mean and sample variance.

1. Test for Mean. In testing

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1\left(\mu_1 > \mu_0\right) \text{ or } H_1 : \mu > \mu_0$$

when σ^2 is known, we reject H_0 when $\bar{X}$ is large. To determine the cut-off point, we note (by Fact 3) that the statistic $Z_0 = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ has an $N(0,1)$ distribution under H_0. Thus, if we decide to reject H_0 when $Z_0 > z_\alpha$, then the probability of committing a type I error is α. As a consequence, we apply the decision rule

$$d_1 : \text{reject } H_0 \text{ if and only if } \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

Similarly, from the distribution of Z_0 under H_0 we can obtain the critical region for the other types of hypotheses. When σ^2 is unknown, then by Fact 4 $T_0 = \sqrt{n}(\bar{X} - \mu_0)/S$ has a $t(n - 1)$ distribution under H_0. Thus the corresponding tests can be obtained by substituting $t_\alpha(n - 1)$ for z_α and S for σ. The tests for the various one-sided and two-sided hypotheses are summarized in Table 19.11.2 below. For each set of hypotheses, the critical region given on the first line is for the case when σ^2 is known, and that

TABLE 19.11.2 One-Sample Tests for Mean

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\mu = \mu_0$ or $\mu \le \mu_0$	$\mu = \mu_1 > \mu_0$ or $\mu > \mu_0$	$\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$
		$\bar{X} > \mu_0 + t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$ or $\mu \ge \mu_0$	$\mu = \mu_1 < \mu_0$ or $\mu < \mu_0$	$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$
		$\bar{X} < \mu_0 - t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert \bar{X} - \mu_0 \rvert > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
		$\lvert \bar{X} - \mu_0 \rvert > t_{\alpha/2} \frac{S}{\sqrt{n}}$

given on the second line is for the case when σ^2 is unknown. Furthermore, t_α and $t_{\alpha/2}$ stand for $t_\alpha(n-1)$ and $t_{\alpha/2}(n-1)$, respectively.

2. Test for Variance. In testing hypotheses concerning the variance σ^2 of a normal distribution, use Fact 5 to assert that, under H_0: $\sigma^2 = \sigma_0^2$, the distribution of $w_0 = (n-1)\,S^2/\sigma_0^2$ is $\chi^2(n-1)$. The corresponding tests and critical regions are summarized in the following table (χ^2_α and $\chi^2_{\alpha/2}$ stand for $\chi^2_\alpha(n-1)$ and $\chi^2_{\alpha/2}(n-1)$, respectively):

TABLE 19.11.3 One-Sample Tests for Variance

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \le \sigma_0^2$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ or $\sigma^2 > \sigma_0^2$	$\left(S^2/\sigma_0^2\right) > \frac{1}{n-1}\chi^2_\alpha$
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \ge \sigma_0^2$	$\sigma^2 = \sigma_1^2 < \sigma_0^2$ or $\sigma^2 < \sigma_0^2$	$\left(S^2/\sigma_0^2\right) < \frac{1}{n-1}\chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	$\left(S^2/\sigma_0^2\right) > \frac{1}{n-1}\chi^2_{\alpha/2}$
		or $\left(S^2/\sigma_0^2\right) < \frac{1}{n-1}\chi^2_{1-\alpha/2}$

Two-Sample Case

In comparing the means and variances of two normal populations, we once again refer to Table 19.11.1 for notation and assumptions.

1. Test for Difference of Two Means. Let $\delta = \mu_1 - \mu_2$ be the difference of the two population means. In testing H_0: $\delta = \delta_0$ vs. a one-sided or two-sided alternative hypothesis, we note that, for

$$\tau = \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^{1/2} \tag{19.11.23}$$

and

$$\nu = S_p\left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.24}$$

$Z_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/\tau$ has an $N(0,1)$ distribution under H_0 and $T_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/\nu$ has a $t(n_1 + n_2 - 2)$ distribution under H_0 when $\sigma_1^2 = \sigma_2^2$. Using these results, the corresponding critical regions for one-sided and two-sided tests can be obtained, and they are listed below. Note that, as in the one-sample case, the critical region given on the first line for each set of hypotheses is for the case of known variances, and that given on the second line is for the case in which the variances are equal but unknown. Further, t_α and $t_{\alpha/2}$ stand for $t_\alpha(n_1 + n_2 - 2)$ and $t_{\alpha/2}(n_1 + n_2 - 2)$, respectively.

TABLE 19.11.4 Two-Sample Tests for Difference of Two Means

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\delta = \delta$ or $\delta \le \delta_0$	$\delta = \delta_1 > \delta_0$ or $\delta > \delta_0$	$(\bar{X}_1 - \bar{X}_2) > \delta_0 + z_\alpha \tau$
		$(\bar{X}_1 - \bar{X}_2) > \delta_0 + t_\alpha \nu$
$\delta = \delta_0$ or $\delta \ge \delta_0$	$\delta = \delta_1 < \delta_0$ or $\delta < \delta_0$	$(\bar{X}_1 - \bar{X}_2) < \delta_0 - z_\alpha \tau$
		$(\bar{X}_1 - \bar{X}_2) < \delta_0 - t_\alpha \nu$
$\delta = \delta_0$	$\delta \ne \delta_0$	$\lvert(\bar{X}_1 - \bar{X}_2) - \delta_0\rvert > z_{\alpha/2} \tau$
		$\lvert(\bar{X}_1 - \bar{X}_2) - \delta_0\rvert > t_{\alpha/2} \nu$

A Numerical Example

In the following we provide a numerical example for illustrating the construction of confidence intervals and hypothesis-testing procedures. The example is given along the line of applications in Wadsworth (1990, p. 4.21) with artificial data.

Suppose that two processes (T_1 and T_2) manufacturing steel pins are in operation, and that a random sample of 4 pins (or 5 pins) was taken from the process T_1 (the process T_2) with the following results (in units of inches):

$$T_1 : 0.7608,\ 0.7596,\ 0.7622,\ 0.7638$$

$$T_2 : 0.7546,\ 0.7561,\ 0.7526,\ 0.7572,\ 0.7565$$

Simple calculation shows that the observed values of sample means sample variances, and sample standard deviations are:

$$\bar{X}_1 = 0.7616, \quad S_1^2 = 3.280 \times 10^{-6}, \quad S_1 = 1.811 \times 10^{-3}$$

$$\bar{X}_2 = 0.7554, \quad S_2^2 = 3.355 \times 10^{-6}, \quad S_2 = 1.832 \times 10^{-3}$$

One-Sample Case

Let us first consider confidence intervals for the parameters of the first process, T_1, only.

1. Assume that, based on previous knowledge of processes of this type, the variance is known to be $\sigma_1^2 = 1.80^2 \times 10^{-6}$ ($\sigma_1 = 0.0018$). Then from the normal table (see, e.g., Ross (1987, p. 482) we have $z_{0.025} = 1.96$. Thus a 95% confidence interval for μ_1 is

$$\left(0.7616 - 1.96 \times 0.0018/\sqrt{4},\ 0.7616 + 1.96 \times 0.0018/\sqrt{4}\right)$$

or (0.7598, 0.7634) (after rounding off to the 4th decimal place).

2. If σ_1^2 is unknown and a 95% confidence interval for μ_1 is needed then, for $t_{0.023}(3) = 3.182$ (see, e.g., Ross, 1987, p. 484) the confidence interval is

$$\left(0.7616 - 3.182 \times 0.001811/\sqrt{4},\ 0.7616 + 3.182 \times 0.001811/\sqrt{4}\right)$$

or (0.7587, 0.7645)

3. From the chi-squared table with 4 – 1 = 3 degrees of freedom, we have (see, e.g., Ross, 1987, p. 483) $\chi^2_{0.975} = 0.216$, $\chi^2_{0.025} = 9.348$. Thus a 95% confidence interval for σ_1^2 is $(3 \times 3.280 \times 10^{-6}/9.348,\ 3 \times 3.280 \times 10^{-6}/0.216)$, or $(1.0526 \times 10^{-6},\ 45{,}5556 \times 10^{-6})$.
4. In testing the hypotheses

$$H_0 : \mu_1 = 0.76 \text{ vs. } H_1 : \mu_1 > 0.76$$

with a = 0.01 when σ_1^2 is unknown, the critical region is $\bar{x}_1 > 0.76 + 4.541 \times 0.001811/\sqrt{4} = 0.7641$. Since the observed value $\bar{x}_1$ is 0.7616, H_0 is accepted. That is, we assert that there is no significant evidence to call for the rejection of H_0.

Two-Sample Case

If we assume that the two populations have a common unknown variance, we can use the Student's *t* distribution (with degree of freedom $\nu = 4 + 5 - 2 = 7$) to obtain confidence intervals and to test hypotheses for $\mu_1 - \mu_2$. We first note that the data given above yield

$$S_p^2 = \frac{1}{7}(3 \times 3.280 + 4 \times 3.355) \times 10^{-6}$$

$$= 3.3229 \times 10^{-6}$$

$$S_p = 1.8229 \times 10^{-3} \qquad \nu = S_p\sqrt{1/4 + 1/5} = 1.2228 \times 10^{-3}$$

and $\bar{X}_1 - \bar{X}_2 = 0.0062$.

1. A 98% confidence interval for $\mu_1 - \mu_2$ is $(0.0062 - 2.998\nu,\ 0.0062 + 2.998\nu)$ or (0.0025, 0.0099).
2. In testing the hypotheses H_0: $\mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$) vs. H_1: $\mu_1 > \mu_2$ with $\alpha = 0.05$, the critical region is $(\bar{X}_1 - \bar{X}_2) > 1.895\nu = 2.3172 \times 10^{-3}$. Thus H_0 is rejected; i.e., we conclude that there is significant evidence to indicate that $\mu_1 > \mu_2$ may be true.
3. In testing the hypotheses H_0: $\mu_1 = \mu_2$ vs. $\mu_1 \neq \mu_2$ with $\alpha = 0.02$, the critical region is $|\bar{X}_1 - \bar{X}_2| > 2.998\nu = 3.6660 \times 10^{-3}$. Thus H_0 is rejected. We note that the conclusion here is consistent with the result that, with confidence probability $1 - \alpha = 0.98$, the confidence interval for $(\mu_1 - \mu_2)$ does not contain the origin.

Concluding Remarks

The history of probability and statistics goes back to the days of the celebrated mathematicians K. F. Gauss and P. S. Laplace. (The normal distribution, in fact, is also called the Gaussian distribution.) The theory and methods of classical statistical analysis began its developments in the late 1800s and early 1900s when F. Galton and R.A. Fisher applied statistics to their research in genetics, when Karl Pearson developed the chi-square goodness-of-fit method for stochastic modeling, and when E.S. Pearson and J. Neyman developed the theory of hypotheses testing. Today statistical methods have been found useful in analyzing experimental data in biological science and medicine, engineering, social sciences, and many other fields. A non-technical review on some of the applications is Hacking (1984).

Applications of statistics in engineering include many topics. In addition to those treated in this section, other important ones include sampling inspection and quality (process) control, reliability, regression analysis and prediction, design of engineering experiments, and analysis of variance. Due to space limitations, these topics are not treated here. The reader is referred to textbooks in this area for further information. There are many well-written books that cover most of these topics, the following short list consists of a small sample of them.

References

Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. *Statistics for Experimenters.* John Wiley & Sons, New York.

Bowker, A.H. and Lieberman, G.J. 1972. *Engineering Statistics*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Hacking, I. 1984. Trial by number, *Science*, 84(5), 69–70.

Hahn, G.J. and Shapiro, S.S. 1967. *Statistical Models in Engineering.* John Wiley & Sons, New York.

Hines, W.W. and Montgomery, D.G. 1980. *Probability and Statistics in Engineering and Management Science.* John Wiley & Sons, New York.

Hogg, R.V. and Ledolter, J. 1992. *Engineering Statistics.* Macmillan, New York.

Ross, S.M. 1987. *Introduction to Probability and Statistics for Engineers and Scientists.* John Wiley & Sons, New York.

Wadsworth, H.M., Ed. 1990. *Handbook of Statistical Methods for Engineers and Scientists.* John Wiley & Sons, New York.

19.12 Numerical Methods

William F. Ames

Introduction

Since many mathematical models of physical phenomena are not solvable by available mathematical methods one must often resort to approximate or numerical methods. These procedures do not yield exact results in the mathematical sense. This inexact nature of numerical results means we must pay attention to the errors. The two errors that concern us here are *round-off errors* and *truncation errors.*

Round-off errors arise as a consequence of using a number specified by m correct digits to approximate a number which requires more than m digits for its exact specification. For example, using 3.14159 to approximate the irrational number π. Such errors may be especially serious in matrix inversion or in any area where a very large number of numerical operations are required. Some attempts at handling these errors are called *enclosure methods.* (Adams and Kulisch, 1993).

Truncation errors arise from the substitution of a finite number of steps for an infinite sequence of steps (usually an iteration) which would yield the exact result. For example, the iteration $y_n(x) = 1 + \int_0^x xty_{n-1}(t)dt$, $y(0) = 1$ is only carried out for a *few steps*, but it converges in *infinitely* many steps.

The study of some errors in a computation is related to the theory of probability. In what follows, a relation for the error will be given in certain instances.

Linear Algebra Equations

A problem often met is the determination of the solution vector $u = (u_1, u_2, \ldots, u_n)^T$ for the set of linear equations $Au = v$ where A is the $n \times n$ square matrix with coefficients, a_{ij} ($i, j = 1, \ldots, n$), $v = (v_1, \ldots, v_n)^T$ and i denotes the row index and j the column index.

There are many numerical methods for finding the solution, u, of $Au = v$. The direct inversion of A is usually too expensive and is not often carried out unless it is needed elsewhere. We shall only list a few methods. One can check the literature for the many methods and computer software available. Some of the software is listed in the References section at the end of this chapter. The methods are usually subdivided into *direct* (once through) or *iterative* (repeated) procedures.

In what follows, it will often be convenient to partition the matrix A into the form $A = U + D + L$, where U, D, and L are matrices having the same elements as A, respectively, above the main diagonal, on the main diagonal, and below the main diagonal, and zeros elsewhere. Thus,

$$U = \begin{bmatrix} 0 & a_{12} & & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ \vdots & \cdots & & \cdots & \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

We also assume the u_js are not all zero and $\det A \neq 0$ so the solution is unique.

Direct Methods

Gauss Reduction. This classical method has spawned many variations. It consists of dividing the first equation by a_{11} (if $a_{11} = 0$, reorder the equations to find an $a_{11} \neq 0$) and using the result to eliminate the terms in u_1 from each of the succeeding equations. Next, the modified second equation is divided by a'_{22} (if $a'_{22} = 0$, a reordering of the modified equations may be necessary) and the resulting equation is used to eliminate all terms in u_2 in the succeeding modified equations. This elimination is done n times resulting in a triangular system:

$$
\begin{aligned}
u_1 + a'_{12}u_2 + \cdots + a'_{1n}u_n &= v'_1 \\
0 + u_2 + \cdots + a'_{2n}u_n &= v'_2 \\
&\cdots \\
0 + \cdots + u_{n-1} + a'_{n-1,n}u_n &= v'_{n-1} \\
u_n &= v'_n
\end{aligned}
$$

where a'_{ij} and v'_j represent the specific numerical values obtained by this process. The solution is obtained by working backward from the last equation. Various modifications, such as the Gauss-Jordan reduction, the Gauss-Doolittle reduction, and the Crout reduction, are described in the classical reference authored by Bodewig (1956). Direct methods prove very useful for sparse matrices and banded matrices that often arise in numerical calculation for differential equations. Many of these are available in computer packages such as IMSL, Maple, Matlab, and Mathematica.

The Tridiagonal Algorithm. When the linear equations are tridiagonal, the system

$$
\begin{aligned}
b_1u_1 + c_1u_2 &= d_1 \\
a_iu_{i-1} + b_iu_i + c_iu_{i+1} &= d_i \\
a_nu_{n-1} + b_nu_n &= d_n, \quad i = 2, 3, \ldots, n-1
\end{aligned}
$$

can be solved explicitly for the unknown, thereby eliminating any matrix operations.

The Gaussian elimination process transforms the system into a simpler one of *upper bidiagonal* form. We designate the coefficients of this new system by a'_i, b'_i c'_i and d'_i, and we note that

$$
\begin{aligned}
a'_i &= 0, \quad i = 2, 3, \ldots, n \\
b'_i &= 1, \quad i = 1, 2, \ldots, n
\end{aligned}
$$

The coefficients c'_i and d'_i are calculated successively from the relations

$$
c'_1 = \frac{c_1}{b_1} \qquad d'_1 = \frac{d_1}{b_1}
$$

$$
c'_{i+1} = \frac{c_{i+1}}{b_{i+1} - a_{i+1}c'_i}
$$

$$
d'_{i+1} = \frac{d_{i+1} - a_{i+1}d'_i}{b_{i+1} - a_{i+1}c'_i}, \quad i = 1, 2, \ldots, n-1
$$

and, of course, $c_n = 0$.

Having completed the elimination we examine the new system and see that the *n*th equation is now

$$
u_n = d'_n
$$

Substituting this value into the $(n - 1)$st equation,

$$
u_{n-1} + c'_{n-1}u_n = d'_{n-1}
$$

we have

$$u_{n-1} = d'_{n-1} - c'_{n-1}u_n$$

Thus, starting with u_n, we have successively the solution for u_i as

$$u_i = d'_i - c'_i u_{i+1}, \quad i = n-1,\ n-2,\ \ldots,\ 1$$

Algorithm for Pentadiagonal Matrix. The equations to be solved are

$$a_i u_{i-2} + b_i u_{i-1} + c_i u_i + d_i u_{i+1} + e_i u_{i+2} = f_i$$

for $1 \le i \le R$ with $a_1 = b_1 = a_2 = e_{R-1} = d_R = e_R = 0$.

The algorithm is as follows. First, compute

$$\delta_1 = d_1/c_1$$
$$\lambda_1 = e_1/c_1$$
$$\gamma_1 = f_1/c_1$$

and

$$\mu_2 = c_2 - b_2\delta_1$$
$$\delta_2 = \left(d_2 - b_2\lambda_1\right)/\mu_2$$
$$\lambda_2 = e_2/\mu_2$$
$$\gamma_2 = \left(f - b_2\gamma_1\right)/\mu_2$$

Then, for $3 \le i \le R-2$, compute

$$\beta_i = b_i - a_i\delta_{i-2}$$
$$\mu_i = c_i - \beta_i\delta_{i-1} - a_i\lambda_{i-2}$$
$$\delta_i = \left(d_i - \beta_i\lambda_{i-1}\right)/\mu_i$$
$$\lambda_i = e_i/\mu_i$$
$$\gamma_i = \left(f_i - \beta_i\gamma_{i-1} - a_i\gamma_{i-2}\right)/\mu_i$$

Next, compute

$$\beta_{R-1} = b_{R-1} - a_{R-1}\delta_{R-3}$$
$$\mu_{R-1} = c_{R-1} - \beta_{R-1}\delta_{R-2} - a_{R-1}\lambda_{R-3}$$
$$\delta_{R-1} = \left(d_{R-1} - \beta_{R-1}\lambda_{R-2}\right)/\mu_{R-1}$$
$$\gamma_{R-1} = \left(f_{R-1} - \beta_{R-1}\gamma_{R-2} - a_{R-1}\gamma_{R-3}\right)/\mu_{R-1}$$

and

$$\beta_R = b_R - a_R\delta_{R-2}$$
$$\mu_R = c_R - \beta_R\delta_{R-1} - a_R\lambda_{R-2}$$
$$\gamma_R = \left(f_R - \beta_R\gamma_{R-1} - a_R\gamma_{R-2}\right)/\mu_R$$

The β_i and μ_i are used only to compute δ_i, λ_i, and γ_i, and need not be stored after they are computed. The δ_i, λ_i, and γ_i, must be stored, as they are used in the back solution. This is

$$u_R = \gamma_R$$
$$u_{R-1} = \gamma_{R-1} - \delta_{R-1}u_R$$

and

$$u_i = \gamma_i - \delta_i u_{i+1} - \lambda_i u_{i+2}$$

for $R - 2 \geq i \geq 1$.

General Band Algorithm. The equations are of the form

$$A_j^{(M)}X_{j-M} + A_j^{(M-1)}X_{j-M+1} + \cdots + A_j^{(2)}X_{j-2} + A_j^{(1)}X_{j-1} + B_jX_j$$
$$+ C_j^{(1)}X_{j+1} + C_j^{(2)}X_{j+2} + \cdots + C_j^{(M-1)}X_{j+M-1} + C_j^{(M)}X_{j+M} = D_j$$

for $1 \leq j \leq N$, $N \geq M$. The algorithm used is as follows:

$$\alpha_j^{(k)} = A_j^{(k)} = 0, \quad \text{for } k \geq j$$
$$C_j^{(k)} = 0, \quad \text{for } k \geq N + 1 - j$$

The forward solution ($j = 1, \ldots, N$) is

$$\alpha_j^{(k)} = A_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p)} W_{j-p}^{(p-k)}, \quad k = M, \ldots, 1$$
$$\beta_j = B_j - \sum_{p=1}^{M} \alpha_j^{(p)} W_{j-p}^{(p)}$$
$$W_j^{(k)} = \left(C_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p-k)} W_{j-(p-k)}^{(p)}\right)\Big/\beta_j, \quad k = 1, \ldots, M$$
$$\gamma_j = \left(D_j - \sum_{p=1}^{M} \alpha_j^{(p)}\gamma_{j-p}\right)\Big/\beta_j$$

The back solution ($j = N, \ldots, 1$) is

$$X_j = \gamma_j - \sum_{p=1}^{M} W_j^{(p)} X_{j+p}$$

Cholesky Decomposition. When the matrix A is a symmetric and positive definite, as it is for many discretizations of self-adjoint positive definite boundary value problems, one can improve considerably on the band procedures by using the Cholesky decomposition. For the system $Au = v$, the Matrix A can be written in the form

$$A = (I + L)D(I + U)$$

where L is lower triangular, U is upper triangular, and D is diagonal. If $A = A'$ (A' represents the transpose of A), then

$$A = A' = (I + U)' D(I + L)'$$

Hence, because of the uniqueness of the decomposition.

$$I + L = (I + U)' = I + U'$$

and therefore,

$$A = (I + U)' D(I + U)$$

that is,

$$A = B'B, \text{ where } B = \sqrt{D}(I + U)$$

The system $Au = v$ is then solved by solving the two triangular system

$$B'w = v$$

followed by

$$Bu = w$$

To carry out the decomposition $A = B'B$, all elements of the first row of A, and of the derived system, are divided by the square root of the (positive) leading coefficient. This yields smaller rounding errors than the banded methods because the relative error of $\sqrt{a}$ is only half as large as that of a itself. Also, taking the square root brings numbers nearer to each other (i.e., the new coefficients do not differ as widely as the original ones do). The actual computation of $B = (b_{ij})$, $j > i$, is given in the following:

$$b_{11} = \left(a_{11}\right)^{1/2}, \qquad b_{1j} = a_{ij}/b_{11}, \quad j \geq 2$$

$$b_{22} = \left(a_{22} - b_{12}^2\right)^{1/2}, \qquad b_{2j} = \left(a_{2j} - b_{12}b_{1j}\right)\Big/b_{22}$$

$$b_{33} = \left(a_{33} - b_{13}^2 - b_{23}^2\right)^{1/2}, \qquad b_{3j} = \left(a_{3j} - b_{13}b_{1j} - b_{23}b_{2j}\right)\Big/b_{33}$$

$$\vdots$$

$$b_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} b_{ki}^2\right)^{1/2}, \qquad b_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} b_{ki}b_{kj}\right)\Big/b_{ii}, \qquad i \geq 2,\ j \geq 2$$

Iterative Methods

Iterative methods consist of repeated application of an often simple algorithm. They yield the exact answer only as the limit of a sequence. They can be programmed to take care of zeros in A and are self-correcting. Their structure permits the use of convergence accelerators, such as overrelaxation, Aitkins acceleration, or Chebyshev acceleration.

Let $a_{ii} > 0$ for all i and $\det A \neq 0$. With $A = U + D + L$ as previously described, several iteration methods are described for $(U + D + L)u = v$.

Jacobi Method (Iteration by total steps). Since $u = -D^{-1}[U + L]u + D^{-1}v$, the iteration $u^{(k)}$ is $u^{(k)} = -D^{-1}[U + L]u^{(k-1)} + D^{-1}v$. This procedure has a slow convergent rate designated by R, $0 < R \ll 1$.

Gauss-Seidel Method (Iteration by single steps). $u^{(k)} = -(L + D)^{-1}Uu^{(k-1)} + (L + D)^{-1}v$. Convergence rate is $2R$, twice as fast as that of the Jacobi method.

Gauss-Seidel with Successive Overrelaxation (SOR). Let $\bar{u}_i^{(k)}$ be the ith components of the Gauss-Seidel iteration. The SOR technique is defined by

$$u_i^{(k)} = (1 - \omega)u_i^{(k-1)} + \omega\bar{u}_i^{(k)}$$

where $1 < \omega < 2$ is the overrelaxation parameter. The full iteration is $u^{(k)} = (D + \omega L)^{-1}\{[(1 - \omega)D - \omega U]u^{(k-1)} + \omega v\}$. Optimal values of ω can be computed and depend upon the properties of A (Ames, 1993). With optimal values of ω, the convergence rate of this method is $2R\sqrt{2}$ which is much larger than that for Gauss-Seidel (R is usually much less than one).

For other acceleration techniques, see the literature (Ames, 1993).

Nonlinear Equations in One Variable

Special Methods for Polynomials

The polynomial $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$, with real coefficients a_j, $j = 0, \ldots, n$, has exactly n roots which may be real or complex.

If all the coefficients of $P(x)$ are integers, then any rational roots, say r/s (r and s are integers with no common factors), of $P(x) = 0$ must be such that r is an integral divisor of a_n and s is an integral division of a_0. Any polynomial with rational coefficients may be converted into one with integral coefficients by multiplying the polynomial by the lowest common multiple of the denominators of the coefficients.

Example. $x^4 - 5x^2/3 + x/5 + 3 = 0$. The lowest common multiple of the denominators is 15. Multiplying by 15, which does not change the roots, gives $15x^4 - 25x^2 + 3x + 45 = 0$. The only possible rational roots r/s are such that r may have the value ±45, ±15, ±5, ±3, and ±1, while s may have the values ±15,

±5, ±3, and ±1. All possible rational roots, with no common factors, are formed using all possible quotients.

If $a_0 > 0$, the first negative coefficient is preceded by k coefficients which are positive or zero, and G is the largest of the absolute values of the negative coefficients, then each real root is less than $1 + \sqrt[k]{G/a_0}$ (upper bound on the real roots). For a lower bound to the real roots, apply the criterion to $P(-x) = 0$.

Example. $P(x) = x^5 + 3x^4 - 2x^3 - 12x + 2 = 0$. Here $a_0 = 1$, $G = 12$, and $k = 2$. Thus, the upper bound for the real roots is $1 + \sqrt[2]{12} \approx 4.464$. For the lower bound, $P(-x) = -x^5 + 3x^4 + 2x^3 + 12x + 2 = 0$, which is equivalent to $x^5 - 3x^4 - 2x^3 - 12x - 2 = 0$. Here $k = 1$, $G = 12$, and $a_0 = 1$. A lower bound is $-(1 + 12) = 13$. Hence all real roots lie in $-13 < x < 1 + \sqrt[2]{12}$.

A useful *Descartes rule of signs* for the number of positive or negative real roots is available by observation for polynomials with real coefficients. The number of positive real roots is either equal to the number of sign changes, n, or is less than n by a positive *even* integer. The number of negative real roots is either equal to the number of sign changes, n, of $P(-x)$, or is less than n by a positive even integer.

Example. $P(x) = x^5 - 3x^3 - 2x^2 + x - 1 = 0$. There are three sign changes, so $P(x)$ has either three or one positive roots. Since $P(-x) = -x^5 + 3x^3 - 2x^2 - 1 = 0$, there are either two or zero negative roots.

The Graeffe Root-Squaring Technique

This is an iterative method for finding the roots of the algebraic equation

$$f(x) = a_0 x^p + a_1 x^{p-1} + \cdots + a_{p-1} x + a_p = 0$$

If the roots are r_1, r_2, r_3, …, then one can write

$$S_p = r_1^p \left(1 + \frac{r_2^p}{r_1^p} + \frac{r_3^p}{r_1^p} + \cdots\right)$$

and if one root is larger than all the others, say r_1, then for large enough p all terms (other than 1) would become negligible. Thus,

$$S_p \approx r_1^p$$

or

$$\lim_{p \to \infty} S_p^{1/p} = r_1$$

The Graeffe procedure provides an efficient way for computing S_p via a sequence of equations such that the roots of each equation are the squares of the roots of the preceding equations in the sequence. This serves the purpose of ultimately obtaining an equation whose roots are so widely separated in magnitude that they may be read approximately from the equation by inspection. The basic procedure is illustrated for a polynomial of degree 4:

$$f(x) = a_0 x^4 + a_1 x^3 + a_2 x^2 + a_3 x + a_4 = 0$$

Rewrite this as

$$a_0 x^4 + a_2 x^2 + a_4 = -a_1 x^3 - a_3 x$$

and square both sides so that upon grouping

$$a_0^2x^8 + \left(2a_0a_2 - a_1^2\right)x^6 + \left(2a_0a_4 - 2a_1a_3 + a_2^2\right)x^4 + \left(2a_2a_4 - a_3^2\right)x^2 + a_4^2 = 0$$

Because this involves only even powers of x, we may set $y = x^2$ and rewrite it as

$$a_0^2y^4 + \left(2a_0a_2 - a_1^2\right)y^3 + \left(2a_0a_4 - 2a_1a_3 + a_2^2\right)y^2 + \left(2a_2a_4 - a_3^2\right)y + a_4^2 = 0$$

whose roots are the squares of the original equation. If we repeat this process again, the new equation has roots which are the fourth power, and so on. After p such operations, the roots are 2^p (original roots). If at any stage we write the coefficients of the unknown in sequence

$$a_0^{(p)} \quad a_1^{(p)} \quad a_2^{(p)} \quad a_3^{(p)} \quad a_4^{(p)}$$

then, to get the new sequence $a_i^{(p+1)}$, write $a_i^{(p+1)} = 2a_0^{(p)}$ (times the symmetric coefficient) with respect to $a_i^{(p)} - 2a_1^{(p)}$ (times the symmetric coefficient) $- \cdots (-1)^i a_i^{(p)2}$. Now if the roots are r_1, r_2, r_3, and r_4, then $a_1/a_0 = -\Sigma_{i-1}^4 r_i$, $a_i^{(1)}/a_0^{(1)} = -\Sigma r_i^2, \ldots, a_i^{(p)}/a_0^{(p)} = -\Sigma r_i^{2p}$. If the roots are all distinct and r_1 is the largest in magnitude, then eventually

$$r_1^{2p} \approx -\frac{a_1^{(p)}}{a_0^{(p)}}$$

And if r_2 is the next largest in magnitude, then

$$r_2^{2p} \approx -\frac{a_2^{(p)}}{a_1^{(p)}}$$

And, in general $a_n^{(p)}/a_{n-1}^{(p)} \approx -r_n^{2p}$. This procedure is easily generalized to polynomials of arbitrary degree and specialized to the case of multiple and complex roots.

Other methods include Bernoulli iteration, Bairstow iteration, and Lin iteration. These may be found in the cited literature. In addition, the methods given below may be used for the numerical solution of polynomials.

General Methods for Nonlinear Equations in One Variable

Successive Substitutions

Let $f(x) = 0$ be the nonlinear equation to be solved. If this is rewritten as $x = F(x)$, then an iterative scheme can be set up in the form $x_{k+1} = F(x_k)$. To start the iteration, an initial guess must be obtained graphically or otherwise. The convergence or divergence of the procedure depends upon the method of writing $x = F(x)$, of which there will usually be several forms. A general rule to ensure convergence cannot be given. However, if a is a root of $f(x) = 0$, a necessary condition for convergence is that $|F'(x)| < 1$ in that interval about a in which the iteration proceeds (this means the iteration cannot converge unless $|F'(x)| < 1$, but it does not ensure convergence). This process is called *first order* because the error in x_{k+1} is proportional to the first power of the error in x_k.

Example. $f(x) = x^3 - x - 1 = 0$. A rough plot shows a real root of approximately 1.3. The equation can be written in the form $x = F(x)$ in several ways, such as $x = x^3 - 1$, $x = 1/(x^2 - 1)$, and $x = (1 + x)^{1/3}$. In the first case, $F'(x) = 3x^2 = 5.07$ at $x = 1.3$; in the second, $F'(1.3) = 5.46$; only in the third case

is $F'(1.3) < 1$. Hence, only the third iterative process has a chance to converge. This is illustrated in the iteration table below.

Step k	$x = \frac{1}{x^2 - 1}$	$x = x^3 - 1$	$x = (1 + x)^{1/3}$
0	1.3	1.3	1.3
1	1.4493	1.197	1.32
2	0.9087	0.7150	1.3238
3	–5.737	–0.6345	1.3247
4	⋯	⋯	1.3247

Numerical Solution of Simultaneous Nonlinear Equations

The techniques illustrated here will be demonstrated for two simultaneous equations — $f(x, y) = 0$ and $g(x, y) = 0$. They immediately generalize to more than two simultaneous equations.

The Method of Successive Substitutions

The two simultaneous equations can be written in various ways in equivalent forms

$$x = F(x,y)$$

$$y = G(x,y)$$

and the method of successive substitutions can be based on

$$x_{k+1} = F(x_k, y_k)$$

$$y_{k+1} = G(x_k, y_k)$$

Again, the procedure is of the first order and a necessary condition for convergence is

$$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \qquad \left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| < 1$$

in the iteration neighborhood of the true solution.

The Newton-Raphson Procedure

Using the two simultaneous equation, start from an approximate, say (x_0, y_0), obtained graphically or from a two-way table. Then, solve successively the linear equations

$$\Delta x_k \frac{\partial f}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial f}{\partial y}(x_k, y_k) = -f(x_k, y_k)$$

$$\Delta x_k \frac{\partial g}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial g}{\partial y}(x_k, y_k) = -g(x_k, y_k)$$

for Δx_k and Δy_k. Then, the $k + 1$ approximation is given from $x_{k+1} = x_k + \Delta x_k, y_{k+1} = y_k + \Delta y_k$. A modification consists in solving the equations with (x_k, y_k) replaced by (x_0, y_0) (or another suitable pair later on in the iteration) in the derivatives. This means the derivatives (and therefore the coefficients of Δx_k, Δy_k) are independent of k. Hence, the results become

$$\Delta x_k = \frac{-f(x_k,y_k)(\partial g/\partial y)(x_0,y_0) + g(x_k,y_k)(\partial f/\partial y)(x_0,y_0)}{(\partial f/\partial x)(x_0,y_0)(\partial g/\partial y)(x_0,y_0) - (\partial f/\partial y)(x_0,y_0)(\partial g/\partial x)(x_0,y_0)}$$

$$\Delta y_k = \frac{-g(x_k,y_k)(\partial f/\partial x)(x_0,y_0) + f(x_k,y_k)(\partial g/\partial x)(x_0,y_0)}{(\partial f/\partial x)(x_0,y_0)(\partial g/\partial y)(x_0,y_0) - (\partial f/\partial y)(x_0,y_0)(\partial g/\partial x)(x_0,y_0)}$$

and $x_{k+1} = \Delta x_k + x_k$, $y_{k+1} = \Delta y_k + y_k$. Such an alteration of the basic technique reduces the rapidity of convergence.

Example

$$f(x,y) = 4x^2 + 6x - 4xy + 2y^2 - 3$$

$$g(x,y) = 2x^2 - 4xy + y^2$$

By plotting, one of the approximate roots is found to be $x_0 = 0.4$, $y_0 = 0.3$. At this point, there results $\partial f/\partial x = 8$, $\partial f/\partial y = -0.4$, $\partial g/\partial x = 0.4$, and $\partial g/\partial y = -1$. Hence,

$$x_{k+1} = x_k + \Delta x_k = x_k + \frac{-f(x_k,y_k) - 0.4g(x_k,y_k)}{8(-1) - (-0.4)(0.4)}$$

$$= x_k - 0.127\,55 f(x_k,y_k) - 0.051\,02 g(x_k,y_k)$$

and

$$y_{k+1} = y_k - 0.051\,02 f(x_k,y_k) + 1.020\,41 g(x_k,y_k)$$

The first few iteration steps are shown in the following table.

Step k	x_k	y_k	$f(x_k, y_k)$	$g(x_k, y_k)$
0	0.4	0.3	–0.26	0.07
1	0.43673	0.24184	0.078	0.0175
2	0.42672	0.25573	–0.0170	–0.007
3	0.42925	0.24943	0.0077	0.0010

Methods of Perturbation

Let $f(x) = 0$ be the equation. In general, the iterative relation is

$$x_{k+1} = x_k - \frac{f(x_k)}{\alpha_k}$$

where the iteration begins with x_0 as an initial approximation and α_k is some functional.

The Newton-Raphson Procedure. This variant chooses $\alpha_k = f'(x_k)$ where $f' = df/dx$ and geometrically consists of replacing the graph of $f(x)$ by the tangent line at $x = x_k$ in each successive step. If $f'(x)$ and $f''(x)$ have the same sign throughout an interval $a \le x \le b$ containing the solution, with $f(a)$ and $f(b)$ of opposite signs, then the process converges starting from any x_0 in the interval $a \le x \le b$. The process is second order.

Example

$$f(x) = x - 1 + \frac{(0.5)^x - 0.5}{0.3}$$

$$f'(x) = 1 - 2.3105[0.5]^x$$

An approximate root (obtained graphically) is 2.

Step k	x_k	$f(x_k)$	$f'(x_k)$
0	2	0.1667	0.4224
1	1.605	–0.002	0.2655
2	1.6125	–0.0005	⋯

The Method of False Position. This variant is commenced by finding x_0 and x_1 such that $f(x_0)$ and $f(x_1)$ are of opposite signs. Then, α_1 = slope of secant line joining $[x_0, f(x_0)]$ and $[x_1, f(x_1)]$ so that

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1)$$

In each following step, α_k is the slope of the line joining $[x_k, f(x_k)]$ to the most recently determined point where $f(x_j)$ has the opposite sign from that of $f(x_k)$. This method is of first order.

The Method of Wegstein

This is a variant of the method of successive substitutions which forces or accelerates convergence. The iterative procedure $x_{k+1} = F(x_k)$ is revised by setting $\hat{x}_{k+1} = F(x_k)$ and then taking $x_{k+1} = qx_k + (1 - q)\ \hat{x}_{k+1}$. Wegstein found that suitably chosen qs are related to the basic process as follows:

Behavior of Successive Substitution Process	Range of Optimum q
Oscillatory convergence	$0 < q < 1/2$
Oscillatory divergence	$1/2 < q < 1$
Monotonic convergence	$q < 0$
Monotonic divergence	$1 < q$

At each step, q may be calculated to give a locally optimum value by setting

$$q = \frac{x_{k+1} - x_k}{x_{k+1} - 2x_k + x_{k-1}}$$

The Method of Continuity

In the case of n equations in n unknowns, when n is large, determining the approximate solution may involve considerable effort. In such a case, the method of continuity is admirably suited for use on either digital or analog computers. It consists basically of the introduction of an extra variable into the n equations

$$f_i(x_1, x_2, \ldots, x_n) = 0, \qquad i = 1, \ldots, n$$

and replacing them by

$$f_i\left(x_1, x_2, \ldots, x_n, \lambda\right) = 0, \qquad i = 1, \ldots, n$$

where λ is introduced in such a way that the functions depend in a simple way upon λ and reduce to an easily solvable system for λ = 0 and to the original equations for λ = 1. A system of ordinary differential equations, with independent variable λ, is then constructed by differentiating with respect to λ. There results

$$\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \frac{dx_j}{d\lambda} + \frac{\partial f_i}{\partial \lambda} = 0$$

where x_1, …, x_n are considered as functions of λ. The equations are integrated, with initial conditions obtained with λ = 0, from λ = 0 to λ = 1. If the solution can be continued to λ = 1, the values of x_1, …, x_n for λ = 1 will be a solution of the original equations. If the integration becomes infinite, the parameter λ must be introduced in a different fashion. Integration of the differential equations (which are usually nonlinear in λ) may be accomplished on an analog computer or by digital means using techniques described in a later section entitled "Numerical Solution of Ordinary Differential Equations."

Example

$$f(x,y) = 2 + x + y - x^2 + 8xy + y^3 = 0$$

$$g(x,y) = 1 + 2x + 3y + x^2 + xy - ye^x = 0$$

Introduce λ as

$$f(x,y,\lambda) = (2 + x + y) + \lambda\left(-x^2 + 8xy + y^3\right) = 0$$

$$g(x,y,\lambda) = (1 + 2x - 3y) + \lambda\left(x^2 + xy - ye^x\right) = 0$$

For λ = 1, these reduce to the original equations, but, for λ = 0, they are the linear systems

$$x + y = -2$$

$$2x - 3y = -1$$

which has the unique solution $x = -1.4$, $y = -0.6$. The differential equations in this case become

$$\frac{\partial f}{\partial x}\frac{dx}{d\lambda} + \frac{\partial f}{\partial y}\frac{dy}{d\lambda} = -\frac{\partial f}{\partial \lambda}$$

$$\frac{\partial g}{\partial x}\frac{dx}{d\lambda} + \frac{\partial g}{\partial y}\frac{dy}{d\lambda} = -\frac{\partial g}{\partial \lambda}$$

or

$$\frac{dx}{d\lambda} = \frac{\dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial \lambda} - \dfrac{\partial f}{\partial \lambda}\dfrac{\partial g}{\partial x}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

$$\frac{dy}{d\lambda} = \frac{\dfrac{\partial f}{\partial \lambda}\dfrac{\partial g}{\partial x} - \dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial \lambda}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

Integrating in λ, with initial values $x = -1.4$ and $y = -0.6$ at $\lambda = 0$, from $\lambda = 0$ to $\lambda = 1$ gives the solution.

Interpolation and Finite Differences

The practicing engineer constantly finds it necessary to refer to tables as sources of information. Consequently, interpolation, or that procedure of "reading between the lines of the table," is a necessary topic in numerical analysis.

Linear Interpolation

If a function $f(x)$ is approximately linear in a certain range, then the ratio $[f(x_1) - f(x_0)]/(x_1 - x_0) = f[x_0, x_1]$ is approximately independent of x_0 and x_1 in the range. The linear approximation to the function $f(x)$, $x_0 < x < x_1$, then leads to the interpolation formula

$$f(x) \approx f(x_0) + (x - x_0) f[x_0, x_1] \approx f(x_0) + \frac{x - x_0}{x_1 - x_0}\left[f(x_1) - f(x_0)\right]$$

$$\approx \frac{1}{x_1 - x_0}\left[(x_1 - x) f(x_0) - (x_0 - x) f(x_1)\right]$$

Divided Differences of Higher Order and Higher-Order Interpolation

The first-order divided difference $f[x_0, x_1]$ was defined above. Divided differences of second and higher order are defined iteratively by

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

$$\vdots$$

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}$$

and a convenient form for computational purposes is

$$f[x_0, x_1, \ldots, x_k] = \sum_{j=0}^{k}{}' \frac{f(x_j)}{(x_j - x_0)(x_j - x_1)\cdots(x_j - x_k)}$$

for any $k \geq 0$, where the $'$ means the term $(x_j - x_j)$ is omitted in the denominator. For example,

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

If the accuracy afforded by a linear approximation is inadequate, a generally more accurate result may be based upon the assumption that $f(x)$ may be approximated by a polynomial of degree 2 or higher over certain ranges. This assumptions leads to *Newton's fundamental interpolation formula* with divided differences:

$$f(x) \approx f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2]$$
$$+ (x - x_0)(x - x_1) \cdots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n] + E_n(x)$$

where $E_n(x)$ = error = $[1/(n + 1)!] f^{(n-1)}(\varepsilon)\pi(x)$ where $\min(x_0, \ldots, x) < \varepsilon < \max(x_0, x_1, \ldots, x_n, x)$ and $\pi(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$. In order to use this most effectively, one may first form a divided-difference table. For example, for third-order interpolation, the difference table is

x_0	$f(x_0)$			
		$f[x_0, x_1]$		
x_1	$f(x_1)$		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	$f(x_2)$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	$f(x_3)$			

where each entry is given by taking the difference between diagonally adjacent entries to the left, divided by the abscissas corresponding to the ordinates intercepted by the diagonals passing through the calculated entry.

Example. Calculate by third-order interpolation the value of cosh 0.83 given cosh 0.60, cosh 0.80, cosh 0.90, and cosh 1.10.

$x_0 = 0.60$	1.185 47			
		0.7598		
$x_1 = 0.80$	1.337 43		0.6560	
		0.9566		0.1586
$x_2 = 0.90$	1.433 09		0.7353	
		1.1772		
$x_3 = 1.10$	1.668 52			

With $n = 3$, we have

$$\cosh 0.83 \approx 1.185\ 47 + (0.23)(0.7598) + (0.23)(0.03)(0.6560)$$
$$+ (0.23)(0.03)(-0.07)(0.1586) = 1.364\ 64$$

which varies from the true value by 0.000 04.

Lagrange Interpolation Formulas

The Newton formulas are expressed in terms of divided differences. It is often useful to have interpolation formulas expressed explicitly in terms of the ordinates involved. This is accomplished by the Lagrange interpolation polynomial of degree n:

$$y(x) = \sum_{j=0}^{n} \frac{\pi(x)}{(x - x_j)\pi'(x_j)} f(x_j)$$

where

$$\pi(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$$

$$\pi'(x_j) = (x_j - x_0)(x_j - x_1)\cdots(x_j - x_n)$$

where $(x_j - x_j)$ is the omitted factor. Thus,

$$f(x) = y(x) + E_n(x)$$

$$E_n(x) = \frac{1}{(n+1)!}\pi(x) f^{(n+1)}(\varepsilon)$$

Example. The interpolation polynomial of degree 3 is

$$y(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} f(x_0) + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} f(x_1)$$
$$+ \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} f(x_2) + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} f(x_3)$$

Thus, directly from the data

x	0	1	3	4
$f(x)$	1	1	–1	2

we have as an interpolation polynomial $y(x)$ for (x):

$$y(x) = 1 \cdot \frac{(x-1)(x-3)(x-4)}{(0-1)(0-3)(0-4)} + 1 \cdot \frac{x(x-3)(x-4)}{(1-0)(1-3)(1-4)}$$
$$-1 \cdot \frac{x(x-1)(x-4)}{(3-0)(3-1)(3-4)} + 2 \cdot \frac{(x-0)(x-1)(x-3)}{(4-0)(4-1)(4-3)}$$

Other Difference Methods (Equally Spaced Ordinates)

Backward Differences. The backward differences denoted by

$$\nabla f(x) = f(x) - f(x-h)$$

$$\nabla^2 f(x) = \nabla f(x) - \nabla f(x-h)$$

$$\cdots$$

$$\nabla f^n(x) = \nabla^{n-1} f(x) - \nabla^{n-1} f(x-h)$$

are useful for calculation near the end of tabulated data.

Central Differences. The central differences denoted by

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right)$$

$$\delta^n f(x) = \delta^{n-1} f\left(x + \frac{h}{2}\right) - \delta^{n-1} f\left(x - \frac{h}{2}\right)$$

are useful for calculating at the interior points of tabulated data.
Also to be found in the literature are Gaussian, Stirling, Bessel, Everett, Comrie differences, and so forth.

Inverse Interpolation

This is the process of finding the value of the independent variable or abscissa corresponding to a given value of the function when the latter is between two tabulated values of the abscissa. One method of accomplishing this is to use Lagrange's interpolation formula in the form

$$x = \psi(y) = \sum_{j=0}^{n} \frac{\pi(y)}{(y - y_j)\pi'(y_j)} x_j$$

where x is expressed as a function of y. Other methods revolve about methods of iteration.

Numerical Differentiation

Numerical differentiation should be avoided wherever possible, particularly when data are empirical and subject to appreciable observation errors. Errors in data can affect numerical derivatives quite strongly (i.e., differentiation is a roughening process). When such a calculation must be made, it is usually desirable first to *smooth* the data to a certain extent.

The Use of Interpolation Formulas

If the data are given over equidistant values of the independent variable x, an interpolation formula, such as the Newton formula, may be used, and the resulting formula differentiated analytically. If the independent variable is not at equidistant values, then Lagrange's formulas must be used. By differentiating three- and five-point Lagrange interpolation formulas, the following differentiation formulas result for equally spaced tabular points.

Three-point Formulas. Let x_0, x_1, and x_2 be the three points

$$f'(x_0) = \frac{1}{2h}\left[-3f(x_0) + 4f(x_1) - f(x_2)\right] + \frac{h^2}{3} f'''(\varepsilon)$$

$$f'(x_1) = \frac{1}{2h}\left[-f(x_0) + f(x_2)\right] + \frac{h^2}{6} f'''(\varepsilon)$$

$$f'(x_2) = \frac{1}{2h}\left[f(x_0) - 4f(x_1) + 3f(x_2)\right] + \frac{h^2}{3} f'''(\varepsilon)$$

where the last term is an error term and $\min_j x_j < \varepsilon < \max_j x_j$.

Five-point Formulas. Let x_0, x_1, x_2, x_3, and x_4 be the five values of the equally spaced independent variable and $f_j = f(x_j)$.

$$f'(x_0) = \frac{1}{12h}\left[-25f_0 + 48f_1 - 36f_2 + 16f_3 - 3f_4\right] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

$$f'(x_1) = \frac{1}{12h}\left[-3f_0 - 10f_1 + 18f_2 - 6f_3 + f_4\right] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_2) = \frac{1}{12h}\left[f_0 - 8f_1 + 8f_3 - f_4\right] + \frac{h^4}{30} f^{(v)}(\varepsilon)$$

$$f'(x_3) = \frac{1}{12h}\left[-f_0 + 6f_1 - 18f_2 + 10f_3 + 3f_4\right] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_4) = \frac{1}{12h}\left[3f_0 - 16f_1 + 36f_2 - 48f_3 + 25f_4\right] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

and the last term is again an error term.

Smoothing Techniques

These techniques involve the approximation of the tabular data by a least squares fit of the data using some known functional form, usually a polynomial. In place of approximating $f(x)$ by a single least squares polynomial of degree n over the entire range of the tabulation, it is often desirable to replace each tabulated value by the value taken on by a last squares polynomial of degree n relevant to a subrange of $2M + 1$ points centered, where possible, at the point for which the entry is to be modified. Thus, each smoothed value replaces a tabulated value. Let $f_i = f(x_i)$ be the tabular points and y_j = smoothed values. A first-degree least squares with three points would be

$$y_0 = \tfrac{1}{6}\left[5f_0 + 2f_1 - f_2\right]$$

$$y_1 = \tfrac{1}{3}\left[f_0 + f_1 + f_2\right]$$

$$y_2 = \tfrac{1}{6}\left[-f_0 + 2f_1 + 5f_2\right]$$

A first-degree least squares with five points would be

$$y_0 = \tfrac{1}{5}\left[3f_0 + 2f_1 + f_2 - f_4\right]$$

$$y_1 = \tfrac{1}{10}\left[4f_0 + 3f_1 + 2f_2 + f_3\right]$$

$$y_2 = \tfrac{1}{5}\left[f_0 + f_1 + f_2 + f_3 + f_4\right]$$

$$y_3 = \tfrac{1}{10}\left[f_0 + 2f_1 + 3f_2 + 4f_3\right]$$

$$y_4 = \tfrac{1}{5}\left[-f_0 + f_2 + 2f_3 + 3f_4\right]$$

Thus, for example, if first-degree, five-point least squares are used, the central formula is used for all values except the first two and the last two, where the off-center formulas are used. A third-degree least squares with seven points would be

$$y_0 = \tfrac{1}{42}\left[39f_0 + 8f_1 - 4f_2 - 4f_3 + f_4 + 4f_5 - 2f_6\right]$$

$$y_1 = \tfrac{1}{42}\left[8f_0 + 19f_1 + 16f_2 + 6f_3 - 4f_4 - 7f_5 + 4f_6\right]$$

$$y_2 = \tfrac{1}{42}\left[-4f_0 + 16f_1 + 19f_2 + 12f_3 + 2f_4 - 4f_5 + f_6\right]$$

$$y_3 = \tfrac{1}{21}\left[-2f_0 + 3f_1 + 6f_2 + 7f_3 + 6f_4 + 3f_5 - 2f_6\right]$$

$$y_4 = \tfrac{1}{42}\left[f_0 - 4f_1 + 2f_2 + 12f_3 + 19f_4 + 16f_5 - 4f_6\right]$$

$$y_5 = \tfrac{1}{42}\left[4f_0 - 7f_1 - 4f_2 + 6f_3 + 16f_4 + 19f_5 + 8f_6\right]$$

$$y_6 = \tfrac{1}{42}\left[-2f_0 + 4f_1 + f_2 - 4f_3 - 4f_4 + 8f_5 + 39f_6\right]$$

Additional smoothing formulas may be found in the references. After the data are smoothed, any of the interpolation polynomials, or an appropriate least squares polynomial, may be fitted and the results used to obtain the derivative.

Least Squares Method

Parabolic. For five evenly spaced neighboring abscissas labeled x_{-2}, x_{-1}, x_0, x_1, and x_2, and their ordinates f_{-2}, f_{-1}, f_0, f_1, and f_2, assume a parabola is fit by least squares. There results for all interior points, except the first and last two points of the data, the formula for the numerical derivative:

$$f_0' = \frac{1}{10h}\left[-2f_{-2} - f_{-1} + f_1 + 2f_2\right]$$

For the first two data points designated by 0 and h:

$$f'(0) = \frac{1}{20h}\left[-21f(0) + 13f(h) + 17f(2h) - 9f(3h)\right]$$

$$f'(h) = \frac{1}{20h}\left[-11f(0) + 3f(h) + 7f(2h) + f(3h)\right]$$

and for the last two given by $\alpha - h$ and α:

$$f'(\alpha - h) = \frac{1}{20h}\left[-11f(\alpha) + 3f(\alpha - h) + 7f(\alpha - 2h) + f(\alpha - 3h)\right]$$

$$f'(\alpha) = \frac{1}{20h}\left[-21f(\alpha) + 13f(\alpha - h) + 17f(\alpha - 2h) - 9(\alpha - 3h)\right]$$

Quartic (Douglas-Avakian). A fourth-degree polynomial $y = a + bx + cx^2 + dx^3 + ex^4$ is fitted to seven adjacent equidistant points (spacing h) after a translation of coordinates has been made so that $x = 0$ corresponds to the central point of the seven. Thus, these may be called $-3h$, $-2h$, $-h$, 0, h, $2h$, and $3h$. Let k = coefficient h for the seven points. This is, in $-3h$, $k = -3$. Then, the coefficients for the polynomial are

$$a = \frac{524\sum f(kh) - 245\sum k^2 f(kh) + 21\sum k^4 f(kh)}{924}$$

$$b = \frac{397\sum kf(kh)}{1512h} - \frac{7\sum k^3 f(kh)}{216h}$$

$$c = \frac{-840\sum f(kh) + 679\sum k^2 f(kh) - 67\sum k^4 f(kh)}{3168h^2}$$

$$d = \frac{-7\sum kf(kh) + \sum k^3 f(kh)}{216h^3}$$

$$e = \frac{72\sum f(kh) - 67\sum k^2 f(kh) + 7\sum k^4 f(kh)}{3168h^4}$$

where all summations run from $k = -3$ to $k = +3$ and $f(kh)$ = tabular value at kh. The slope of the polynomial at $x = 0$ is $dy/dx = b$.

Numerical Integration

Numerical evaluation of the finite integral $\int_a^b f(x)\,dx$ is carried out by a variety of methods. A few are given here.

Newton-Cotes Formulas (Equally Spaced Ordinates)

Trapezoidal Rule. This formula consists of subdividing the interval $a \le x \le b$ into n subintervals a to $a + h$, $a + h$ to $a + 2h$, …, and replacing the graph of $f(x)$ by the result of joining the ends of adjacent ordinates by line segments. If $f_j = f(x_j) = f(a + jh)$, $f_0 = f(a)$, and $f_n = f(b)$, the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n\right] + E_n$$

where $|E_n| = (nh^3/12)|f''(\varepsilon)| = [(b - a)^3/12n^2]|f''(\varepsilon)|$, $a < \varepsilon < b$. This procedure is not of high accuracy. However, if $f''(x)$ is continuous in $a < x < b$, the error goes to zero as $1/n^2$, $n \to \infty$.

Parabolic Rule (Simpson's Rule). This procedure consists of subdividing the interval $a < x < b$ into $n/2$ subintervals, each of length $2h$, where n is an even integer. Using the notation as above the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{n-3} + 2f_{n-2} + 4f_{n-1} + f_n\right] + E_n$$

where

$$\left|E_n\right| = \frac{nh^5}{180}\left|f^{(iv)}(\varepsilon)\right| = \frac{(b-a)^3}{180n^4}\left|f^{(iv)}(\varepsilon)\right| \qquad a < \varepsilon < b$$

This method approximates $f(x)$ by a parabola on each subinterval. This rule is generally more accurate than the trapezoidal rule. It is the most widely used integration formula.

Weddle's Rule. This procedure consists of subdividing the integral $a < x < b$ into $n/6$ subintervals, each of length $6h$, where n is a multiple of 6. Using the notation from the trapezoidal rule, there results

$$\int_a^b f(x)\,dx = \frac{3h}{10}\left[f_0 + 5f_1 + f_2 + 6f_3 + f_4 + 5f_5 + 2f_6 + 5f_7 + f_8 + \cdots + 6f_{n-3} + f_{n-2} + 5f_{n-1} + f_n\right] + E_n$$

Note that the coefficients of f_j follow the rule 1, 5, 1, 6, 1, 5, 2, 5, 1, 6, 1, 5, 2, 5, etc.… This procedure consists of approximately $f(x)$ by a polynomial of degree 6 on each subinterval. Here,

$$E_n = \frac{nh^7}{1400}\left[10f^{(6)}(\varepsilon_1) + 9h^2 f^{(8)}(\varepsilon_2)\right]$$

Gaussian Integration Formulas (Unequally Spaced Abscissas)

These formulas are capable of yielding comparable accuracy with fewer ordinates than the equally spaced formulas. The ordinates are obtained by optimizing the distribution of the abscissas rather than by arbitrary choice. For the details of these formulas, Hildebrand (1956) is an excellent reference.

Two-Dimensional Formula

Formulas for two-way integration over a rectangle, circle, ellipse, and so forth, may be developed by a double application of one-dimensional integration formulas. The two-dimensional generalization of the parabolic rule is given here. Consider the iterated integral $\int_a^b \int_c^d f(x, y)\,dx\,dy$. Subdivide $c < x < d$ into m (even) subintervals of length $h = (d - c)/m$, and $a < y < b$ into n (even) subintervals of length $k = (b - a)/n$. This gives a subdivision of the rectangle $a \le y \le b$ and $c \le x \le d$ into subrectangles. Let $x_j = c + jh$, $y_j = a + jk$, and $f_{i,j} = f(x_i, y_j)$. Then,

$$\int_a^b \int_c^d f(x,y)\,dx\,dy = \frac{hk}{9}\left[\left(f_{0,0} + 4f_{1,0} + 2f_{2,0} + \cdots + f_{m,0}\right) + 4\left(f_{0,1} + 4f_{1,1} + 2f_{2,1} + \cdots + f_{m,1}\right)\right.$$
$$\left. + 2\left(f_{0,2} + 4f_{1,2} + 2f_{2,2} + \cdots + f_{m,2}\right) + \cdots + \left(f_{0,n} + 4f_{1,n} + 2f_{2,n} + \cdots + f_{m,n}\right)\right] + E_{m,n}$$

where

$$E_{m,n} = -\frac{hk}{90}\left[mh^4 \frac{\partial^4 f(\varepsilon_1, \eta_1)}{\partial x^4} + nk^4 \frac{\partial^4 f(\varepsilon_2, \eta_2)}{\partial y^4}\right]$$

where ε_1 and ε_2 lie in $c < x < d$, and η_1 and η_2 lie in $a < y < b$.

Numerical Solution of Ordinary Differential Equations

A number of methods have been devised to solve ordinary differential equations numerically. The general references contain some information. A numerical solution of a differential equation means a table of values of the function y and its derivatives over only a limited part of the range of the independent variable. Every differential equation of order n can be rewritten as n first-order differential equations. Therefore, the methods given below will be for first-order equations, and the generalization to simultaneous systems will be developed later.

The Modified Euler Method

This method is simple and yields modest accuracy. If extreme accuracy is desired, a more sophisticated method should be selected. Let the first-order differential equation be $dy/dx = f(x, y)$ with the initial condition (x_0, y_0) (i.e., $y = y_0$ when $x = x_0$). The procedure is as follows.

Step 1. From the given initial conditions (x_0, y_0) compute $y_0' = f(x_0, y_0)$ and $y_0'' = [\partial f(x_0, y_0)/\partial x] + [\partial f(x_0, y_0)/\partial y]\, y_0'$. Then, determine $y_1 = y_0 + h\, y_0' + (h^2/2)\, y_0''$, where h = subdivision of the independent variable.

Step 2. Determine $y_1' = f(x_1, y_1)$ where $x_1 = x_0 + h$. These prepare us for the following.

Predictor Steps.

Step 3. For $n \geq 1$, calculate $(y_{n+1})_1 = y_{n-1} + 2h\, y_n'$.

Step 4. Calculate $(y_{n+1}')_1 = f[x_{n+1}, (y_{n+1})_1]$.

Corrector Steps.

Step 5. Calculate $(y_{n+1})_2 = y_n + (h/2)\,[(y_{n+1}')_1 + y_n']$, where y_n and y_n' without the subscripts are the previous values obtained by this process (or by steps 1 and 2).

Step 6. $(y_{n+1}')_2 = f[x_{n+1}, (y_{n+1})_2]$.

Step 7. Repeat the corrector steps 5 and 6 if necessary until the desired accuracy is produced in y_{n+1}, y_{n+1}'.

Example. Consider the equation $y' = 2y^2 + x$ with the initial conditions $y_0 = 1$ when $x_0 = 0$. Let $h = 0.1$. A few steps of the computation are illustrated.

Step	
1	$y_0' = 2y_0^2 + x_0 = 2$
	$y_0'' = 1 + 4y_0y_0' = 1 + 8 = 9$
	$y_1 = 1 + (0.1)(2) + \left[(0.1)^2/2\right]9 = 1.245$
2	$y_1' = 2y_1^2 + x_1 = 3.100 + 0.1 = 3.200$
3	$(y_2)_1 = y_0 + 2hy_1' = 1 + 2(0.1)3.200 = 1.640$
4	$(y_2')_1 = 2(y_2)_1^2 + x_2 = 5.592$
5	$(y_2)_2 = y_1 + (0.1/2)\left[(y_2')_1 + y_1'\right] = 1.685$
6	$(y_2')_2 = 2(y_2)_2^2 + x_2 = 5.878$
5 (repeat)	$(y_2)_3 = y_1 + (0.05)\left[(y_2')_2 + y_1'\right] = 1.699$
6 (repeat)	$(y_2')_3 = 2(y_2)_3^2 + x_2 = 5.974$

and so forth. This procedure. may be programmed for a computer. A discussion of the truncation error of this process may be found in Milne (1953).

Modified Adam's Method

The procedure given here was developed retaining third differences. It can then be considered as a more exact predictor-corrector method than the Euler method. The procedure is as follows for $dy/dx = f(x, y)$ and h = interval size.

Steps 1 and 2 are the same as in Euler method.

Predictor Steps.

Step 3. $(y_{n+1})_1 = y_n + (h/24)\,[55y_n' - 59y_{n-1}' + 37y_{n-2}' - 9y_{n-3}']$, where y_n', y_{n-1}', etc…, are calculated in step 1.

Step 4. $(y_{n+1}')_1 = f[x_{n+1}, (y_{n+1})_1]$.

Corrector Steps.

Step 5. $(y_{n+1})_2 = y_n + (h/24)\,[9(y'_{n+1})_1 + 19y'_n - 5y'_{n-1} + y'_{n-2}]$.
Step 6. $(y'_{n+1})_2 = f[x_{n+1}, (y_{n+1})_2]$.
Step 7. Iterate steps 5 and 6 if necessary.

Runge-Kutta Methods

These methods are self-starting and are inherently stable. Kopal (1955) is a good reference for their derivation and discussion. Third- and fourth-order procedures are given below for $dy/dx = f(x, y)$ and h = interval size.

For third-order (error $\approx h^4$).

$$k_0 = hf\left(x_n, y_n\right)$$

$$k_1 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_0\right)$$

$$k_2 = hf\left(x_n + h, y_n + 2k_1 - k_0\right)$$

and

$$y_{n+1} = y_n + \tfrac{1}{6}\left(k_0 + 4k_1 + k_2\right)$$

for all $n \geq 0$, with initial condition (x_0, y_0).

For fourth-order (error $\approx h^5$),

$$k_0 = hf\left(x_n, y_n\right)$$

$$k_1 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_0\right)$$

$$k_2 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1\right)$$

$$k_3 = hf\left(x_n + h, y_n + k_2\right)$$

and

$$y_{n+1} = y_n + \tfrac{1}{6}\left(k_0 + 2k_1 + 2k_2 + k_3\right)$$

Example. (Third-order) Let $dy/dx = x - 2y$, with initial condition $y_0 = 1$ when $x_0 = 0$, and let $h = 0.1$. Clearly, $x_n = nh$. To calculate y_1, proceed as follows:

$$k_0 = 0.1\left[x_0 - 2y_0\right] = -0.2$$

$$k_1 = 0.1\left[0.05 - 2(1 - 0.1)\right] = -0.175$$

$$k_2 = 0.1\left[0.1 - 2(1 - 0.35 + 0.2)\right] = -0.16$$

$$y_1 = 1 + \tfrac{1}{6}(-0.2 - 0.7 - 0.16) = 0.8234$$

Equations of Higher Order and Simultaneous Differential Equations

Any differential equation of second- or higher order can be reduced to a simultaneous system of first-order equations by the introduction of auxiliary variables. Consider the following equations:

$$\frac{d^2x}{dt^2} + xy\frac{dx}{dt} + z = e^x$$

$$\frac{d^2y}{dt^2} + xy\frac{dy}{dt} = 7 + t^2$$

$$\frac{d^2z}{dt^2} + xz\frac{dz}{dt} + x = e^x$$

In the new variables $x_1 = x$, $x_2 = y$, $x_3 = z$, $x_4 = dx_1/dt$, $x_5 = dx_2/dt$, and $x_6 = dx_3/dt$, the equations become

$$\frac{dx_1}{dt} = x_4$$

$$\frac{dx_2}{dt} = x_5$$

$$\frac{dx_3}{dt} = x_6$$

$$\frac{dx_4}{dt} = -x_1x_2x_4 - x_3 + e^{x_3}$$

$$\frac{dx_5}{dt} = -x_3x_2x_5 + 7 + t^2$$

$$\frac{dx_6}{dt} = -x_1x_3x_6 - x_1 + e^{x_1}$$

which is a system of the general form

$$\frac{dx_i}{dt} = f_i\left(t, x_1, x_2, x_3, \ldots, x_n\right)$$

where $i = 1, 2, \ldots, n$. Such systems may be solved by simultaneous application of any of the above numerical techniques. A Runge-Kutta method for

$$\frac{dx}{dt} = f(t, x, y)$$

$$\frac{dy}{dt} = g(t, x, y)$$

is given below. The fourth-order procedure is shown.

Starting at the initial conditions x_0, y_0, and t_0, the next values x_1 and y_1 are computed via the equations below (where $\Delta t = h$, $t_j = h + t_{j-1}$):

$$k_0 = hf(t_0, x_0, y_0) \qquad l_0 = hg(t_0, x_0, y_0)$$

$$k_1 = hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right) \qquad l_1 = hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right)$$

$$k_2 = hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right) \qquad l_2 = hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right)$$

$$k_3 = hf(t_0 + h, x_0 + k_2, y_0 + l_2) \qquad l_3 = hg(t_0 + h, x_0 + k_2, y_0 + l_2)$$

and

$$x_1 = x_0 + \tfrac{1}{6}(k_0 + 2k_1 + 2k_2 + k_3)$$

$$y_1 = y_0 + \tfrac{1}{6}(l_0 + 2l_1 + 2l_2 + l_3)$$

To continue the computation, replace t_0, x_0, and y_0 in the above formulas by $t_1 = t_0 + h$, x_1, and y_1 just calculated. Extension of this method to more than two equations follows precisely this same pattern.

Numerical Solution of Integral Equations

This section considers a method of numerically solving the Fredholm integral equation of the second kind:

$$u(x) = f(x) + \lambda \int_a^b k(x,t)u(t)\,dt \qquad \text{for } u(x)$$

The method discussed arises because a definite integral can be closely approximated by any of several numerical integration formulas (each of which arises by approximating the function by some polynomial over an interval). Thus, the definite integral can be replaced by an integration formula which becomes

$$u(x) = f(x) + \lambda(b-a)\left[\sum_{i=1}^{n} c_i k(x, t_i) u(t_i)\right]$$

where $t_1, \ldots, t_n$ are points of subdivision of the t axis, $a \le t \le b$, and the cs are coefficients whose values depend upon the type of numerical integration formula used. Now, this must hold for all values of x, where $a \le x \le b$; so it must hold for $x = t_1$, $x = t_2, \ldots, x = t_n$. Substituting for x successively $t_1, t_2, \ldots, t_n$, and setting $u(t_i) = u_i$ and $f(t_i) = f_i$, we get n linear algebraic equations for the n unknowns $u_1, \ldots, u_n$. That is,

$$u_i = f_i + (b-a)\left[c_1 k(t_i, t_1)u_1 + c_2 k(t_i, t_2)u_2 + \cdots + c_n k(t_i, t_n)u_n\right], \qquad i = 1, 2, \ldots, n$$

These u_j may be solved for by the methods under the section entitled "Numerical Solution of Linear Equations."

Numerical Methods for Partial Differential Equations

The ultimate goal of numerical (discrete) methods for partial differential equations (PDEs) is the reduction of continuous systems (projections) to discrete systems that are suitable for high-speed computer solutions. The user must be cautioned that the seeming elementary nature of the techniques holds pitfalls that can be seriously misleading. These approximations often lead to difficult mathematical questions of adequacy, accuracy, convergence, stability, and consistency. Convergence is concerned with

the approach of the approximate numerical solution to the exact solution as the number of mesh units increase indefinitely in some sense. Unless the numerical method can be shown to converge to the exact solution, the chosen method is unsatisfactory.

Stability deals in general with error growth in the calculation. As stated before, any numerical method involves truncation and round-off errors. These errors are not serious unless they grow as the computation proceeds (i.e., the method is unstable).

Finite Difference Methods

In these methods, the derivatives are replaced by various finite differences. The methods will be illustrated for problems in two space dimensions (x, y) or (x, t) where t is timelike. Using subdivisions $\Delta x = h$ and $\Delta y = k$ with $u(i\,h, jk) = u_{i,j}$, approximate $u_x|_{i,j} = [(u_{i+1,j} - u_{i,j})/h] + O(h)$ (forward difference), a first-order $[O(h)]$ method, or $u_x|_{i,j} = [(u_{i+1,j} - u_{i-1,j})/2h] + O(h^2)$ (central difference), a second-order method. The second derivative is usually approximated with the second-order method $[u_{xx}|_{i,j} = [(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})/h^2] + O(h^2)]$.

Example. Using second-order differences for u_{xx} and u_{yy}, the five-point difference equation (with $h = k$) for Laplace's equation $u_{xx} + u_{yy} = 0$ is $u_{i,j} = 1/4[u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}]$. The accuracy is $O(h^2)$. This model is called *implicit* because one must solve for the total number of unknowns at the unknown grid points (i, j) in terms of the given boundary data. In this case, the system of equations is a linear system.

Example. Using a forward-difference approximation for u_t and a second-order approximation for u_{xx}, the diffusion equation $u_t = u_{xx}$ is approximated by the *explicit* formula $u_{i,j+1} = ru_{i-1,j} + (1 - 2r)u_{i,j} + ru_{i+1,j}$. This classic result permits step-by-step advancement in the t direction beginning with the initial data at $t = 0$ $(j = 0)$ and guided by the boundary data. Here, the term $r = \Delta t/(\Delta x)^2 = k/h^2$ is restricted to be less than or equal to 1/2 for stability and the truncation error is $O(k^2 + kh^2)$.

The Crank-Nicolson implicit formula which approximates the diffusion equation $u_t = u_{xx}$ is

$$-r\lambda u_{i-1,j+1} + (1 + 2r\lambda)u_{i,j+1} - r\lambda u_{i+1,j+1} = r(1-\lambda)u_{i-1,j} + [1 - 2r(1-\lambda)]u_{i,j} + r(1-\lambda)u_{i+1,j}$$

The stability of this numerical method was analyzed by Crandall (Ames, 1993) where the λ, r stability diagram is given.

Approximation of the time derivative in $u_t = u_{xx}$ by a central difference leads to an always unstable approximation — the useless approximation

$$u_{i,j+1} = u_{i,j-1} + 2r(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})$$

which is a warning to be careful.

The foregoing method is *symmetric* with respect to the point (i, j), where the method is centered. Asymmetric methods have some computational advantages, so the Saul'yev method is described (Ames, 1993). The algorithms $(r = k/h^2)$

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i-1,j+1} - u_{i,j} + u_{i+1,j}) \qquad \text{(Saul' yev A)}$$

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i+1,j+1} - u_{i,j} + u_{i-1,j}) \qquad \text{(Saul' yev B)}$$

are used as in any one of the following options:

1. Use Saul'yev A only and proceed line-by-line in the $t(j)$ direction, but *always* from the left boundary on a line.
2. Use Saul'yev B only and proceed line-by-line in the $t(j)$ direction, but *always* from the right boundary to the left on a line.

3. Alternate from line to line by first using Saul'yev A and then B, or the reverse. This is related to *alternating direction methods.*
4. Use Saul'yev A and Saul'yev B on the same line and average the results for the final answer (A first, and then B). This is equivalent to introducing the dummy variables $P_{i,j}$ and $Q_{i,j}$ such that

$$(1+r)P_{i,j+1} = U_{i,j} + r\left(P_{i-1,j+1} - U_{i,j} + U_{i+1,j}\right)$$

$$(1+r)Q_{i,j+1} = U_{i,j} + r\left(Q_{i+1,j+1} - U_{i,j} + U_{i-1,j}\right)$$

and

$$U_{i,j+1} = \tfrac{1}{2}\left(P_{i,j+1} + Q_{i,j+1}\right)$$

This averaging method has some computational advantage because of the possibility of truncation error cancellation. As an alternative, one can retain the $P_{i,j}$ and $Q_{i,j}$ from the previous step and replace $U_{i,j}$ and $U_{i+1,j}$ by $P_{i,j}$ and $P_{i+1,j}$, respectively, and $U_{i,j}$ and $U_{i-1,j}$ by $Q_{i,j}$ and $Q_{i-1,j}$, respectively.

Weighted Residual Methods (WRMs)

To set the stage for the method of finite elements, we briefly describe the WRMs, which have several variations — the interior, boundary, and mixed methods. Suppose the equation is $Lu = f$, where L is the partial differential operator and f is a known function, of say x and y. The first step in WRM is to select a class of known basis functions b_i (e.g., trigonometric, Bessel, Legendre) to approximate $u(x, y)$ as $\sim \Sigma\, a_i\, b_i\, (x, y) = U(x, y, a)$. Often, the b_i are selected so that $U(x, y, a)$ satisfy the boundary conditions. This is essentially the *interior method*. If the b_i, in $U(x, y, a)$ are selected to satisfy the differential equations, but not the boundary conditions, the variant is called the *boundary method*. When neither the equation nor the boundary conditions are satisfied, the method is said to be *mixed*. The least ingenuity is required here. The usual method of choice is the interior method.

The second step is to select an optimal set of constants a_i, $i = 1,2, \ldots, n$, by using the residual $R_I(U) = LU - f$. This is done here for the interior method. In the boundary method, there are a set of boundary residual R_B, and, in the mixed method. Both R_I and R_B. Using the spatial average $(w, v) = \int_v wvdV$, the criterion for selecting the values of a_i is the requirement that the n spatial averages

$$\left(b_i, R_E(U)\right) = 0, \qquad i = 1,2,\ldots,n$$

These represent n equations (linear if the operator L is linear and nonlinear otherwise) for the a_i.

Particular WRMs differ because of the choice of the b_js. The most common follow.

1. *Subdomain* The domain V is divided into n smaller, not necessarily disjoint, subdomains V_j with $w_j(x, y) = 1$ if (x, y) is in V_j, and 0 if (x, y) is not in V_j.
2. *Collocation* Select n points $P_j = (x_j, y_j)$ in V with $w_j(P_j) = \delta(P - P_j)$, where $\int_v \phi(P)\delta(P - P_j)dP = \phi(P_j)$ for all test functions $\phi(P)$ which vanish outside the compact set V. Thus, $(w_j, R_E) = \int_v \delta(P - P_j)R_E dV = R_E[U(P_j) \equiv 0$ (i.e., the residual is set equal to zero at the n points P_j).
3. *Least squares* Here, the functional $I(a) = \int_V R_E^2\, dV$, where $a = (a_1, \ldots, a_n)$, is to be made stationary with respect to the a_j. Thus, $0 = \partial I/\partial a_j = 2\int_v R_E(\partial R_E/\partial a_j)dV$, with $j = 1, 2, \ldots, n$. The w_j in this case are $\partial R_E/\partial a_j$.
4. *Bubnov-Galerkin* Choose $w_j(P) = b_j(P)$. This is perhaps the best-known method.
5. *Stationary Functional (Variational) Method* With ϕ a variational integral (or other functional), set $\partial\phi[U]/\partial a_j = 0$, where $j = 1, \ldots, n$, to generate the n algebraic equations.

Example. $u_{xx} + u_{yy} = -2$, with $u = 0$ on the boundaries of the square $x = \pm 1$, $y = \pm 1$. Select an interior method with $U = a_1(1 - x^2)(1 - y^2) + a_2x^2y^2(1 - x^2)(1 - y^2)$, whereupon the residual $R_E(U) = 2a_1(2 - x^2 - y^2) + 2a_2[(1 - 6x^2)y^2(1 - y^2) + (1 - 6y^2) - (1 - x^2)] + 2$. Collocating at (1/3, 1/3) and (2/3, 2/3) gives the two linear equations $-32a_1/9 + 32a_2/243x^2 + 2 = 0$ and $-20a_1/9 - 400a_2/243 + 2 = 0$ for a_1 and a_2.

WRM methods can obviously be used as approximate methods. We have now set the stage for *finite elements.*

Finite Elements

The WRM methods are more general than the *finite elements* (FE) methods. FE methods require, in addition, that the basis functions be finite elements (i.e., functions that are zero except on a small part of the domain under consideration). A typical example of an often used basis is that of triangular elements. For a triangular element with Cartesian coordinates (x_1, y_1), (x_2, y_2), and (x_3, y_3), define natural coordinates L_1, L_2, and L_3 (L_i, $\leftrightarrow (x_i, y_i)$) so that $L_i = A_i/A$ where

$$A = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

is the area of the triangle and

$$A_1 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_2 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x & y \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_3 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x & y \end{bmatrix}$$

Clearly $L_1 + L_2 + L_3 = 1$, and the L_i are one at node i and zero at the other nodes. In terms of the Cartesian coordinates,

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \frac{1}{2A}\begin{bmatrix} x_2y_3 - x_3y_2, & y_2 - y_3, & x_3 - x_2 \\ x_3y_1 - x_1y_3, & y_3 - y_1, & x_1 - x_3 \\ x_1y_2 - x_2y_1, & y_1 - y_2, & x_2 - x_1 \end{bmatrix}\begin{bmatrix} 1 \\ x \\ y \end{bmatrix}$$

is the linear triangular element relation.

Tables of linear, quadratic, and cubic basis functions are given in the literature. Notice that while the linear basis needs three nodes, the quadratic requires six and the cubic basis ten. Various modifications, such as the Hermite basis, are described in the literature. Triangular elements are useful in approximating irregular domains.

For rectangular elements, the *chapeau* functions are often used. Let us illustrate with an example. Let $u_{xx} + u_{yy} = Q$, $0 < x < 2$, $0 < y < 2$, $u(x, 2) = 1$, $u(0, y) = 1$, $u_y(x, 0) = 0$, $u_x(2, y) = 0$, and $Q(x, y) = Qw\delta(x - 1)\delta(y - 1)$,

$$\delta(x) = \begin{cases} 0 & x \neq 0 \\ 1 & x = 0 \end{cases}$$

Using four equal rectangular elements, map the element I with vertices at (0, 0) (0, 1), (1, 1), and (1, 0) into the local (canonical) coordinates (ξ, η), $-1 \le \xi \le 1$, $-1 \le \eta \le 1$, by means of $x = 1/2(\xi + 1)$, $y = 1/2(\eta + 1)$. This mapping permits one to develop software that standardizes the treatment of all elements. Converting to (ξ, η) coordinates, our problem becomes $u_{\xi\xi} + u_{\eta\eta} = 1/4Q$, $-1 \le \xi \le 1$, $-1 \le \eta \le 1$, $Q = Qw\delta(\xi - 1)\delta(\eta - 1)$.

First, a trial function $\bar{u}(\xi, \eta)$ is defined as $u(\xi, \eta) \approx \bar{\mu}(\xi, \eta) = \Sigma_{j=1}^{4} A_j \phi_j(\xi, \eta)$ (in element I) where the ϕ_j are the two-dimensional chapeau functions

$$\phi_1 = \left[\tfrac{1}{2}(1-\xi)\tfrac{1}{2}(1-\eta)\right] \qquad \phi_2 = \left[\tfrac{1}{2}(1+\xi)\tfrac{1}{2}(1-\eta)\right]$$

$$\phi_3 = \left[\tfrac{1}{2}(1+\xi)\tfrac{1}{2}(1+\eta)\right] \qquad \phi_4 = \left[\tfrac{1}{2}(1-\xi)\tfrac{1}{2}(1+\eta)\right]$$

Clearly ϕ_i take the value one at node i, provide a bilinear approximation, and are nonzero only over elements adjacent to node i.

Second, the equation residual $R_E = \nabla^2 \bar{u} - 1/4Q$ is formed and a WRM procedure is selected to formulate the algebraic equations for the A_i. This is indicated using the Galerkin method. Thus, for element I, we have

$$\iint_{D_I} \left(\bar{u}_{\xi\xi} + \bar{u}_{\eta\eta} - Q\right)\phi_i(\xi,\eta)\, d\xi\, d\eta = 0, \qquad i = 1,\ldots,4$$

Applying Green's theorem, this result becomes

$$\iint_{D_I} \left[\bar{u}_\xi(\phi_i)_\xi + \bar{u}_\eta(\phi_i)_\eta + \tfrac{1}{4}Q\phi_i\right] d\xi\, d\eta - \int_{\partial D_I} \left(\bar{u}_\xi c_\xi + \bar{u}_\eta c_\eta\right)\phi_i\, ds = 0, \qquad i = 1,2,\ldots,4$$

Using the same procedure in all four elements and recalling the property that the ϕ_i in each element are nonzero only over elements adjacent to node i gives the following nine equations:

$$\sum_{e=1}^{4} \left\{ \iint_{D_e} \sum_{j=1}^{9} A_j \left[(\phi_j)_\xi(\phi_i)_\xi + (\phi_j)_\eta(\phi_i)_\eta\right] + \tfrac{1}{4}Q\phi_i \right\} d\xi\, d\eta$$

$$- \sum_{e=1}^{4} \int_{\partial D_e} \left(\bar{u}_\xi c_\xi + \bar{u}_\eta c_\eta\right)\phi\, ds = 0, \quad n = 1,2,\ldots,9$$

where the c_ξ and c_η are the direction cosines of the appropriate element (e) boundary.

Method of Lines

The *method of lines*, when used on PDEs in two dimensions, reduces the PDE to a system of ordinary differential equations (ODEs), usually by finite difference or finite element techniques. If the original problem is an initial value (boundary value) problem, then the resulting ODEs form an initial value (boundary value) problem. These ODEs are solved by ODE numerical methods.

Example. $u_t = u_{xx} + u^2$, $0 < x < 1$, $0 < t$, with the initial value $u(x, 0) = x$, and boundary data $u(0, t) = 0$, $u(1, t) = \sin t$. A discretization of the space variable (x) is introduced and the time variable is left continuous. The approximation is $\dot{u}_i = (u_{i+1} - 2u_i + u_{i-1})/h^2 + u_i^2$. With $h = 1/5$, the equations become

$$u_0(t) = 0$$

$$\dot{u}_1 = \tfrac{1}{25}\left[u_2 - 2u_1\right] + u_1^2$$

$$\dot{u}_2 = \tfrac{1}{25}\left[u_3 - 2u_2 + u_1\right] + u_2^2$$

$$\dot{u}_3 = \tfrac{1}{25}\left[u_4 - 2u_3 + u_2\right] + u_3^2$$

$$\dot{u}_4 = \tfrac{1}{25}\left[\sin t - 2u_4 + u_3\right] + u_4^2$$

$$u_5 = \sin t$$

and $u_1(0) = 0.2$, $u_2(0) = 0.4$, $u_3(0) = 0.6$, and $u_4(0) = 0.8$.

Discrete and Fast Fourier Transforms

Let $x(n)$ be a sequence that is nonzero only for a finite number of samples in the interval $0 \le n \le N - 1$. The quantity

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i(2\pi/N)nk}, \quad k = 0,1,\ldots,N-1$$

is called the *discrete Fourier transform* (DFT) of the sequence $x(n)$. Its inverse (IDFT) is given by

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{i(2\pi/N)nk}, \quad n = 0,1,\ldots,N-1 \quad \left(i^2 = -1\right)$$

Clearly, DFT and IDFT are finite sums and there are N frequency values. Also, $X(k)$ is periodic in k with period N.

Example. $x(0) = 1$, $x(1) = 2$, $x(2) = 3$, $x(3) = 4$

$$X(k) = \sum_{n=0}^{3} x(n) e^{-i(2\pi/4)nk}, \quad k = 0,\ 1,\ 2,\ 3,\ 4$$

Thus,

$$X(0) = \sum_{n=0}^{3} x(n) = 10$$

and $X(1) = x(0) + x(1)e^{-i\pi/2} + x(2)e^{-i\pi} + x(3)e^{-i3\pi/2} = 1 - 2i - 3 + 4i = -2 + 2i$; $X(2) = -2$; $X(3) = -2 - 2i$.

DFT Properties

1. Linearity: If $x_3(n) = ax_1(n) + bx_2(n)$, then $X_3(k) = aX_1(k) + bX_2(k)$.
2. Symmetry: For $x(n)$ real, $\mathrm{Re}[X(k)] = \mathrm{Re}[X(N - k)]$, $\mathrm{Im}[(k)] = -\mathrm{Im}[X(N - k)]$.

3. Circular shift: By a circular shift of a sequence defined in the interval $0 \le n \le N-1$, we mean that, as values *fall off* from one end of the sequence, they are appended to the other end. Denoting this by $x(n \oplus m)$, we see that positive m means shift left and negative m means shift right. Thus, $x_2(n) = x_1(n \oplus m) \Leftrightarrow X_2(k) = X_1(k)e^{i(2\pi/N)km}$.
4. Duality: $x(n) \Leftrightarrow X(k)$ implies $(1/N)X(n) \Leftrightarrow x(-k)$.
5. Z-transform relation: $X(k) = X(z)|_{z=e^{i(2\pi k/N)}}$, $k = 0,1, \ldots, N-1$.
6. Circular convolution: $x_3(n) = \sum_{m=0}^{N-1} x_1(m)x_2(n \ominus m) = \sum_{\ell=0}^{N-1} x_1(n \ominus \ell)x_2(\ell)$ where $x_2(n \ominus m)$ corresponds to a circular shift to the right for positive m.

One fast algorithm for calculating DFTs is the radix-2 *fast Fourier transform* developed by J. W. Cooley and J. W. Tucker. Consider the two-point DFT $X(k) = \sum_{n=0}^{1} x(n)e^{-i(2\pi/2)nk}$, $k = 0, 1$. Clearly, $X(k) = x(0) + x(1)e^{-i\pi k}$. So, $X(0) = x(0) + x(1)$ and $X(1) = x(0) - x(1)$. This process can be extended to DFTs of length $N = 2^r$, where r is a positive integer. For $N = 2^r$, decompose the N-point DFT into *two* $N/2$-point DFTs. Then, decompose each $N/2$-point DFT into *two* $N/4$-point DFTs, and so on until eventually we have $N/2$ *two*-point DFTs. Computing these as indicated above, we combine them into $N/4$ four-point DFTs and then $N/8$ eight-point DFTs, and so on, until the DFT is computed. The total number of DFT operations (for large N) is $O(N^2)$, and that of the FFT is $O(N \log_2 N)$, quite a saving for large N.

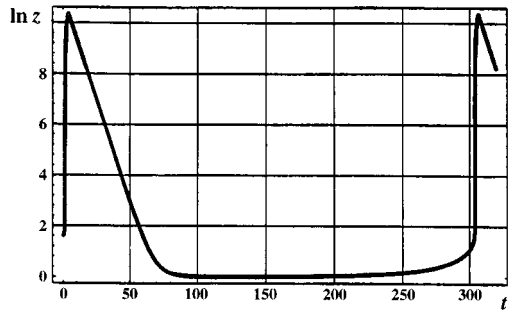

Figure 19.12.1

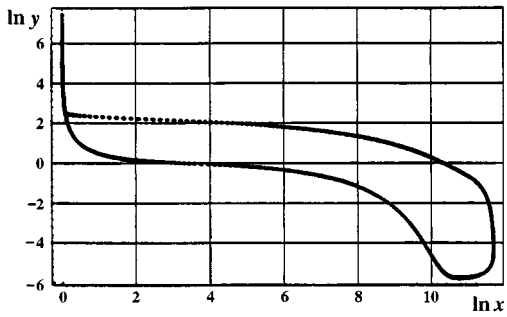

Figure 19.12.2

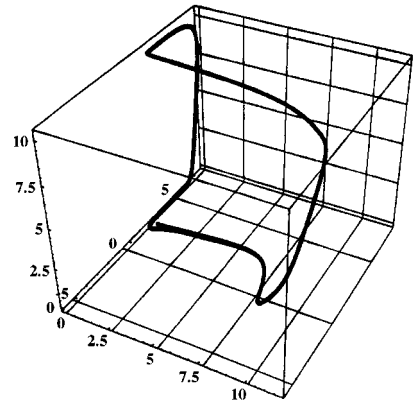

Figure 19.12.3

FIGURES 19.12.1 to 19.12.3 The "Oregonator" is a periodic chemical reaction describable by three nonlinear first-order differential equations. The results (Figure 19.12.1) illustrate the periodic nature of the major chemical versus time. Figure 19.12.2 shows the phase diagram of two of the reactants, and Figure 19.12.3 is the three-dimensional phase diagram of all reactants. The numerical computation was done using a fourth-order Runge-Kuta method on Mathematica by Waltraud Rufeger at the Georgia Institute of Technology.

Software

Some available software is listed here.

General Packages

General software packages include Maple, Mathematica, and Matlab. All contain algorithms for handling a large variety of both numerical and symbolic computations.

Special Packages for Linear Systems

In the IMSL Library, there are three complementary linear system packages of note.

LINPACK is a collection of programs concerned with *direct* methods for general (or full) symmetric, symmetric positive definite, triangular, and tridiagonal matrices. There are also programs for least squares problems, along with the QR algorithm for eigensystems and the singular value decompositions of rectangular matrices. The programs are intended to be completely machine independent, fully portable, and run with good efficiency in most computing environments. The LINPACK User's Guide by Dongarra *et al.* is the basic reference.

ITPACK is a modular set of programs for iterative methods. The package is oriented toward the sparse matrices that arise in the solution of PDEs and other applications. While the programs apply to full matrices, that is rarely profitable. Four basic iteration methods and two convergence acceleration methods are in the package. There is a Jacobi, SOR (with optimum relaxation parameter estimated), symmetric SOR, and reduced system (red-black ordering) iteration, each with semi-iteration and conjugate gradient acceleration. All parameters for these iterations are automatically estimated. The practical and theoretical background for ITPACK is found in Hagemen and Young (1981).

YALEPACK is a substantial collection of programs for sparse matrix computations.

Ordinary Differential Equations Packages

Also in IMSL, one finds such sophisticated software as DVERK, DGEAR, or DREBS for initial value problems. For two-point boundary value problems, one finds DTPTB (use of DVERK and multiple shooting) or DVCPR.

Partial Differential Equations Packages

DISPL was developed and written at Argonne National Laboratory. DISPL is designed for nonlinear second-order PDEs (parabolic, elliptic, hyperbolic (some cases), and parabolic-elliptic). Boundary conditions of a general nature and material interfaces are allowed. The spatial dimension can be either one or two and in Cartesian, cylindrical, or spherical (one dimension only) geometry. The PDEs are reduced to ordinary DEs by Galerkin discretization of the spatial variables. The resulting ordinary DEs in the timelike variable are then solved by an ODE software package (such as GEAR). Software features include graphics capabilities, printed output, dump/restart/facilities, and free format input. DISPL is intended to be an engineering and scientific tool and is not a finely tuned production code for a small set of problems. DISPL makes no effort to control the spatial discretization errors. It has been used to successfully solve a variety of problems in chemical transport, heat and mass transfer, pipe flow, etc.

PDELIB was developed and written at Los Alamos Scientific Laboratory. PDELIB is a library of subroutines to support the numerical solution of evolution equations with a timelike variable and one or two space variables. The routines are grouped into a dozen independent modules according to their function (i.e., accepting initial data, approximating spatial derivatives, advancing the solution in time). Each task is isolated in a distinct module. Within a module, the basic task is further refined into general-purpose flexible lower-level routines. PDELIB can be understood and used at different levels. Within a small period of time, a large class of problems can be solved by a novice. Moreover, it can provide a wide variety of outputs.

DSS/2 is a differential systems simulator developed at Lehigh University as a transportable numerical method of lines (NMOL) code. See also LEANS.

FORSIM is designed for the automated solution of sets of implicitly coupled PDEs of the form

$$\frac{\partial u_i}{\partial t} = \phi_i\left(x, t, u_i, u_j, \ldots, (u_i)_x, \ldots, (u_i)_{xx}, (u_j)_{xx}, \ldots\right), \quad \text{for } i = 1, \ldots, N$$

The user specifies the ϕ_i in a simple FORTRAN subroutine. Finite difference formulas of any order may be selected for the spatial discretization and the spatial grid need not be equidistant. The resulting system of time-dependent ODEs is solved by the method of lines.

SLDGL is a program package for the self-adaptive solution of nonlinear systems of elliptic and parabolic PDEs in up to three space dimensions. Variable step size and variable order are permitted. The discretization error is estimated and used for the determination of the optimum grid and optimum orders. This is the most general of the codes described here (not for hyperbolic systems, of course). This package has seen extensive use in Europe.

FIDISOL (finite difference solver) is a program package for nonlinear systems of two-or three-dimensional elliptic and parabolic systems in rectangular domains or in domains that can be transformed analytically to rectangular domains. This package is actually a redesign of parts of SLDGL, primarily for the solution of large problems on vector computers. It has been tested on the CYBER 205, CRAY-IM, CRAY X-MP/22, and VP 200. The program vectorizes very well and uses the vector arithmetic efficiently. In addition to the numerical solution, a reliable error estimate is computed.

CAVE is a program package for conduction analysis via eigenvalues for three-dimensional geometries using the method of lines. In many problems, much time is saved because only a few terms suffice.

Many industrial and university computing services subscribe to the IMSL Software Library. Announcements of new software appear in *Directions*, a publication of IMSL. A brief description of some IMSL packages applicable to PDEs and associated problems is now given. In addition to those packages just described, two additional software packages bear mention. The first of these, the ELLPACK system, solves elliptic problems in two dimensions with general domains and in three dimensions with box-shaped domains. The system contains over 30 numerical methods modules, thereby providing a means of evaluating and comparing different methods for solving elliptic problems. ELLPACK has a special high-level language making it easy to use. New algorithms can be added or deleted from the system with ease.

Second, TWODEPEP is IMSL's general finite element system for two-dimensional elliptic, parabolic, and eigenvalue problems. The Galerkin finite elements available are triangles with quadratic, cubic, or quartic basic functions, with one edge curved when adjacent to a curved boundary, according to the isoparametric method. Nonlinear equatons are solved by Newton's method, with the resulting linear system solved directly by Gauss elimination. PDE/PROTRAN is also available. It uses triangular elements with piecewise polynomials of degree 2, 3, or 4 to solve quite general steady state, time-dependent, and eigenvalue problems in general two-dimensional regions. There is a simple user input. Additional information may be obtained from IMSL. NASTRAN and STRUDL are two advanced finite element computer systems available from a variety of sources. Another, UNAFEM, has been extensively used.

References

General

Adams, E. and Kulisch, U. (Eds.) 1993. *Scientific Computing with Automatic Result Verification*, Academic Press, Boton, MA.

Gerald, C. F. and Wheatley, P. O. 1984. *Applied Numerical Analysis*, Addison-Wesley, Reading, MA.

Hamming, R. W. 1962. *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York.

Hildebrand, F. B. 1956. *Introduction to Numerical Analysis*, McGraw-Hill, New York.

Isaacson, E. and Keller, H. B. 1966. *Analysis of Numerical Methods*, John Wiley & Sons, New York.

Kopal, Z. 1955. *Numerical Analysis*, John Wiley & Sons, New York.

Rice, J. R. 1993. *Numerical Methods, Software and Analysis*, 2nd ed. Academic Press, Boston, MA.

Stoer, J. and Bulirsch, R. 1976. *Introduction to Numerical Analysis*, Springer, New York.

Linear Equations

Bodewig, E. 1956. *Matrix Calculus*, Wiley (Interscience), New York.
Hageman, L. A. and Young, D. M. 1981. *Applied Iterative Methods*, Academic Press, Boston, MA.
Varga, R. S. 1962. *Matrix Iterative Numerical Analysis*, John Wiley & Sons, New York.
Young, D. M. 1971. *Iterative Solution of Large-Linear Systems*, Academic Press, Boston, MA.

Ordinary Differential Equations

Aiken, R. C. 1985. *Stiff Computation*, Oxford Unitersity Press, New York.
Gear, C. W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, NJ.
Keller, H. B. 1976. *Numerical Solutions of Two Point Boundary Value Problems*, SIAM, Philadelphia, PA.
Lambert, J. D. 1973. *Computational Methods in Ordinary Differential Equations*, Cambridge University Press, New York.
Milne, W.E. 1953. *Numerical Solution of Differential Equations*, John Wiley & Sons, New York.
Rickey, K. C., Evans, H. R., Griffiths, D. W., and Nethercot, D. A. 1983. *The Finite Element Method — A Basic Introduction for Engineers*, 2nd ed. Halstead Press, New York.
Shampine, L. and Gear, C. W. 1979. A User's View of Solving Stiff Ordinary Differential Equatons, *SIAM Rev.* 21:1–17.

Partial Differential Equations

Ames, W. F. 1993. *Numerical Methods for Partial Differential Equations*, 3d ed. Academic Press, Boston, MA.
Brebbia, C. A. 1984. *Boundary Element Techniques in Computer Aided Engineering*, Martinus Nijhoff, Boston, MA.
Burnett, D. S. 1987. *Finite Element Analysis*, Addison-Wesley, Reading, MA.
Lapidus, L. and Pinder, G. F. 1982. *Numerical Solution of Partial Differential Equations in Science and Engineering*, John Wiley & Sons, New York.
Roache, P. 1972. *Computational Fluid Dynamics*, Hermosa, Albuquerque, NM.

19.13 Experimental Uncertainty Analysis

W.G. Steele and H.W. Coleman

Introduction

The goal of an experiment is to answer a question by measuring a specific variable, X_i, or by determining a result, r, from a functional relationship among measured variables

$$r = r\left(X_1, X_2, \ldots, X_i, \ldots, X_J\right) \tag{19.13.1}$$

In all experiments there is some error that prevents the measurement of the true value of each variable, and therefore, prevents the determination of r_{true}.

Uncertainty analysis is a technique that is used to estimate the interval about a measured variable or a determined result within which the true value is thought to lie with a certain degree of confidence. As discussed by Coleman and Steele (1989), uncertainty analysis is an extremely useful tool for all phases of an experimental program from initial planning (general uncertainty analysis) to detailed design, debugging, test operation, and data analysis (detailed uncertainty analysis).

The application of uncertainty analysis in engineering has evolved considerably since the classic paper of Kline and McClintock (1953). Developments in the field have been especially rapid and significant over the past decade, with the methods formulated by Abernethy and co-workers (1985) that were incorporated into ANSI/ASME Standards in (1984) and (1986) being superseded by the more rigorous approach presented in the International Organization for Standardization (ISO) *Guide to the Expression of Uncertainty in Measurement* (1993). This guide, published in the name of ISO and six other international organizations, has in everything but name established a new international experimental uncertainty standard.

The approach in the ISO *Guide* deals with "Type A" and "Type B" categories of uncertainties, not the more traditional engineering categories of systematic (bias) and precision (random) uncertainties, and is of sufficient complexity that its application in normal engineering practice is unlikely. This issue has been addressed by AGARD Working Group 15 on Quality Assessment for Wind Tunnel Testing, by the Standards Subcommittee of the AIAA Ground Test Technical Committee, and by the ASME Committee PTC 19.1 that is revising the ANSI/ASME Standard (1986). The documents issued by two of these groups (AGARD-AR-304, 1994) and (AIAA Standard S-071-1995, 1995) and in preparation by the ASME Committee present and discuss the additional assumptions necessary to achieve a less complex "large sample" methodology that is consistent with the ISO *Guide*, that is applicable to the vast majority of engineering testing, including most single-sample tests, and that retains the use of the traditional engineering concepts of systematic and precision uncertainties. The range of validity of this "large sample" approximation has been presented by Steele et al. (1994) and by Coleman and Steele (1995). The authors of this section are also preparing a second edition of Coleman and Steele (1989), which will incorporate the ISO *Guide* methodology and will illustrate its use in all aspects of engineering experimentation.

In the following, the uncertainties of individual measured variables and of determined results are discussed. This section concludes with an overview of the use of uncertainty analysis in all phases of an experimental program.

Uncertainty of a Measured Variable

For a measured variable, X_i, the total error is caused by both precision (random) and systematic (bias) errors. This relationship is shown in Figure 19.13.1. The possible measurement values of the variable are scattered in a distribution (here assumed Gaussian) around the parent population mean, μ_i. The parent population mean differs from $(X_i)_{true}$ by an amount called the systematic (or bias) error, β_i. The quantity

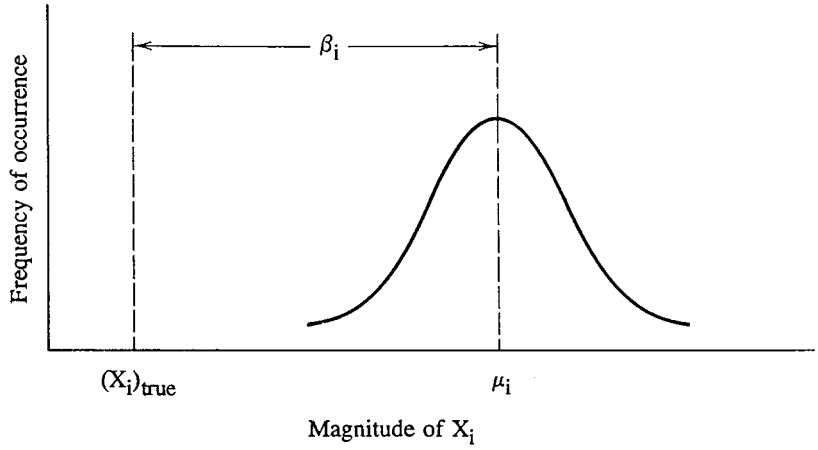

FIGURE 19.13.1 Errors in the measurement of variable X_i.

β_i is the total fixed error that remains in the measurement process after all calibration corrections have been made. In general, there will be several sources of bias error such as calibration standard errors, data acquisition errors, data reduction errors, and test technique errors. There is usually no direct way to measure these errors, so they must be estimated.

For each bias error source, $(\beta_i)_k$, the experimenter must estimate a systematic uncertainty (or bias limit), $(B_i)_k$, such that there is about a 95% confidence that $(B_i)_k \geq |(\beta_i)_k|$. Systematic uncertainties are usually estimated from previous experience, calibration data, analytical models, and the application of sound engineering judgment. For each variable, there will be a set, K_i, of elemental systematic uncertainties, $(B_i)_k$, for the significant fixed error sources. The overall systematic uncertainty for variable X_i is determined from these estimates as

$$B_i^2 = \sum_{k=1}^{K_i} \left(B_i\right)_k^2 \tag{19.13.2}$$

For a discussion on estimating systematic uncertainties (bias limits), see Coleman and Steele (1989).

The estimate of the precision error for a variable is the sample standard deviation, or the estimate of the error associated with the repeatability of a particular measurement. Unlike the systematic error, the precision error varies from reading to reading. As the number of readings, N_i, of a particular variable tends to infinity, the distribution of these readings becomes Gaussian.

The readings used to calculate the sample standard deviation for each variable must be taken over the time frame and conditions which cover the variation in the variable. For example, taking multiple samples of data as a function of time while holding all other conditions constant will identify the random variation associated with the measurement system and the unsteadiness of the test condition. If the sample standard deviation of the variable being measured is also expected to be representative of other possible variations in the measurement, e.g., repeatability of test conditions, variation in test configuration, then these additional error sources will have to be varied while the multiple data samples are taken to determine the standard deviation.

When the value of a variable is determined as the mean, $\overline{X}_i$, of N_i readings, then the sample standard deviation of the mean is

$$S_{\overline{X}_i} = \left\{ \left[\frac{1}{N_i\left(N_i - 1\right)} \right] \sum_{k=1}^{N_i} \left[\left(X_i\right)_k - \overline{X}_i \right]^2 \right\}^{1/2} \tag{19.13.3}$$

where

$$\overline{X}_i = \frac{\sum_{k=1}^{N_i} \left(X_i\right)_k}{N_i} \tag{19.13.4}$$

It must be stressed that these N_i readings have to be taken over the appropriate range of variations for X_i as described above.

When only a single reading of a variable is available so that the value used for the variable is X_i, then N_{P_i} previous readings, $(X_{Pi})_k$, must be used to find the standard deviation for the variable as

$$S_{X_i} = \left\{ \frac{1}{N_{Pi} - 1} \sum_{k=1}^{N_{P_i}} \left[\left(X_{P_i}\right)_k - \overline{X}_{P_i} \right]^2 \right\}^{1/2} \tag{19.13.5}$$

where

$$\overline{X}_{P_i} = \frac{1}{N_{P_i}} \sum_{k=1}^{N_{P_i}} \left(X_{P_i}\right)_k \tag{19.13.6}$$

Another situation where previous readings of a variable are useful is when a small current sample size, N_i, is used to calculate the mean value, $\overline{X}_i$, of a variable. If a much larger set of previous readings for the same test conditions is available, then it can be used to calculate a more appropriate standard deviation for the variable (Steele et al., 1993) as

$$S_{\overline{X}_i} = \frac{S_{X_i}}{\sqrt{N_i}} \tag{19.13.7}$$

where N_i is the number of current readings averaged to determine $\overline{X}_i$, and S_{X_i} is computed from N_{P_i} previous readings using Equation (19.13.5). Typically, these larger data sets are taken in the early "shake-down" or "debugging" phases of an experimental program.

For many engineering applications, the "large sample" approximation applies, and the uncertainty for variable i (X_i or $\overline{X}_i$) is

$$U_i = \sqrt{B_i^2 + \left(2S_i\right)^2} \tag{19.13.8}$$

where S_i is found from the applicable Equation (19.13.3), (19.13.5) or (19.13.7). The interval $X_i \pm U_i$ or $\overline{X}_i + U_i$, as appropriate, should contain $(X_i)_{\text{true}}$ 95 times out of 100. If a small number of samples (N_i or $N_{P_i} < 10$) is used to determine $S_{\overline{X}_i}$ or S_{X_i}, then the "large sample" approximation may not apply and the methods in ISO (1993) or Coleman and Steele (1995) should be used to find U_i.

Uncertainty of a Result

Consider an experimental result that is determined for J measured variables as

$$r = r\left(X_1, X_2, \ldots, X_i, \ldots, X_J\right)$$

where some variables may be single readings and others may be mean values. A typical mechanical engineering experiment would be the determination of the heat transfer in a heat exchanger as

$$q = \dot{m} c_p \left(T_o - T_i\right) \tag{19.13.9}$$

where q is the heat rate, $\dot{m}$ is the flow rate, c_p is the fluid specific heat, and T_o and T_i are the heated fluid outlet and inlet temperatures, respectively. For the "large sample" approximation, U_r is found as

$$U_r = \sqrt{B_r^2 + \left(2S_r\right)^2} \tag{19.13.10}$$

where B_r is the systematic uncertainty of the result

$$B_r^2 = \sum_{i=1}^{J} \left(\theta_i B_i\right)^2 + 2\sum_{i=1}^{J-1}\sum_{k=i+1}^{J} \theta_i \theta_k B_{ik} \tag{19.13.11}$$

with

$$\theta_i = \frac{\partial r}{\partial X_i} \tag{19.13.12}$$

and S_r is the standard deviation of the result

$$S_r^2 = \sum_{i=1}^{J} \left(\theta_i S_i\right)^2 \tag{19.13.13}$$

The term B_{ik} in Equation (19.13.11) is the covariance of the systematic uncertainties. When the elemental systematic uncertainties for two separately measured variables are related, for instance when the transducers used to measure different variables are each calibrated against the same standard, the systematic uncertainties are said to be correlated and the covariance of the systematic errors is nonzero. The significance of correlated systematic uncertainties is that they can have the effect of either decreasing or increasing the uncertainty in the result. B_{ik} is determined by summing the products of the elemental systematic uncertainties for variables i and k that arise from the same source and are therefore perfectly correlated (Brown et al., 1996) as

$$B_{ik} = \sum_{\alpha=1}^{L} \left(B_i\right)_\alpha \left(B_k\right)_\alpha \tag{19.13.14}$$

where L is the number of elemental systematic error sources that are common for measurements X_i and X_k.

If, for example,

$$r = r\left(X_1, X_2\right) \tag{19.13.15}$$

and it is possible for portions of the systematic uncertainties B_1 and B_2 to arise from the same source(s), Equation (19.13.11) gives

$$B_r^2 = \theta_1^2 B_1^2 + \theta_2^2 B_2^2 + 2\theta_1 \theta_2 B_{12} \tag{19.13.16}$$

For a case in which the measurements of X_1 and X_2 are each influenced by four elemental systematic error sources and sources two and three are the same for both X_1 and X_2, Equation (19.13.2) gives

$$B_1^2 = \left(B_1\right)_1^2 + \left(B_1\right)_2^2 + \left(B_1\right)_3^2 + \left(B_1\right)_4^2 \tag{19.13.17}$$

and

$$B_2^2 = \left(B_2\right)_1^2 + \left(B_2\right)_2^2 + \left(B_2\right)_3^2 + \left(B_2\right)_4^2 \tag{19.13.18}$$

while Equation (19.13.14) gives

$$B_{12} = \left(B_1\right)_2 \left(B_2\right)_2 + \left(B_1\right)_3 \left(B_2\right)_3 \tag{19.13.19}$$

In the general case, there would be additional terms in the expression for the standard deviation of the result, S_r, (Equation 19.13.13) to take into account the possibility of precision errors in different variables being correlated. These terms have traditionally been neglected, although precision errors in different variables caused by the same uncontrolled factor(s) are certainly possible and can have a substantial impact on the value of S_r (Hudson et al., 1996). In such cases, one would need to acquire sufficient data to allow a valid estimate of the precision covariance terms using standard statistical techniques (ISO, 1993). Note, however, that if multiple test results over an appropriate time period are available, these can be used to directly determine S_r. This value of the standard deviation of the result implicitly includes the correlated error effect.

If a test is performed so that M multiple sets of measurements $(X_1, X_2, \ldots, X_J)_k$ at the same test condition are obtained, then M results can be determined using Equation (19.13.1) and a mean result, $\bar{r}$, can be determined using

$$\bar{r} = \frac{1}{M}\sum_{k=1}^{M} r_k \tag{19.13.20}$$

The standard deviation of the sample of M results, S_r, is calculated as

$$S_r = \left[\frac{1}{M-1}\sum_{k=1}^{M}\left(r_k - \bar{r}\right)^2\right]^{1/2} \tag{19.13.21}$$

The uncertainty associated with the mean result, $\bar{r}$, for the "large sample" approximation is then

$$U_{\bar{r}} = \sqrt{B_r^2 + \left(2S_{\bar{r}}\right)^2} \tag{19.13.22}$$

where

$$S_{\bar{r}} = \frac{S_r}{\sqrt{M}} \tag{19.13.23}$$

and where B_r is given by Equation (19.13.11).

The "large sample" approximation for the uncertainty of a determined result (Equations (19.13.10) or (19.13.22)) applies for most engineering applications even when some of the variables have fewer

than 10 samples. A detailed discussion of the applicability of this approximation is given in Steele et al. (1994) and Coleman and Steele (1995).

The determination of U_r from S_r (or $S_{\bar{r}}$) and B_r using the "large sample" approximation is called detailed uncertainty analysis (Coleman and Steele, 1989). The interval r (or $\bar{r}$) $\pm$ U_r (or $U_{\bar{r}}$) should contain r_{true} 95 times out of 100. As discussed in the next section, detailed uncertainty analysis is an extremely useful tool in an experimental program. However, in the early stages of the program, it is also useful to estimate the overall uncertainty for each variable, U_i. The overall uncertainty of the result is then determined as

$$U_r^2 = \sum_{k=1}^{J} \left(\theta_i U_i\right)^2 \tag{19.13.24}$$

This determination of U_r is called general uncertainty analysis.

Using Uncertainty Analysis in Experimentation

The first item that should be considered in any experimental program is "What question are we trying to answer?" Another key item is how accurately do we need to know the answer, or what "degree of goodness" is required? With these two items specified, general uncertainty analysis can be used in the planning phase of an experiment to evaluate the possible uncertainties from the various approaches that might be used to answer the question being addressed. Critical measurements that will contribute most to the uncertainty of the result can also be identified.

Once past the planning, or preliminary design phase of the experiment, the effects of systematic errors and precision errors are considered separately using the techniques of detailed uncertainty analysis. In the design phase of the experiment, estimates are made of the systematic and precision uncertainties, B_r and $2S_r$, expected in the experimental result. These detailed design considerations guide the decisions made during the construction phase of the experiment.

After the test is constructed, a debugging phase is required before production tests are begun. In the debugging phase, multiple tests are run and the precision uncertainty determined from them is compared with the $2S_r$ value estimated in the design phase. Also, a check is made to see if the test results plus and minus U_r compare favorably with known results for certain ranges of operation. If these checks are not successful, then further test design, construction, and debugging is required.

Once the test operation is fully understood, the execution phase can begin. In this phase, balance checks can be used to monitor the operation of the test apparatus. In a balance check, a quantity, such as flow rate, is determined by different means and the difference in the two determinations, z, is compared to the ideal value of zero. For the balance check to be satisfied, the quantity z must be less than or equal to U_z.

Uncertainty analysis will of course play a key role in the data analysis and reporting phases of an experiment. When the experimental results are reported, the uncertainties should be given along with the systematic uncertainty, B_r, the precision uncertainty, $2S_r$, and the associated confidence level, usually 95%.

References

Abernethy, R.B., Benedict, R.P., and Dowdell, R.B. 1985. ASME Measurement Uncertainty. *J. Fluids Eng.*, 107, 161–164.

AGARD-AR-304. 1994. *Quality Assessment for Wind Tunnel Testing*. AGARD, Neuilly Sur Seine, France.

AIAA Standard S-071-1995. 1995. *Assessment of Wind Tunnel Data Uncertainty.* AIAA, Washington, D.C.

ANSI/ASME MFC-2M-1983. 1984. *Measurement Uncertainty for Fluid Flow in Closed Conduits.* ASME, New York.

ANSI/ASME PTC 19.1-1985, Part 1. 1986. *Measurement Uncertainty.* ASME, New York.

Coleman, H.W. and Steele, W.G. 1989. *Experimentation and Uncertainty Analysis for Engineers*. John Wiley & Sons, New York.

Coleman, H.W. and Steele, W.G. 1995. Engineering Application of Experimental Uncertainty Analysis. *AIAA Journal*, 33(10), 1888–1896.

Brown, K.B., Coleman, H.W., Steele, W.G., and Taylor, R.P. 1996. Evaluation of Correlated Bias Approximations in Experimental Uncertainty Analysis. *AIAA J.*, 34(5), 1013–1018.

Hudson, S.T., Bordelon, Jr., W.J., and Coleman, H.W. 1996. Effect of Correlated Precision Errors on the Uncertainty of a Subsonic Venturi Calibration. *AIAA J.*, 34(9), 1862–1867.

Kline, S.J. and McClintock, F.A. 1953. Describing Uncertainties in Single-Sample Experiments. *Mech. Eng.*, 75, 3–8.

ISO. 1993. *Guide to the Expression of Uncertainty in Measurement.* ISO, Geneva, Switzerland.

Steele, W.G., Taylor, R.P., Burrell, R.E., and Coleman, H.W. 1993. Use of Previous Experience to Estimate Precision Uncertainty of Small Sample Experiments. *AIAA J.*, 31(10), 1891–1896.

Steele, W.G., Ferguson, R.A., Taylor, R.P., and Coleman, H.W. 1994. Comparison of ANSI/ASME and ISO Models for Calculation of Uncertainty. *ISA Trans.*, 33, 339–352.

19.14 Chaos

R. L. Kautz

Introduction

Since the time of Newton, the science of dynamics has provided quantitative descriptions of regular motion, from a pendulum's swing to a planet's orbit, expressed in terms of differential equations. However, the role of Newtonian mechanics has recently expanded with the realization that it can also describe chaotic motion. In elementary terms, **chaos** can be defined as **pseudorandom** behavior observed in the steady-state dynamics of a deterministic **nonlinear system**. How can motion be pseudorandom, or random according to statistical tests and yet be entirely predictable? This is just one of the paradoxes of chaotic motion, which is globally stable but locally unstable, predictable in principle but not in practice, and geometrically complex but derived from simple equations.

The strange nature of chaotic motion was first understood by Henri Poincaré, who established the mathematical foundations of chaos in a treatise published in 1890 (Holmes, 1990). However, the practical importance of chaos began to be widely appreciated only in the 1960s, beginning with the work of Edward Lorenz (1963), a meteorologist who discovered chaos in a simple model for fluid convection. Today, chaos is understood to explain a wide variety of apparently random natural phenomena, ranging from dripping faucets (Martien et al., 1985), to the flutter of a falling leaf (Tanabe and Kaneko, 1994), to the irregular rotation of a moon of Saturn (Wisdom et al., 1984).

Although chaos is used purposely to provide an element of unpredictability in some toys and carnival rides (Kautz and Huggard, 1994), it is important from an engineering point of view primarily as a phenomenon to be avoided. Perhaps the simplest scenario arises when a nonlinear mechanism is used to achieve a desired effect, such as the synchronization of two oscillators. In many such cases, the degree of nonlinearity must be chosen carefully: strong enough to ensure the desired effect but not so strong that chaos results. In another scenario, an engineer might be required to deal with an intrinsically chaotic system. In this case, if the system can be modeled mathematically, then a small feedback signal can often be applied to eliminate the chaos (Ott et al., 1990). For example, low-energy feedback has been used to suppress chaotic behavior in a thermal convection loop (Singer et al., 1991). As such considerations suggest, chaos in rapidly becoming an important topic for engineers.

Flows, Attractors, and Liapunov Exponents

Dynamic systems can generally be described mathematically in terms of a set of differential equations of the form.

$$d\mathbf{x}(t)/dt = \mathbf{F}\left[\mathbf{x}(t)\right] \tag{19.14.1}$$

where $x = (x_1, \ldots, x_N)$ is an N-dimensional vector called the **state vector** and the vector function $\mathbf{F} = F_1(\mathrm{x}), \ldots, F_N(\mathbf{x})$ defines how the state vector changes with time. In mechanics, the state variables x_i are typically the positions and velocities associated with the potential and kinetic energies of the system. Because the state vector at times $t > 0$ depends only on the initial state vector $\mathbf{x}(0)$, the system defined by Equation (19.14.1) is deterministic, and its motion is in principle exactly predictable.

The properties of a dynamic system are often visualized most readily in terms of trajectories $\mathbf{x}(t)$ plotted in **state space**, where points are defined by the coordinates $(x_1, \ldots, x_N)$. As an example, consider the motion of a damped pendulum defined by the normalized equation

$$d^2\theta/dt^2 = -\sin\theta - \rho d\theta/dt \tag{19.14.2}$$

which expresses the angular acceleration $d^2\theta/dt^2$ in terms of the gravitational torque – sin θ and a damping torque – ρ$d\theta/dt$ proportional to the angular velocity ν = $d\theta/dt$. If we define the state vector as x = (x_1,x_2) = (θ,ν), then Equation (19.14.2) can be written in the form of Equation (19.14.1) with **F** = (x_2, – sin x_1 – ρx_2). In this case, the state space is two dimensional, and a typical trajectory is a spiral, as shown in Figure 19.14.1 for the initial condition x(0) = (0,1). If additional trajectories, corresponding to other initial conditions, were plotted in Figure 19.14.1, we would obtain a set of interleaved spirals, all converging on the point **x** = (0,0). Because the direction of a trajectory passing through a given point is uniquely defined by Equation (19.14.1), state–space trajectories can never cross, and, by analogy with the motion of a fluid, the set of all of trajectories is called a flow.

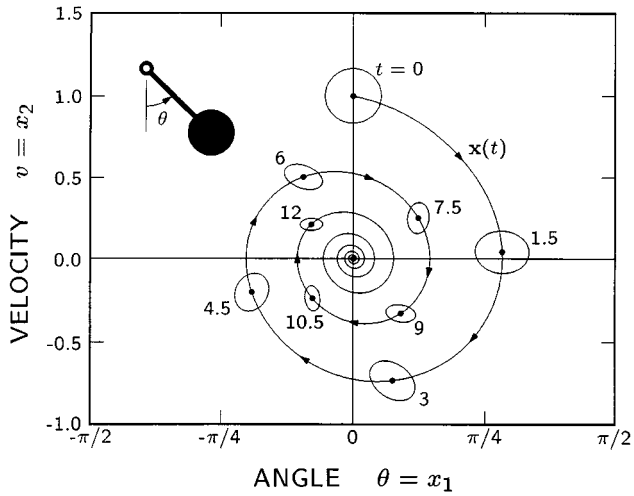

FIGURE 19.14.1 The state-space trajectory **x**(*t*) for a pendulum with a damping coefficient ρ = 0.2 for the initial condition **x**(0) = (0,1). The evolution of trajectories initialized in a small circle surrounding **x** = (0,1) is indicated by the ellipses plotted at time intervals of Δ*t* = 1.5.

The tendency of a flow to converge toward a single point or other restricted subset of state space is characteristic of dissipative systems like the damped pendulum. Such an asymptotic set, called an attracting set or **attractor**, can be a fixed point (for which **F(x)** = 0) as in Figure 19.14.1, but might also be a periodic or chaotic trajectory. The convergence of neighboring trajectories is suggested in Figure 19.14.1 by a series of ellipses spaced at time intervals Δ*t* = 1.5 that track the flow of all trajectories originating within the circle specified at *t* = 0. In general, the contraction of an infinitesimal state–space volume *V* as it moves with the flow is given by

$$V^{-1}\,\partial V/\partial t = \nabla \cdot \mathbf{F} \tag{19.14.3}$$

where $\nabla \cdot \mathbf{F} = \Sigma_{i=1}^{N} \partial F_i/x_i$ is the divergence of **F**. For the damped pendulum, $\nabla \cdot \mathbf{F} = \rho$, so the area of the ellipse shown in Figure 19.14.1 shrinks exponentially as $V(t) = V(0)\exp(-\rho t)$. The contraction of state–space volumes explains the existence of attractors in dissipative systems, but in conservative systems such as the pendulum with ρ = 0, state–space volumes are preserved, and trajectories are instead confined to constant-energy surfaces.

While the existence of chaotic behavior is generally difficult to predict, two essential conditions are easily stated. First, the complex topology of a chaotic trajectory can exist only in a state–space of dimension $N \geq 3$. Thus, the pendulum defined by Equation (19.14.2) cannot be chaotic because $N = 2$ for this system. Second, a system must be nonlinear to exhibit chaotic behavior. Linear systems, for which any linear combination $c_1x_a(t) + c_2x_b(t)$ of two solutions $x_a(t)$ and $x_b(t)$ is also a solution, are mathematically simple and amenable to analysis. In contrast, nonlinear systems are noted for their

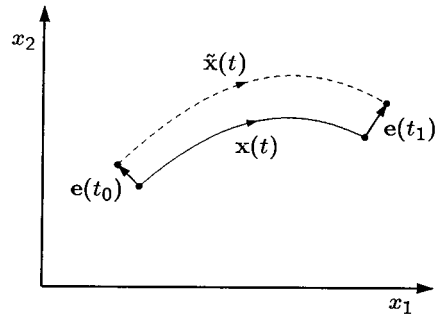

FIGURE 19.14.2 A trajectory $\mathbf{x}(t)$ and a neighboring trajectory $\tilde{\mathbf{x}}(t)$ plotted in state space from time t_0 to t_1. The vectors $\mathbf{e}(t_0)$ and $\mathbf{e}(t_1)$ indicate the deviation of $\tilde{\mathbf{x}}(t)$ from $\mathbf{x}(t)$ at times t_0 and t_1.

intractability. Thus, chaotic behavior is of necessity explored more frequently by numerical simulation than mathematical analysis, a fact that helps explain why the prevalence of chaos was discovered only after the advent of efficient computation.

A useful criterion for the existence of chaos can be developed from an analysis of a trajectory's local stability. As sketched in Figure 19.14.2, the local stability of a trajectory $\mathrm{x}(t)$ is determined by considering a neighboring trajectory $\tilde{x}(t)$ initiated by an infinitesimal deviation $\mathrm{e}(t_0)$ from $\mathrm{x}(t)$ at time t_0. The deviation vector $\mathrm{e}(t) = \tilde{x}(t) - \mathrm{x}(t)$ at times $t_1 > t_0$ can be expressed in terms of the Jacobian matrix

$$J_{ij}(t_1,t_0) = \partial x_i(t_1)/\partial x_j(t_0) \tag{19.14.4}$$

which measures the change in state variable x_i at time t_1 due to a change in x_j at time t_0. From the Jacobian's definition, we have $\mathbf{e}(t_1) = \mathbf{J}(t_1,t_0)\mathbf{e}(t_0)$. Although the local stability of $\mathbf{x}(t)$ is determined simply by whether deviations grow or decay in time, the analysis is complicated by the fact that deviation vectors can also rotate, as suggested in Figure 19.14.2. Fortunately, an arbitrary deviation can be written in terms of the eigenvectors $\mathrm{e}^{(i)}$ of the Jacobian, defined by

$$\mathbf{J}(t_1,t_0)\mathbf{e}^{(i)} = \mu_i(t_1,t_0)\mathbf{e}^{(i)} \tag{19.14.5}$$

which are simply scaled by the eigenvalues $\mu_i(t_1,t_0)$ without rotation. Thus, the N eigenvalues of the Jacobian matrix provide complete information about the growth of deviations. Anticipating that the asymptotic growth will be exponential in time, we define the **Liapunov exponents**,

$$\lambda_i = \lim_{t_1\to\infty} \frac{\ln|\mu_i(t_1,t_0)|}{t_1 - t_0} \tag{19.14.6}$$

Because any deviation can be broken into components that grow or decay asymptotically as $\exp(\lambda_i t)$, the N exponents associated with a trajectory determine its local stability.

In dissipative systems, chaos can be defined as motion on an attractor for which one or more Liapunov exponents are positive. Chaotic motion thus combines global stability with local instability in that motion is confined to the attractor, generally a bounded region of state space, but small deviations grow exponentially in time. This mixture of stability and instability in chaotic motion is evident in the behavior of an infinitesimal deviation ellipsoid similar to the finite ellipse shown in Figure 19.14.1. Because some λ_i are positive, an ellipsoid centered on a chaotic trajectory will expand exponentially in some directions. On the other hand, because state-space volumes always contract in dissipative systems and the asymptotic volume of the ellipsoid scales as $\exp(\Lambda t)$, where $\Lambda = \Sigma_{i=1}^{N} \lambda_i$, the sum of the negative exponents must be greater in magnitude than the sum of the positive exponents. Thus, a deviation ellipsoid tracking a

chaotic trajectory expands in some directions while contracting in others. However, because an arbitrary deviation almost always includes a component in a direction of expansion, nearly all trajectories neighboring a chaotic trajectory diverge exponentially.

According to our definition of chaos, neighboring trajectories must diverge exponentially and yet remain on the attractor. How is this possible? Given that the attractor is confined to a bounded region of state space, perpetual divergence can occur only for trajectories that differ infinitesimally. Finite deviations grow exponentially at first but are limited by the bounds of the chaotic attractor and eventually shrink again. The full picture can be understood by following the evolution of a small state-space volume selected in the neighborhood of the chaotic attractor. Initially, the volume expands in some directions and contracts in others. When the expansion becomes too great, however, the volume begins to fold back on itself so that trajectories initially separated by the expansion are brought close together again. As time passes, this stretching and folding is repeated over and over in a process that is often likened to kneading bread or pulling taffy.

Because all neighboring volumes approach the attractor, the stretching and folding process leads to an attracting set that is an infinitely complex filigree of interleaved surfaces. Thus, while the differential equation that defines chaotic motion can be very simple, the resulting attractor is highly complex. Chaotic attractors fall into a class of geometric objects called **fractals**, which are characterized by the presence of structure at arbitrarily small scales and by a dimension that is generally fractional. While the existence of objects with dimensions falling between those of a point and a line, a line and a surface, or a surface and a volume may seem mysterious, fractional dimensions result when dimension is defined by how much of an object is apparent at various scales of resolution. For the dynamical systems encompassed by Equation (19.14.1), the fractal dimension D of a chaotic attractor falls in the range of $2 < D < N$ where N is the dimension of the state space. Thus, the dimension of a chaotic attractor is large enough that trajectories can continually explore new territory within a bounded region of state space but small enough that the attractor occupies no volume of the space.

Synchronous Motor

As an example of a system that exhibits chaos, we consider a simple model for a synchronous motor that might be used in a clock. As shown in Figure 19.14.3, the motor consists of a permanent-magnet rotor subjected to a uniform oscillatory magnetic field $B \sin t$ provided by the stator. In dimensionless notation, its equation of motion is

$$d^2\theta/dt^2 = -f \sin t \sin\theta - \rho d\theta/dt \tag{19.14.7}$$

where $d^2\theta/dt^2$ is the angular acceleration of the rotor, $-f \sin t \sin\theta$ is the torque due to the interaction of the rotor's magnetic moment with the stator field, and $-\rho d\theta/dt$ is a viscous damping torque. Although Equation (19.14.7) is explicitly time dependent, it can be case in the form of Equation (19.14.1) by

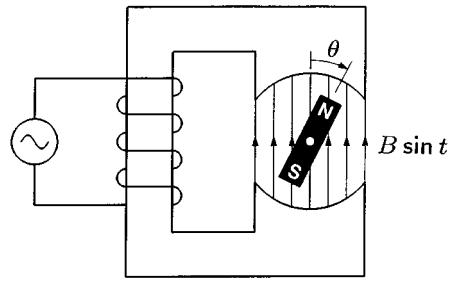

FIGURE 19.14.3 A synchronous motor, consisting of a permanent magnet free to rotate in a uniform magnetic field $B \sin t$ with an amplitude that varies sinusoidally in time.

defining the state vector as $\mathbf{x} = (x_1,x_2,x_3) = (\theta,\nu,t)$, where $\nu = d\theta/dt$ is the angular velocity, and by defining the flow as $\mathbf{F} = (x_2, -f \sin x_3 \sin x_1 - \rho x_2, 1)$. The state space is thus three dimensional and large enough to allow chaotic motion. Equation (19.14.7) is also nonlinear due to the term $-f \sin t \sin \theta$, since $\sin(\theta_a + \theta_b)$ is not generally equal to $\sin \theta_a + \sin \theta_b$. Chaos in this system has been investigated by several authors (Ballico et al., 1990).

By intent, the motor uses nonlinearity to synchronize the motion of the rotor with the oscillatory stator field, so it evolves exactly once during each field oscillation. Although synchronization can occur over a range of system parameters, proper operation requires that the drive amplitude f, which measures the strength of the nonlinearity, be chosen large enough to produce the desired rotation but not so large that chaos results. Calculating the motor's dynamics for $\rho = 0.2$, we find that the rotor oscillates without rotating for f less than 0.40 and that the intended rotation is obtained for $0.40 < \rho < 1.87$. The periodic attractor corresponding to synchronized rotation is shown for $f = 1$ in Figure 19.14.4(a). Here the three-dimensional state-space trajectory is projected onto the (x_1,x_2) or (θ,ν) plane, and a dot marks the point in the rotation cycle at which $t = 0$ modulo 2π. As Figure 19.14.4(a) indicates, the rotor advances by exactly 2π during each drive cycle.

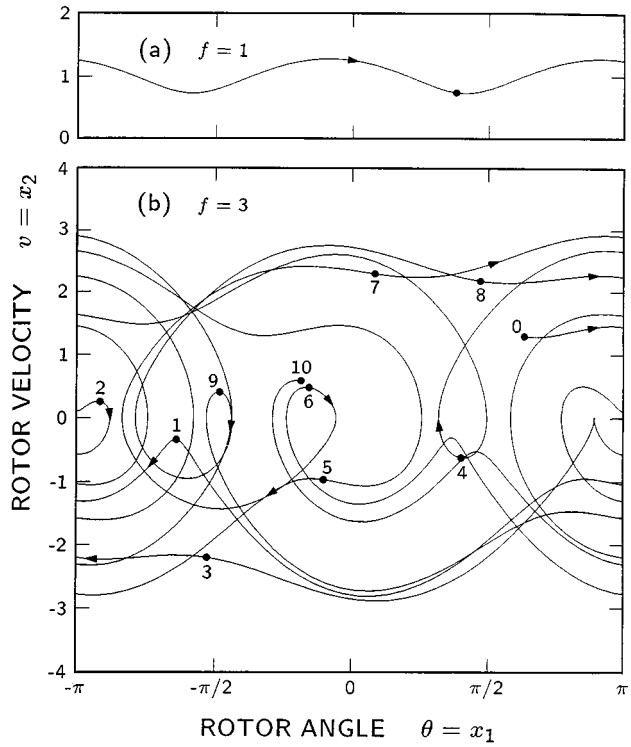

FIGURE 19.14.4 State-space trajectories projected onto the (x_1,x_2) or (θ,ν) plane, showing attractors of the synchronous motor for $\rho = 0.2$ and two drive amplitudes, $f = 1$ and 3. Dots mark the state of the system at the beginning of each drive cycle ($t = 0$ modulo 2π). The angles $\theta = \pi$ and $-\pi$ are equivalent.

The utility of the motor hinges on the stability of the synchronous rotation pattern shown in Figure 19.14.4(a). This periodic pattern is the steady–state motion that develops after initial transients decay and represents the final asymptotic trajectory resulting for initial conditions chosen from a wide area of state space. Because the flow approaches this attracting set from all neighboring points, the effect of a perturbation that displaces the system from the attractor is short lived. This stability is reflected in the Liapunov exponents of the attractor: $\lambda_1 = 0$ and $\lambda_2 = \lambda_3 = -0.100$. The zero exponent is associated with deviations coincident with the direction of the trajectory and is a feature common to all bounded attractors

other than fixed points. The zero exponent results because the system is neutrally stable with respect to offsets in the time coordinate. The exponents of –0.100 are associated with deviations transverse to the trajectory and indicate that these deviations decay exponentially with a characteristic time of 1.6 drive cycles. The negative exponents imply that the synchrony between the rotor and the field is maintained in spite of noise or small variations in system parameters, as required of a clock motor.

For drive amplitudes greater than $f = 1.87$, the rotor generally does not advance by precisely 2π during every drive cycle, and its motion is commonly chaotic. An example of chaotic behavior is illustrated for $f = 3$ by the trajectory plotted in Figure 19.14.4(b) over an interval of 10 drive cycles. In this figure, sequentially numbered dots mark he beginning of each drive cycle. When considered cycle by cycle, the trajectory proves to be a haphazard sequence of oscillations, forward rotations, and reverse rotations. Although we might suppose that this motion is just an initial transient, it is instead characteristic of the steady–state behavior of the motor. If extended, the trajectory continued with an apparently random mixture of oscillation and rotation, without approaching a repetitive cycle. The motion is aptly described as chaotic.

The geometry of the chaotic attractor sampled in Figure 19.14.4(b) is revealed more fully in Figure 19.14.5. Here we plot points (θ,ν) recording the instantaneous angle and velocity of the rotor at the beginning of each drive cycle for 100,000 successive cycles, Figure 19.14.5 displays the three-dimensional attractor called a **Poincaré section**, at its intersection with the planes $t = x_3 = 0$ modulo 2π, corresponding to equivalent times in the drive cycle. For the periodic attractor of Figure 19.14.4(a), the rotor returns to the same position and velocity at the beginning of each drive cycle, so its Poincaré section is a single point, the dot in this figure. For chaotic motion, in contrast, we obtain the complex swirl of points shown in Figure 19.14.5. If the system is initialized at a point far from the swirl, the motion quickly converges to this attracting set. On succeeding drive cycles, the state of the system jumps from one part of the swirl to another in an apparently random fashion that continues indefinitely. As the number of plotted points approaches infinity, the swirl becomes a cross section of the chaotic attractor. Thus, Figure 19.14.5 approximates a slice through the infinite filigree of interleaved surfaces that compose the attracting set. In this case, the fractal dimension of the attractor is 2.52 and that of its Poincaré section is 1.52.

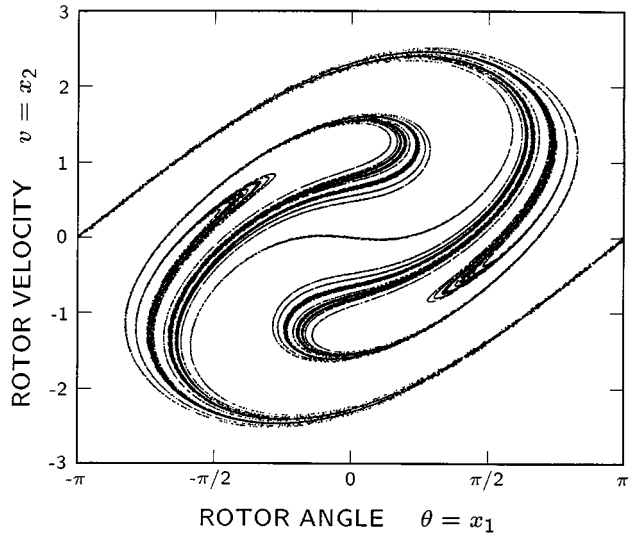

FIGURE 19.14.5 Poincaré section of a chaotic attractor of the synchronous motor with $\rho = 0.2$ and $f = 3$, obtained by plotting points $(x_1,x_2) = (\theta,\nu)$ corresponding to the position and velocity of the rotor at the beginning of 100,000 successive drive cycles.

The computed Liapunov exponents of the chaotic solution at $\rho = 0.2$ and $f = 3$ are $\lambda_1 = 0$, $\lambda_2 = 0.213$, and $\lambda_3 = -0.413$. As for the periodic solution, the zero exponent implies neutral stability associated with deviations directed along a given trajectory. The positive exponent, which signifies the presence of chaos, is associated with deviations transverse to the given trajectory but tangent to the surface of the attracting set in which it is embedded. The positive exponent implies that such deviations grow exponentially in time and that neighboring trajectories on the chaotic attractor diverge exponentially, a property characteristic of chaotic motion. The negative exponent is associated with deviations transverse to the surface of the attractor and assures the exponential decay of displacements from the attracting set. Thus, the Liapunov exponents reflect both the stability of the chaotic attractor and the instability of a given chaotic trajectory with respect to neighboring trajectories.

One sequence of a positive Liapunov exponent is a practical limitation on our ability to predict the future state of a chaotic system. This limitation is illustrated in Figure 19.14.6, where we plot a given chaotic trajectory (solid line) and three perturbed trajectories (dashed lines) that result by offsetting the initial phase of the given solution by various deviations $e_1(0)$. When the initial angular offset is $e_1(0) = 10^{-3}$ radian, the perturbed trajectory (short dash) closely tracks the given trajectory for about seven drive cycles before the deviation become significant. After seven drive cycles, the perturbed trajectory is virtually independent of the given trajectory, even though it is confined to the same attractor. Similarly, initial offsets of 10^{-6} and 10^{-9} radian lead to perturbed trajectories (medium and long dash) that track the given trajectory for about 12 and 17 drive cycles, respectively, before deviations become significant. These results reflect the fact that small deviations grow exponentially and, in the present case, increase on average by a factor of 10 every 1.7 drive cycles. If the position of the rotor is to be predicted with an accuracy of 10^{-1} radian after 20 drive cycles, its initial angle must be known to better than 10^{-13} radian, and the calculation must be carried out with at least 14 significant digits. If a similar prediction is to be made over 40 drive cycles, then 25 significant digits are required. Thus, even though chaotic motion is predictable in principle, the state of a chaotic system can be accurately predicted in practice for only a short time into the future. According to Lorenz (1993), this effect explains why weather forecasts are of limited significance beyond a few days.

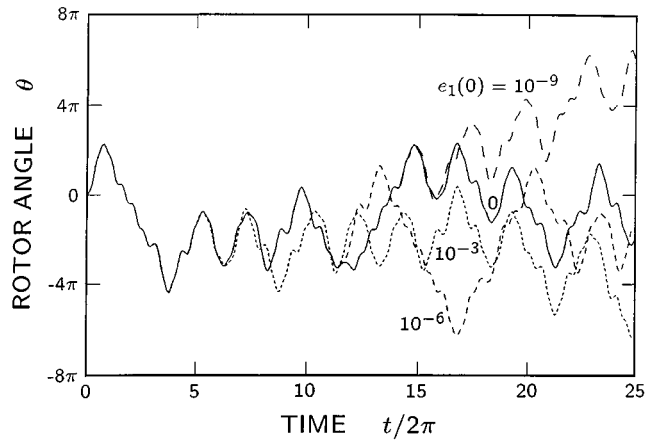

FIGURE 19.14.6 Rotor angle as a function of time for chaotic trajectories of the synchronous motor with $\rho = 0.2$ and $f = 3$. Solid line shows a given trajectory and dashed lines show perturbed trajectories resulting from initial angular deviations of $e_1(0) = 10^{-3}$ (short dash), 10^{-6} (medium dash), and 10^{-9} (long dash).

This pseudorandom nature of chaotic motion is illustrated in Figure 19.14.7 for the synchronous motor by a plot of the net rotation during each of 100 successive drive cycles. Although this sequence of rotations results from solving a deterministic equation, it is apparently random, jumping erratically between forward and reverse rotations of various magnitudes up to about 1.3 revolutions. The situation

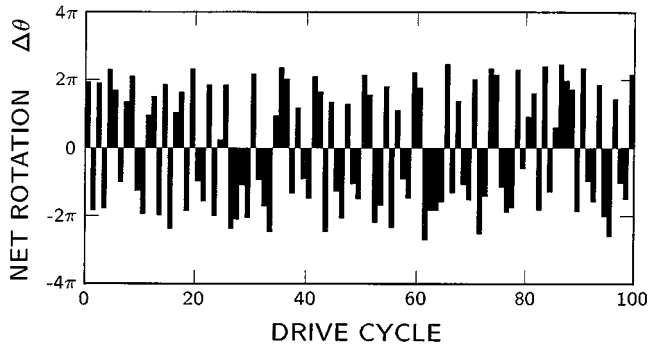

FIGURE 19.14.7 Net rotation of a synchronous motor during each of 100 successive drive cycles, illustrating chaotic motion for $\rho = 0.2$ and $f = 3$. By definition, $\Delta\theta = \theta(2\pi n) - \theta(2\pi(n-1))$ on the nth drive cycle.

is similar to that of a digital random number generator, in which a deterministic algorithm is used to produce a sequence of pseudorandom numbers. In fact, the similarity is not coincidental since chaotic processes often underlie such algorithms (Li, 1978). For the synchronous motor, statistical analysis reveals almost no correlation between rotations separated by more than a few drive cycles. This statistical independence is a result of the motor's positive Liapunov exponent. Because neighboring trajectories diverge exponentially, a small region of the attractor can quickly expand to cover the entire attractor, and a small range of rotations on one drive cycle can lead to almost any possible rotation a few cycles later. Thus, there is little correlation between rotations separated by a few drive cycles, and on this time scale the motor appears to select randomly between the possible rotations.

From an engineering point of view, the problem of chaotic behavior in the synchronous motor can be solved simply by selecting a drive amplitude in the range of $0.40 < f < 1.87$. Within this range, the strength of the nonlinearity is large enough to produce synchronization but not so large as to produce chaos. As this example suggests, it is important to recognize that erratic, apparently random motion can be an intrinsic property of a dynamic system and is not necessarily a product of external noise. Searching a real motor for a source of noise to explain the behavior shown in Figure 19.14.7 would be wasted effort since the cause is hidden in a noise-free differential equation. Clearly, chaotic motion is a possibility that every engineer should understand.

Defining Terms

Attractor: A set of points in state space to which neighboring trajectories converge in the limit of large time.

Chaos: Pseudorandom behavior observed in the steady–state dynamics of a deterministic nonlinear system.

Fractal: A geometric object characterized by the presence of structure at arbitrarily small scales and by a dimension that is generally fractional.

Liapunov exponent: One of N constants λ_i that characterize the asymptotic exponential growth of infinitesimal deviations from a trajectory in an N-dimensional state space. Various components of a deviation grow or decay on average in proportion to $\exp(\lambda_i t)$.

Nonlinear system: A system of equations for which a linear combination of two solutions is not generally a solution.

Poincaré section: A cross section of a state-space trajectory formed by the intersection of the trajectory with a plane defined by a specified value of one state variable.

Pseudorandom: Random according to statistical tests but derived from a deterministic process.

State space: The space spanned by state vectors.

State vector: A vector **x** whose components are the variables, generally positions and velocities, that define the time evolution of a dynamical system through an equation of the form $\dot{x}/dt = \mathbf{F}(x)$, where **F** is a vector function.

References

Ballico, M.J., Sawley, M.L., and Skiff, F. 1990. The bipolar motor: A simple demonstration of deterministic chaos. *Am. J. Phys.*, 58, 58–61.

Holmes, P. 1990. Poincaré, celestial mechanics, dynamical-systems theory and "chaos." *Phys. Reports*, 193, 137–163.

Kautz, R.L. and Huggard, B.M. 1994. Chaos at the amusement park: dynamics of the Tilt-A-Whirl. *Am. J. Phys.*, 62, 69–66.

Li, T.Y. and Yorke, J.A. 1978. Ergodic maps on [0,1] and nonlinear pseudo-random number generators. *Nonlinear Anal. Theory Methods Appl.*, 2, 473–481.

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20, 130–141.

Lorenz, E.N. 1993. *The Essence of Chaos*, University of Washington Press, Seattle, WA.

Martien, P., Pope, S.C., Scott, P.L., and Shaw, R.S. 1985. The chaotic behavior of a dripping faucet. *Phys. Lett. A.*, 110, 399–404.

Ott, E., Grebogi, C., and Yorke, J.A. 1990. Controlling chaos. *Phys. Rev. Lett.*, 64, 1196–1199.

Singer, J., Wang, Y.Z., and Bau, H.H. 1991. Controlling a chaotic system. *Phys. Rev. Lett.*, 66, 1123–1125.

Tanabe, Y. and Kaneko, K. 1994. Behavior of falling paper. *Phys. Rev. Lett.*, 73, 1372–1375.

Wisdom, J., Peale, S.J., and Mignard, F. 1984. The chaotic rotation of Hyperion. *Icarus*, 58, 137–152.

For Further Information

A good introduction to deterministic chaos for undergraduates is provided by *Chaotic and Fractal Dynamics: An Introduction for Applied Scientists and Engineers* by Francis C. Moon. This book presents numerous examples drawn from mechanical and electrical engineering.

Chaos in Dynamical Systems by Edward Ott provides a more rigorous introduction to chaotic dynamics at the graduate level.

Practical methods for experimental analysis and control of chaotic systems are presented in *Coping with Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, a reprint volume edited by Edward Ott, Tim Sauer, and James A. York.

19.15 Fuzzy Sets and Fuzzy Logic

Dan M. Frangopol

Introduction

In the sixties, Zaheh (1965) introduced the concept of fuzzy sets. Since its inception more than 30 years ago, the theory and methods of fuzzy sets have developed considerably. The demands for treating situations in engineering, social sciences, and medicine, among other applications that are complex and not crisp have been strong driving forces behind these developments.

The concept of the fuzzy set is a generalization of the concept of the ordinary (or crisp) set. It introduces vagueness by eliminating the clear boundary, defined by the ordinary set theory, between full nonmembers (i.e., grade of membership equals zero) and full members (i.e., grade of membership equals one). According to Zaheh (1965) a fuzzy set A, defined as a collection of elements (also called objects) $x \in X$, where X denotes the universal set (also called universe of discourse) and the symbol $\in$ denotes that the element x is a member of X, is characterized by a membership (also called characteristic) function $\mu_A(x)$ which associates each point in X a real member in the unit interval [0,1]. The value of $\mu_A(x)$ at x represents the grade of membership of x in A. Larger values of $\mu_A(x)$ denote higher grades of membership of x in A. For example, a fuzzy set representing the concept of control might assign a degree of membership of 0.0 for no control, 0.1 for weak control, 0.5 for moderate control, 0.9 for strong control, and 1.0 for full control. From this example, it is clear that the two-valued crisp set [i.e., no control (grade of membership 0.0) and full control (grade of membership 1.0)] is a particular case of the general multivalued fuzzy set A in which $\mu_A(x)$ takes its values in the interval [0,1].

Problems in engineering could be very complex and involve various concepts of uncertainty. The use of fuzzy sets in engineering has been quite extensive during this decade. The area of fuzzy control is one of the most developed applications of fuzzy set theory in engineering (Klir and Folger, 1988). Fuzzy controllers have been created for the control of robots, aircraft autopilots, and industrial processes, among others. In Japan, for example, so-called "fuzzy electric appliances," have gained great success from both technological and commercial points of view (Furuta, 1995). Efforts are underway to develop and introduce fuzzy sets as a technical basis for solving various real-world engineering problems in which the underlying information is complex and imprecise. In order to achieve this, a mathematical background in the theory of fuzzy sets is necessary. A brief summary of the fundamental mathematical aspects of the theory of fuzzy sets is presented herein.

Fundamental Notions

A fuzzy set A is represented by all its elements x_i and associated grades of membership $\mu_A(x_i)$ (Klir and Folger, 1988).

$$A = \left\{\mu_A(x_1)|x_1, \mu_A(x_2)|x_2, \ldots, \mu_A(x_n)|x_n\right\} \tag{19.15.1}$$

where x_i is an element of the fuzzy set, $\mu_A(x_i)$ is its grade of membership in A, and the vertical bar is employed to link the element with their grades of membership in A. Equation (19.15.1) shows a discrete form of a fuzzy set. For a continuous fuzzy set, the membership function $\mu_A(x)$ is a continuous function of x.

Figure 19.15.1 illustrates a discrete and a continuous fuzzy set. The larger membership grade *max* $(\mu_A(x_i))$ represents the height of a fuzzy set.

If at least one element of the fuzzy set has a membership grade of 1.0, the fuzzy set is called normalized. Figure 19.15.2 illustrates both a nonnormalized and a normalized fuzzy set.

The following properties of fuzzy sets, which are obvious extensions of the corresponding definitions for ordinary (crisp) sets, are defined herein according to Zaheh (1965) and Klir and Folger (1988).

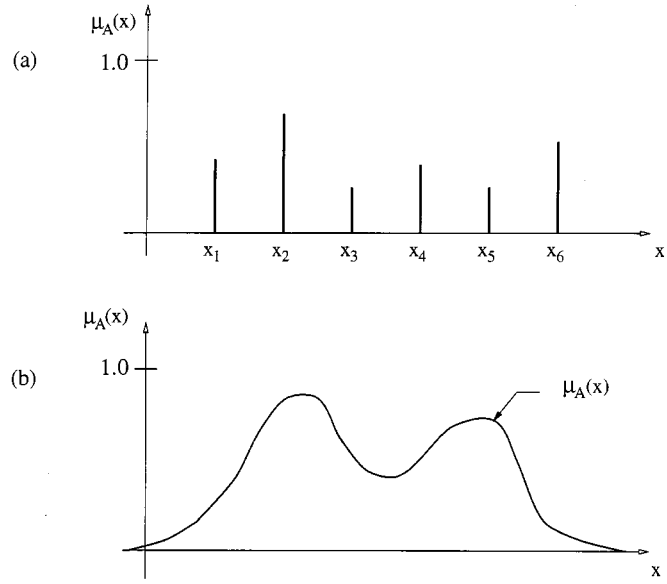

FIGURE 19.15.1 (a) Discrete and (b) continuous fuzzy set.

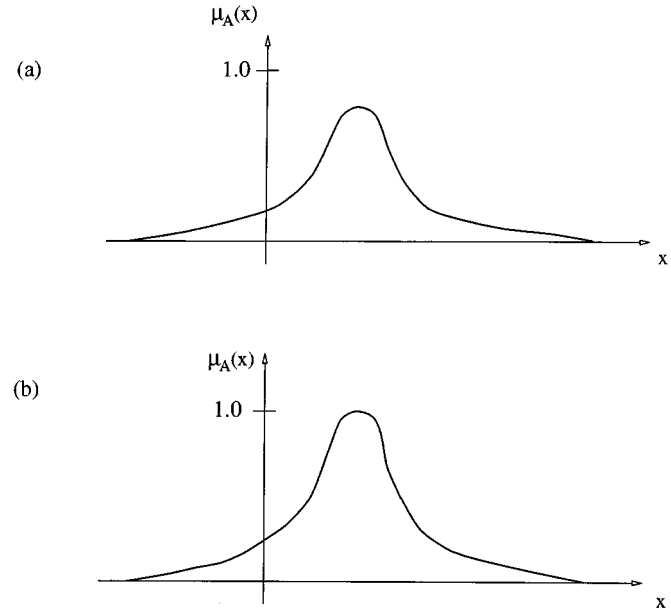

FIGURE 19.15.2 (a) Nonnormalized and (b) normalized fuzzy set.

Two fuzzy sets A and B are equal, A = B, if and only if $\mu_A(x) = \mu_B(x)$ for every element x in X (see Figure 19.15.3).

The complement of a fuzzy set A is a fuzzy set $\overline{A}$ defined as

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x) \tag{19.15.2}$$

Figure 19.15.4 shows both discrete and continuous fuzzy sets and their complements.

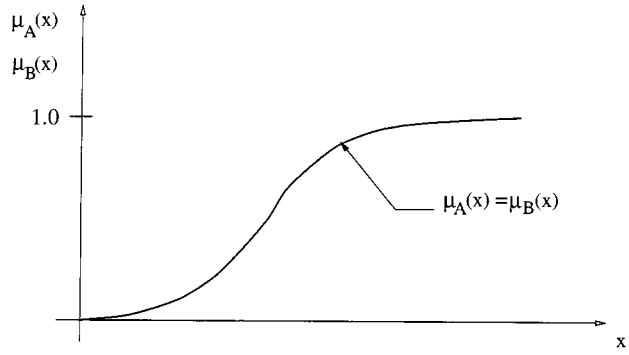

FIGURE 19.15.3 Two equal fuzzy sets, A = B.

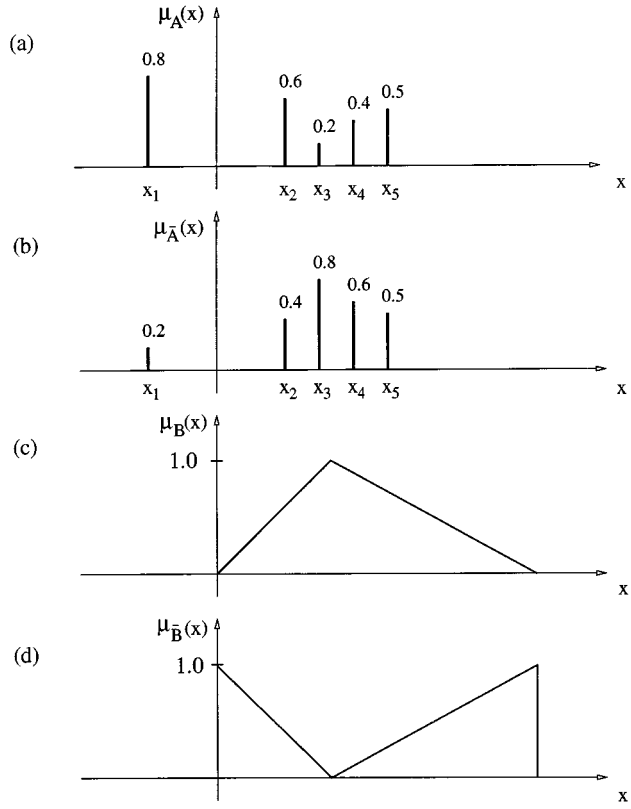

FIGURE 19.15.4 (a) Discrete fuzzy set A, (b) complement $\bar{A}$ of fuzzy set A, (c) continuous fuzzy set B, and (d) complement $\bar{B}$ of fuzzy set B.

If the membership grade of each element of the universal set X in fuzzy set B is less than or equal to its membership grade in fuzzy set A, then B is called a subset of A. This is denoted $B \subseteq A$. Figure 19.15.5 illustrates this situation.

The union of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set $C = A \cup B$ such that

$$\mu_C(x) = \max\left[\mu_A(x), \mu_B(x)\right] \tag{19.15.3}$$

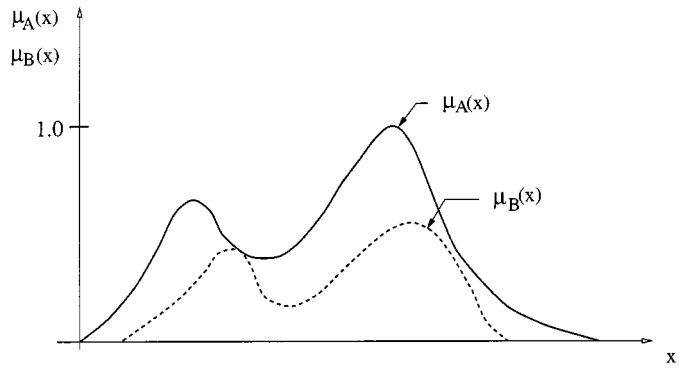

FIGURE 19.15.5 Fuzzy set A and its subset B.

for all x in X.

Conversely, the intersection of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$, respectively, is a fuzzy set $C = A \cap B$ such that

$$\mu_C(x) = \min\left[\mu_A(x), \mu_B(x)\right] \tag{19.15.4}$$

for all x in X.

Figure 19.15.6 illustrates two fuzzy sets A and B, the union set $A \cup B$ and the intersection set $A \cap B$.

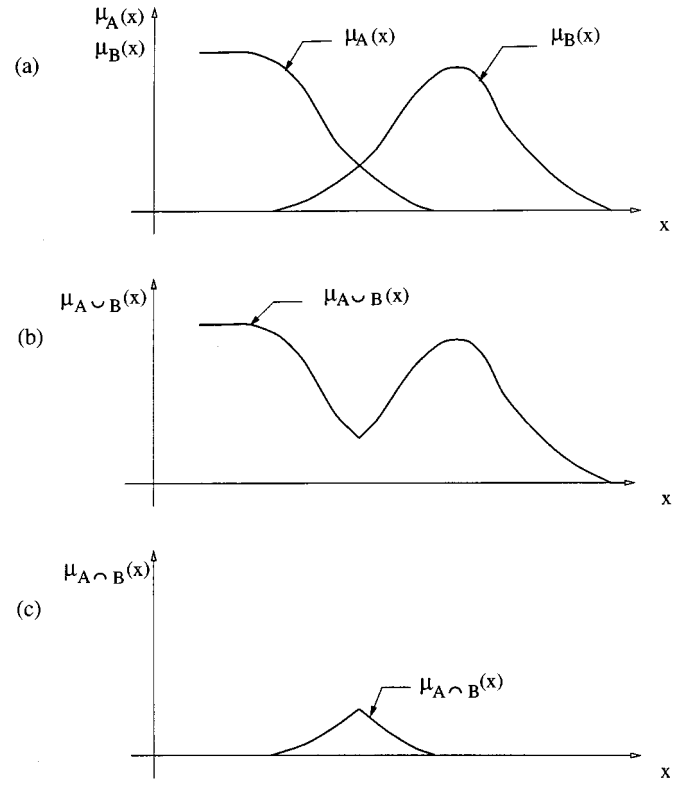

FIGURE 19.15.6 (a) Two fuzzy sets, (b) union of fuzzy sets $A \cup B$, and (c) intersection of fuzzy sets $A \cap B$.

An empty fuzzy set A is a fuzzy set with a membership function $\mu_A(x) = 0$ for all elements x in X (see Figure 19.15.7).

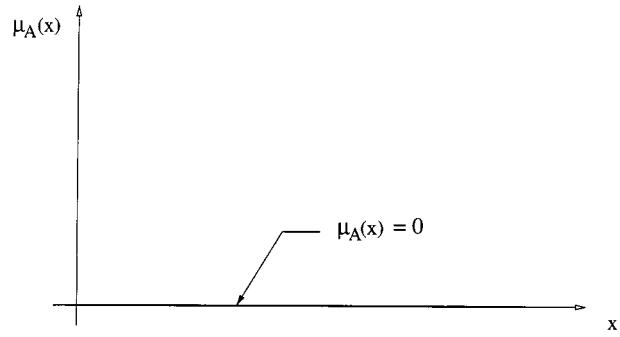

FIGURE 19.15.7 Empty fuzzy set.

Two fuzzy sets A and B with respective membership function $\mu_A(x)$ and $\mu_B(x)$ are disjoint if their intersection is empty (see Figure 19.15.8).

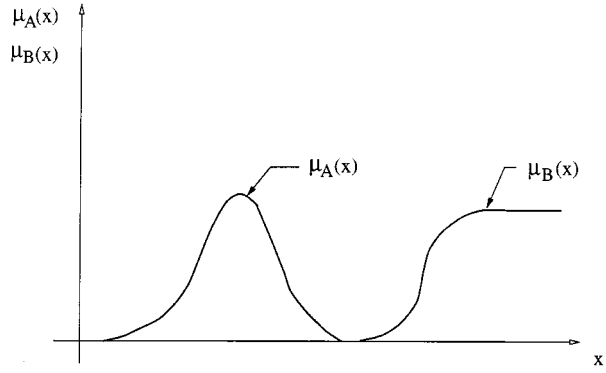

FIGURE 19.15.8 Disjoint fuzzy sets.

An α-cut of a fuzzy set A is an ordinary (crisp) set A_α containing all elements that have a membership grade in A greater or equal to α. Therefore,

$$A_\alpha = \{x | \mu_A(x) \geq \alpha\} \tag{19.15.5}$$

From Figure 19.15.9, it is clear that $\alpha = 0.5$, the α-cut of the fuzzy set A is the crisp set $A_{0.5} = \{x_5, x_6, x_7, x_8\}$ and for $\alpha = 0.8$, the α-cut of the fuzzy set A is the crisp set $A_{0.8} = \{x_7, x_8\}$.

A fuzzy set is convex if and only if all of its α-cuts are convex for all α in the interval [0,1]. Figure 19.15.10 shows both a convex and a nonconvex fuzzy set.

A fuzzy number $\tilde{N}$ is a normalized and convex fuzzy set of the real line whose membership function is piecewise continuous and for which it exists exactly one element with $\mu_{\tilde{N}}(x_0) = 1$. As an example, the real numbers close to 50 are shown by four membership functions in Figure 19.15.11.

The scalar cardinality of a fuzzy set A is the summation of membership grades of all elements of X in A. Therefore,

$$|A| = \sum_x \mu_A(x) \tag{19.15.6}$$

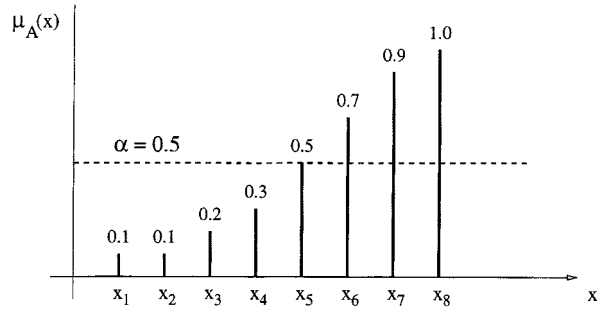

FIGURE 19.15.9 α-cut of a fuzzy set.

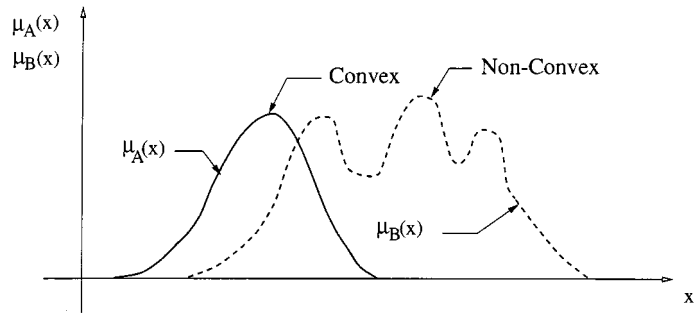

FIGURE 19.15.10 Convex and non-convex fuzzy set.

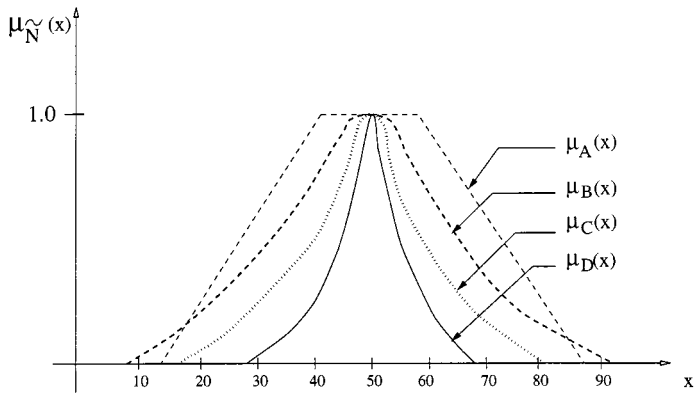

FIGURE 19.15.11 Membership functions of fuzzy sets of real numbers close to 50.

For example, the scalar cardinality of the fuzzy set A in Figure 19.15.4(a) is 2.5. Obviously, an empty fuzzy set has a scalar cardinality equal to zero. Also, the scalar cardinality of the fuzzy complement set is equal to scalar cardinality of the original set. Therefore,

$$|A| = |\overline{A}| \tag{19.15.7}$$

One of the basic concepts of fuzzy set theory is the extension principle. According to this principle (Dubois and Prade, 1980), given (a) a function f mapping points in the ordinary set X to points in the ordinary set Y, and (b) any fuzzy set A defined on X,

$$A = \left\{\mu_A(x_1)|x_1, \mu_A(x_2)|x_2, \ldots, \mu_A(x_n)|x_n\right\}$$

then the fuzzy set B = f(A) is given as

$$B = f(A) = \left\{\mu_A(x_1)|f(x_1), \mu_A(x_2)|f(x_2), \ldots, \mu_A(x_n)|f(x_n)\right\} \tag{19.15.8}$$

If more than one element of the ordinary set X is mapped by *f* to the same element y in Y, then the maximum of the membership grades in the fuzzy set A is considered as the membership grade of y in f(A).

As an example, consider the fuzzy set in Figure 19.15.4(a), where $x_1 = -2$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, and $x_5 = 5$. Therefore, A = {0.8|–2, 0.6|2, 0.2|3, 0.4|4, 0.5|5} and $f(x) = x^4$. By using the extension principle, we obtain

$$f(A) = \left\{\max(0.8, 0.6)|2^4, 0.2|3^4, 0.4|4^4, 0.5|5^4\right\}$$

$$= \left\{0.8|16, 0.2|81, 0.4|256, 0.5|625\right\}$$

As shown by Klir and Folger (1988), degrees of association can be represented by membership grades in a fuzzy relation. Such a relation can be considered a general case for a crisp relation.

Let P be a binary fuzzy relation between the two crisp sets X = {4, 8, 11} and Y = {4, 7} that represents the relational concept “very close.” This relation can be expressed as:

$$P(X,Y) = \left\{1|(4,4), 0.7|(4,7), 0.6|(8,4), 0.9|(8,7), 0.3|(11,4), 0.6|(11,7)\right\}$$

or it can be represented by the two dimensional membership matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \end{array}$$

Fuzzy relations, especially binary relations, are important for many engineering applications.

The concepts of domain, range, and the inverse of a binary fuzzy relation are clearly defined in Zadeh (1971), and Klir and Folger (1988).

The max-min composition operation for fuzzy relations is as follows (Zadeh, 1991; Klir and Folger, 1988):

$$\mu_{P\circ Q}(x,z) = \max_{y\in Y} \min\left[\mu_P(x,y), \mu_Q(y,z)\right] \tag{19.15.9}$$

for all x in X, y in Y, and z in Z, where the composition of the two binary relations P(X,Y) and Q(Y,Z) is defined as follows:

$$R(X,Z) = P(X,Y) \circ Q(Y,Z) \tag{19.15.10}$$

As an example, consider the two binary relations

$$P(X,Y) = \left\{1.0|(4,4), 0.7|(4,7), 0.6|(8,4), 0.9|(8,7), 0.3|(11,4), 0.6|(11,7)\right\}$$

$$Q(Y,Z) = \{0.8|(4,6),\ 0.5|(4,9),\ 0.2|(4,12),\ 0.0|(4,15),\ 0.9|(7,6),\ 0.8|(7,9),\ 0.5|(7,12),\ 0.2|(7,15)\}$$

The following matrix equations illustrate the max-min composition for these binary relations

$$\begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \circ \begin{bmatrix} 0.8 & 0.5 & 0.2 & 0.0 \\ 0.9 & 0.8 & 0.5 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.7 & 0.5 & 0.2 \\ 0.9 & 0.8 & 0.5 & 0.2 \\ 0.6 & 0.6 & 0.5 & 0.2 \end{bmatrix}$$

Zadeh (1971) and Klir and Folger (1988), define also an alternative form of operation on fuzzy relations, called max-product composition. It is denoted as $P(X,Y) \otimes Q(Y,Z)$ and is defined by

$$\mu_{P\otimes Q}(x,z) = \max_{y\in Y}\left[\mu_P(x,y), \mu_Q(y,z)\right] \tag{19.15.11}$$

for all x in X, y in Y, and z in Z. The matrix equation

$$\begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \times \begin{bmatrix} 0.8 & 0.5 & 0.2 & 0.0 \\ 0.9 & 0.8 & 0.5 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.7 & 0.5 & 0.2 \\ 0.9 & 0.8 & 0.5 & 0.2 \\ 0.6 & 0.6 & 0.5 & 0.2 \end{bmatrix}$$

illustrates the max product composition for the pair of binary relations P(X,Y) and Q(Y,Z) previously considered.

A crisp binary relation among the elements of a single set can be denoted by R(X,X). If this relation is reflexive, symmetric, and transistive, it is called an equivalence relation (Klir and Folger, 1988).

A fuzzy binary relation S that is reflexive

$$\mu_S(x,x) = 1 \tag{19.15.12}$$

symmetric

$$\mu_S(x,y) = \mu_S(y,x) \tag{19.15.13}$$

and transitive

$$\mu_S(x,z) = \max_y \min\left[\mu_S(x,y), \mu_S(y,z)\right] \tag{19.15.14}$$

is called a similarity relation (Zadeh, 1971). Equations (19.15.12), (19.15.13), and (19.15.14) are valid for all x,y,z in the domain of S. A similarity relation is a generalization of the notion of equivalence relation.

Fuzzy orderings play a very important role in decision-making in a fuzzy environment. Zadeh (1971) defines fuzzy ordering as a fuzzy relation which is transitive. Fuzzy partial ordering, fuzzy linear ordering, fuzzy preordering, and fuzzy weak ordering are also mathematically defined by Zaheh (1971) and Zimmermann (1991).

The notion of fuzzy relation equation, proposed by Sanchez (1976), is an important notion with various applications. In the context of the max-min composition of two binary relations P(X,Y) and Q(Y,Z), the fuzzy relation equation is as follows

$$\mathbf{P} \circ \mathbf{Q} = \mathbf{R} \tag{19.15.15}$$

where **P** and **Q** are matrices of membership functions $\mu_P(x,y)$ and $\mu_Q(y,z)$, respectively, and **R** is a matrix whose elements are determined from Equation (19.15.9). The solution in this case in unique. However, when **R** and one of the matrices **P**, **Q** are given, the solution is neither guaranteed to exit nor to be unique (Klir and Folger, 1988).

Another important notion is the notion of fuzzy measure. It was introduced by Sugeno (1977). A fuzzy measure is defined by a function which assigns to each crisp subset of X a number in the unit interval [0,1]. This member represents the ambiguity associated with our belief that the crisp subset of X belongs to the subset A. For instance, suppose we are trying to diagnose a mechanical system with a failed component. In other terms, we are trying to assess whether this system belongs to the set of systems with, say, safety problems with regard to failure, serviceability problems with respect to deflections, and serviceability problems with respect to vibrations. Therefore, we might assign a low value, say 0.2 to failure problems, 0.3 to deflection problems, and 0.8 to vibration problems. The collection of these values constitutes a fuzzy measure of the state of the system.

Other measures including plausibility, belief, probability, and possibility measures are also used for defining the ambiguity associated with several crisp defined alternatives. For an excellent treatment of these measures and of the relationship among classes of fuzzy measures see Klir and Folger (1988).

Measures of fuzziness are used to indicate the degree of fuzziness of a fuzzy set (Zimmermann, 1991). One of the most used measures of fuzziness is the entropy. This measure is defined (Zimmermann, 1991) as

$$d(A) = h\sum_{i=1}^{n} S\left(\mu_A(x_i)\right) \tag{19.15.16}$$

where h is a positive constant and $S(\alpha)$ is the Shannon function defined as
$S(\alpha) = -\alpha \ln\alpha - (1-\alpha)\ln(1-\alpha)$ for rational α. For the fuzzy set in Figure 19.15.4(a), defined as

$$A = \{0.8|-2,\ 0.6|2,\ 0.2|3,\ 0.4|4,\ 0.5|5\}$$

the entropy is

$$\begin{aligned} d(A) &= h(0.5004 + 0.6730 + 0.5004 + 0.6730 + 0.6931) \\ &= 3.0399\ h \end{aligned}$$

Therefore, for h = 1, the entropy of the fuzzy set A is 3.0399.

The notion of linguistic variable, introduced by Zadeh (1973), is a fundamental notion in the development of fuzzy logic and approximate reasoning. According to Zadeh (1973), linguistic variables are "variables whose values are not members but words or sentences in a natural or artificial language. The motivation for the use of words or sentences rather than numbers is that linguistic characterizations are, in general, less specific than numerical ones." The main differences between fuzzy logic and classical two-valued (e.g., true or false) or multivalued (e.g., true, false, and indeterminate) logic are that (a) fuzzy logic can deal with fuzzy quantities (e.g., most, few, quite a few, many, almost all) which are in general represented by fuzzy numbers (see Figure 19.15.11), fuzzy predicates (e.g., expensive, rare), and fuzzy modifiers (e.g., extremely, unlikely), and (b) the notions of truth and false are both allowed to be fuzzy using fuzzy true/false values (e.g., very true, mostly false). As Klir and Folger (1988) stated, the ultimate goal of fuzzy logic is to provide foundations for approximate reasoning. For a general background on fuzzy logic and approximate reasoning and their applications to expert systems, the reader is referred to Zadeh (1973, 1987), Kaufmann (1975), Negoita (1985), and Zimmermann (1991), among others.

Decision making in a fuzzy environment is an area of continuous growth in engineering and other fields such as economics and medicine. Bellman and Zadeh (1970) define this process as a "decision

process in which the goals and/or the constraints, but not necessarily the system under control, are fuzzy in nature."

According to Bellman and Zadeh (1970), a fuzzy goal G associated with a given set of alternatives X = {x} is identified with a given fuzzy set G in X. For example, the goal associated with the statement "x should be in the vicinity of 50" might be represented by a fuzzy set whose membership function is equal to one of the four membership functions shown in Figure 19.15.11. Similarly, a fuzzy constraint C in X is also a fuzzy set in X, such as "x should be substantially larger than 20."

Bellman and Zadeh (1970) define a fuzzy decision D as the confluence of goals and constraints, assuming, of course, that the goals and constraints conflict with one another. Situations in which the goals and constraints are fuzzy sets in different spaces, multistage decision processes, stochastic systems with implicitly defined termination time, and their associated optimal policies are also studied in Bellman and Zadeh (1970).

References

Bellman, R.E. and Zadeh, L.A. 1970. Decision-making in a fuzzy environment. *Management Science*, 17(4), 141–164.

Dubois, D. and Prade, H. 1980. *Fuzzy Sets and Systems: Theory and Applications.* Academic Press, New York.

Furuta, H. 1995. Fuzzy logic and its contribution to reliability analysis. In *Reliability and Optimization of Structural Systems*, R. Rackwitz, G. Augusti, and A. Borri, Eds., Chapman & Hall, London, pp. 61–76.

Kaufmann, A. 1975. *Introduction to the Theory of Fuzzy Subsets*, Vol. 1, Academic Press, New York.

Klir, G.J. and Folger, T.A. 1988. *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, Englewood Cliffs, New Jersey.

Negoita, C.V. 1985. *Expert Systems and Fuzzy Systems*, Benjamin/Cummings, Menlo Park, California.

Sanchez, E. 1976. Resolution of composite fuzzy relation equations. *Information and Control*, 30, 38–48.

Sugeno, M. 1977. Fuzzy measures and fuzzy integrals — a survey, in *Fuzzy Automata and Decision Processes*, M.M. Gupta, R.K. Ragade, and R.R. Yager, Eds., North Holland, New York, pp. 89–102.

Zadeh, L.A. 1965. Fuzzy sets. *Information and Control*, 8, 338–353.

Zadeh, L.A. 1971. Similarity relations and fuzzy orderings. *Information Sciences.* 3, 177–200.

Zadeh, L.A. 1973. The concept of a linguistic variable and its applications to approximate reasoning. Memorandum ERL-M 411, Berkeley, California.

Zadeh, L.A. 1987. *Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh*, R.R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen, Eds., John Wiley & Sons, New York.

Zimmerman, H.-J. 1991. *Fuzzy Set Theory – and Its Applications*, 2nd ed., Kluwer Academic Publishers, Boston.

Further Information

The more than 5000 publications that exist in the field of fuzzy sets are widely scattered in many books, journals, and conference proceedings. For newcomers, good introductions to the theory and applications of fuzzy sets are presented in (a) *Introduction to the Theory of Fuzzy Sets*, Volume I, Academic Press, New York, 1975, by Arnold Kaufmann; (b) *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York, 1980, by Didier Dubois and Henri Prade; (c) *Fuzzy Sets, Uncertainty and Information*, Prentice Hall, Englewood Cliffs, NJ, 1988, by George Klir and Tina Folger, and (d) *Fuzzy Set Theory and Its Applications*,, 2nd ed., Kluwer Academic Publishers, Boston, 1991, by H.-J. Zimmerman, among others.

The eighteen selected papers by Lotfi A. Zadeh grouped in *Fuzzy Sets and Applications*, John Wiley & Sons, New York, 1987, edited by R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen are particularly helpful for understanding the developments of issues in fuzzy set and possibility theory. Also, the interview with Professor Zadeh published in this book illustrates the basic philosophy of the founder of fuzzy set theory.