

Whirlwind Exposition of Mathematics

In this appendix we will very quickly review all the mathematical background needed for complete understanding of the text. Depending on who you are, this chapter may be entirely superfluous, or it may be one of the most useful chapters in the entire book. You should probably at least look it over before starting to read Chapter 1. If most of the material in this chapter is unfamiliar to you, then you are probably not ready to continue reading. You should definitely consult it whenever you feel uncomfortable with the mathematics being used in any of the chapters, and it is written to be useful in a more general setting. Under no conditions should you read it in its natural place, after the last chapter; if you have already finished the book, you don't need it any more.

A.1 Numbers

Since DSP involves a lot of 'number crunching', we had better at least know what a number is! The simplest type of number is the 'whole number' or positive integer. These are 1, 2, 3, . . . You probably learned about them in kindergarten. Kronecker (the same guy who invented the delta) once said that the whole numbers were created by God, while all the rest are human inventions. Indeed the whole numbers are taken as basic entities in most of mathematics, but in axiomatic set theory their existence can actually be derived based on even simpler axioms.

So how did people create the rest of the numbers? The basic idea is to write equations using whatever numbers we already have, and try to solve them. Whenever we can't solve an equation using the numbers we already know about, we invent new ones. For instance, $1 + 1 = x$ leads us to $x = 2$, which is a whole number and thus no news, but when we try $x + 1 = 1$ we discover we need the first extension—we have to invent 'zero'. In case you think zero is no big deal, try writing large numbers in Hebrew numerals (where 10, 20, . . . 90 have their own symbols) or dividing Roman numerals.

Next we try to solve $x + 1 = 0$ and discover that we must invent -1 , an idea even more abstract than zero (you might recall how negative numbers perplexed you in grade school). Continuing in this fashion we discover all the ‘integers’ $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$

Now we try solving equations which contain multiplications. $2x = 4$ causes no problems, but $4x = 2$ does. We are thus led to the discovery of fractions, which together with the integers form all of the ‘rational numbers’.

We next try solving equations involving powers: $x^2 = 4$ is easy, but $x^2 = 2$ leads us to difficulties. The idea of $\sqrt{2}$ not being a rational number was once considered so important that the Pythagoreans killed to keep it secret. It turns out that not only aren’t these ‘irrational numbers’ rare, but there are more of them than there are rationals among the ‘real numbers’.

We’re almost done. The final kind of equation to observe is $x^2 = -1$, which leads to $i = \sqrt{-1}$, to the imaginary numbers, and to the combination of everything we have seen so far—the ‘complex numbers’. It turns out that complex numbers are sufficient to solve all equations expressible in terms of complex numbers, so our search is over.

EXERCISES

- A.1.1 Prove that $\sqrt{2}$ is irrational. (Hint: Assume that $\sqrt{2} = \frac{n}{m}$ and find a contradiction.)
- A.1.2 Prove that the set of real numbers is not denumerable, and that most real numbers are irrational.
- A.1.3 Hamilton invented ‘quaternions’, which are like complex numbers but with four real components. Why did he do this if complex numbers are sufficient?

A.2 Integers

Although most of us are used to decimal numbers, where we count from 1 to 9 before incrementing the next decimal place to the left, digital computers prefer binary numbers. Counting from zero up in binary numbers is done as follows.

0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111,
1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111, ...

Each 0 or 1 is called a bit, the rightmost bit in a number is called the Least Significant Bit (LSB), while the leftmost bit in a number (which can be zero since we assume that a constant number of bits are used) is the Most Significant Bit (MSB). There are several ways of extending binary numbers to negative numbers without using a separate minus sign, the most popular of which is two's complement. The two's complement of a number with $b + 1$ bits (the MSB is interpreted as the sign) is obtained by subtracting the number from 2^{b+1} ; hence addition of negative numbers is automatically correct assuming we just discard the overflow bit. We assume that the reader is reasonably proficient in using the integers, including the operations addition, subtraction, multiplication, division with remainder, and raising to a power, (particularly in binary) and understands the connection between binary arithmetic and logic, and how all this facilitates the building of digital computers.

There is another operation over the integers that we will require. We say that two whole numbers i and j are 'equal modulo' m

$$i = j \pmod{m} \tag{A.1}$$

if when they are divided by m they give the same remainder. This operation principle can be extended to real numbers as well, and is related to *periodicity*. Given an integer i , the 'reduction modulo' m of i

$$i \pmod{m} = j \tag{A.2}$$

means finding the minimum whole number j to which i is equal modulo m . Thus $15 = 8 \pmod{7}$ since $15 \pmod{7} = 1$ and $8 \pmod{7} = 1$.

If i divided by m leaves no remainder (i.e., $i \pmod{m} = 0$), we say that m is a 'factor' of i . A whole number is prime if it has no factors other than itself and 1. The 'fundamental theorem of arithmetic' states that every whole number has a unique factorization as the product of powers of primes.

$$i = p_1^{n_1} \cdot p_2^{n_2} \cdots p_m^{n_m} \tag{A.3}$$

A set is said to be 'finite' if the number of its elements is some whole number. A set is said to be 'denumerably infinite' if its elements can be placed in a list labeled by whole numbers. The interpretation is that there are in some sense the same number of elements as there are whole numbers. In particular the set of all integers is denumerable, since it can be listed in the following way,

$$a_1 = 0, a_2 = 1, a_3 = -1, a_4 = 2, a_5 = -2, \dots a_{2k} = k, a_{2k+1} = -k, \dots$$

and the set of all rational numbers between 0 and 1 is denumerable, as can be seen by the following order.

$$0, 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \dots$$

The set of all real numbers is nondenumerably infinite.

EXERCISES

- A.2.1 Show that there are an infinite number of primes. (Hint: Assume that there is a largest prime and find a contradiction.)
- A.2.2 You are given two input electrical devices that perform AND, OR, and NOT on bits. Show how to build a binary adder that inputs two 2-bit numbers and outputs a 3-bit number. How can this be extended to b -bit numbers?
- A.2.3 In one's complement notation the negative of a number is obtained by flipping all its bits. What are the advantages and disadvantage of this method?

A.3 Real Numbers

The reader should also know about real numbers, including the operations addition, subtraction, multiplication, division, and raising to an integer power. Some reals are rational (i.e., can be written as the ratio of two integers), but most are not.

Rational numbers can be represented as binary numbers with a decimal point (or should it be called a binary point?); for example, the decimal number $\frac{1}{2}$ is written 0.1, $\frac{1}{4}$ is 0.01, and $\frac{3}{4}$ is 0.11. This is called 'fixed point' notation. Some rational numbers and all irrational numbers require an infinite number of bits to the right of the point, and must be truncated in all practical situations when only a finite number of bits is available. Such truncation leads to numeric error. In order to increase the range of real numbers representable without adding too many bits, 'floating point' notation can be used. In floating point notation numbers are multiplied by positive or negative powers of 2 until they are between 0 and 1, and the power (called the exponent) and fraction (called the mantissa) are used together. For example, $\frac{3}{256} = 3 \cdot 2^{-8}$ is represented by mantissa 3 and binary exponent -8 .

Two specific irrational numbers tend to turn up everywhere.

$$\begin{aligned} e &= \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots \approx 2.718281828 \\ \pi &= 4 \tan^{-1}(1) = 4 \left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots\right) \approx 3.141592653 \end{aligned}$$

These series expansions are important from a theoretical point of view, but there are more efficient computational algorithms for approximating these numbers.

EXERCISES

- A.3.1 Compare several methods for computing e and π . See exercise A.8.5 below.
- A.3.2 How can you tell a rational number from an irrational one based on its binary representation?
- A.3.3 Another interesting irrational number is the golden ratio $\gamma = \frac{1+\sqrt{5}}{2} \approx 1.618$. Show that if a line segment of length l is divided in two segments of lengths a and b such that the ratio of l to a equals the ratio of a to b , then $\frac{a}{b} = \gamma$. Show that if a nonsquare rectangle has sides of length a and b such that if a square is removed the remaining rectangle has the same proportions, then $\frac{a}{b} = \gamma$. Show that $\cos\left(\frac{\pi}{5}\right) = \frac{\gamma}{2}$.
- A.3.4 Given a decimal representation r and a tolerance ϵ , how can the smallest a and b such that $r \approx \frac{a}{b}$ to within ϵ be found?

A.4 Complex Numbers

We assume that the reader has some knowledge of complex numbers, including how to convert a complex number z between the Cartesian form $z = x + iy$ and the polar form $z = re^{i\theta}$.

$$\begin{aligned} |z| &= \sqrt{x^2 + y^2} & \theta &= \tan^{-1}\left(\frac{y}{x}\right) \\ x &= z \cos \theta & y &= z \sin \theta \end{aligned} \quad (\text{A.4})$$

We will use the notations

$$x = \Re z \quad y = \Im z \quad r = |z| \quad \theta = \angle z$$

for the real part, imaginary part, absolute value (magnitude) and angle of the complex number z .

The arctangent function $\tan^{-1}(\varphi)$ is usually defined only for $-\frac{\pi}{2} < \varphi < \frac{\pi}{2}$. For equation (A.4) we need the 'four-quadrant arctangent', computable via the following algorithm:

```

a ← tan-1(y/x)
if x < 0
  if y > 0
    a ← a + π
  else
    a ← a - π
if a < 0
  a ← a + 2π

```

The complex operations of addition and multiplication are simple when addition is performed on the Cartesian form

$$z = z_1 + z_2 \quad \text{means} \quad \begin{aligned} x &= x_1 + x_2 \\ y &= y_1 + y_2 \end{aligned}$$

and multiplication in polar form,

$$z = z_1 z_2 \quad \text{means} \quad \begin{aligned} r &= r_1 r_2 \\ \theta &= \theta_1 + \theta_2 \end{aligned}$$

although multiplication can be done on the Cartesian forms as well.

$$z = z_1 z_2 \quad \text{means} \quad \begin{aligned} x &= x_1 x_2 - y_1 y_2 \\ y &= x_1 y_2 + x_2 y_1 \end{aligned}$$

Raising to a power, like multiplication, is also simplest in polar form, and in this form is called DeMoivre's theorem.

$$(r e^{i\theta})^a = r^a e^{ia\theta} \quad (\text{A.5})$$

There is a certain isomorphism between complex numbers and a two-dimensional vector space, but multiplication isn't defined in the same way for the two and complex numbers can't be extended to three dimensions. Nonetheless, it is often useful to picture the complex numbers as residing in a plane, called the 'complex plane', especially when dealing with functions defined over the complex plane.

Euler discovered a most elegant relation between four important numbers, -1 , i , e , and π , namely

$$e^{i\pi} = -1 \quad (\text{A.6})$$

which is a special case of the more general connection between imaginary exponentials and sinusoids,

$$e^{it} = \cos t + i \sin t \quad (\text{A.7})$$

a relation that can be reversed as well.

$$\sin(t) = \frac{e^{it} - e^{-it}}{2i} \quad \cos(t) = \frac{e^{it} + e^{-it}}{2} \quad (\text{A.8})$$

A very important family of complex numbers are the N^{th} ‘roots of unity’. These are the N solutions to equation $W^N = 1$. Thus for $N = 2$ we have the two square roots of unity $W = \pm 1$, while for $N = 4$ the four fourth roots of unity are $W = \pm 1, \pm i$. It is obvious that the N^{th} roots must all reside on the unit circle, $|W| = 1$, and it is not hard to show that they are given by $W = e^{i\frac{2\pi n}{N}} = W_N^n$, where the principle root is:

$$W_N = e^{i\frac{2\pi}{N}} \quad (\text{A.9})$$

EXERCISES

A.4.1 When a complex multiplication is performed using the Cartesian forms, it would seem that we need to perform four multiplications and two additions. Show that this same multiplication can be performed using three multiplications and five additions.

A.4.2 Express the power of a complex number in Cartesian form.

A.4.3 Find the square roots of i in Cartesian form.

A.4.4 Give geometric interpretations for the following:

1. All complex numbers with the same magnitude
2. All complex numbers with the same real part
3. All complex numbers with the same imaginary part
4. All complex numbers with the same angle
5. All complex numbers equidistant from a given complex number

A.5 Abstract Algebra

Numbers have so many different characteristics that it is hard to study them all at one time. For example, they have many inherent features (e.g., absolute value, positiveness), there are several operations that can be performed between two numbers (e.g., addition and multiplication), and these operations have many attributes (e.g., commutativity, associativity). As is customary in such complex situations mathematicians start their investigations with simpler objects that have only a small number of the many characteristics, and then advance to more and more complex systems.

The simplest such system is a ‘group’, which is a set of elements between which a single binary operation \cdot is defined. This operation must have the following properties:

closure: for all a and b in the group, $c = a \cdot b$ is in the group

associativity: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$

identity: there is a unique element i in the group such that $a \cdot i = i \cdot a = a$ for all a

inverse: for every a in the group there is a unique element b in the group such that $a \cdot b = b \cdot a = i$ where i is the identity element.

If in addition the operation obeys:

commutativity: $a \cdot b = b \cdot a$

then we call the group ‘commutative’ or ‘Abelian’.

The integers, the rationals, the real numbers, and the complex numbers are all groups with respect to the operation of addition; zero is the identity and $-a$ is the inverse. Likewise the set of polynomials of degree n (see Appendix A.6) and $m \times n$ matrices (Appendix A.15) are groups with respect to addition. The nonzero reals and complex numbers are also groups with respect to multiplication, with unity being the identity and $\frac{1}{a}$ the inverse. Not all groups have an infinite number of elements; for any prime number p , the set consisting of the integers $0, 1, \dots, (p - 1)$ is a finite group with p elements if we use the operation $a \cdot b \equiv (a + b) \pmod{p}$.

A field is more complex than a group in that it has two operations, usually called addition and multiplication. The field is a group under both operations, and in addition a new relation involving both addition and multiplication must hold.

distributivity: $a \cdot (b + c) = a \cdot b + a \cdot c$

The real numbers are a field, as are the rationals and the complex numbers. There are also finite fields (e.g., the binary numbers and more generally the integers $0 \dots p - 1$ under modulo arithmetic).

Given a field we can define a ‘vector space’ over that field. A vector space is a set of elements called ‘vectors’; our convention is to symbolize vectors by an underline, such as \underline{v} . The elements of the field are called ‘scalars’ in this context. Between the vectors in a vector space there is an operation of addition; and the vectors are a commutative group under this operation. In addition, there is a multiplication operation between a scalar and a vector that yields a vector.

Multiplication by unity must yield the same vector

- $1\underline{v} = \underline{v}$

and several types of distributivity must be obeyed.

- $a(\underline{u} + \underline{v}) = a\underline{u} + a\underline{v}$
- $(a + b)\underline{v} = a\underline{v} + b\underline{v}$
- $(ab)\underline{v} = a(b\underline{v})$

There is another kind of multiplication operation that may be defined for vector spaces that goes under several names including scalar product, inner product, and dot product. This operation is between two vectors and yields a scalar. If the underlying field is that of the reals, the dot product must have the following properties:

nonnegativity: $\underline{u} \cdot \underline{v} \geq 0$

self-orthogonality: $\underline{v} \cdot \underline{v} = 0$ if and only if $\underline{v} = \underline{0}$

commutativity: $\underline{u} \cdot \underline{v} = \underline{v} \cdot \underline{u}$

distributivity: $(\underline{u} + \underline{v}) \cdot \underline{w} = \underline{u} \cdot \underline{w} + \underline{v} \cdot \underline{w}$

scalar removal: $(a\underline{u}) \cdot \underline{v} = a(\underline{u} \cdot \underline{v})$

Two vectors for which $\underline{u} \cdot \underline{v} = 0$ are called ‘orthogonal’. If the underlying field is of the complex numbers, the commutativity relation requires modification.

conjugate commutativity: $\underline{u} \cdot \underline{v} = (\underline{v} \cdot \underline{u})^*$

The prototypical example of a vector space is the set of ordered n -tuples of numbers, and this used as the definition of ‘vector’ in computer science. The number n is called the dimension and the operations are then defined by the following recipes:

- $(u_1, u_2 \dots u_n) + (v_1, v_2 \dots v_n) = (u_1 + v_1, u_2 + v_2, \dots u_n + v_n)$
- $a(v_1, v_2 \dots v_n) = (av_1, av_2, \dots av_n)$
- $(u_1, u_2 \dots u_n) \cdot (v_1, v_2 \dots v_n) = u_1v_1 + u_2v_2 + \dots u_nv_n$

The usual two-dimensional and three-dimensional vectors are easily seen to be vector spaces of this sort. We can similarly define vector spaces of any finite dimension over the reals or complex numbers, and by letting n go to infinity we can define vector spaces of denumerably infinite dimension. These ‘vectors’ are in reality infinite sequences of real or complex numbers. It is also possible to define vector spaces with nondenumerable dimension, but then the interpretation must be that of a function defined on the real axis, rather than an n -tuple of numbers or an infinite sequence.

A ‘metric space’ is a set of elements, between every two of which is defined a metric (distance). The metric is a nonnegative number $d(x, y)$ that has the following three properties:

symmetry: $d(y, x) = d(x, y)$,

identity: $d(x, y) = 0$ if and only if $x = y$,

triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Metric spaces and linear vector spaces capture different aspects of Euclidean vectors, and it is not surprising that we can define ‘normed spaces’ that are both metric and vector spaces. The norm of a vector is defined to be $|v| = \sqrt{v \cdot v}$, which is easily seen to be a nonnegative number and to fulfill all the requirements of a metric.

EXERCISES

- A.5.1 Find groups with small numbers of elements.
- A.5.2 Show that **true** and **false** with **or** as addition and **and** as multiplication form a vector space.
- A.5.3 Prove formally that three-dimensional space is a vector space and a metric space.

A.6 Functions and Polynomials

Functions uniquely map one or more numbers (called arguments) onto other numbers (called returned values). For example, the function $f(x) = x^2$ returns a unique real number for every real argument x , although all positive x^2 are returned by two different real arguments (x and $-x$) and negative numbers are not returned for any real argument. Hence $f(x) = \sqrt{x}$ is not a function unless we define it as returning only the positive square root (since a function can't return two values at the same time) and even then it is undefined for negative arguments unless we allow it to return complex values.

A 'symmetry' of a function is a transformation of the argument that does not change the returned value. For example, $x \rightarrow -x$ is a symmetry of the function $f(x) = x^2$ and $x \rightarrow x + 2\pi n$ for any n are symmetries of the function $f(x) = \sin(x)$.

'Polynomials' are functions built by weighted summing of powers of the argument

$$a(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots a_nx^n \quad (\text{A.10})$$

the weights a_i are called 'coefficients' and the highest power n is called the 'degree' of the polynomial.

The straightforward algorithm for evaluating polynomials,

```

a ← a0
for i ← 1 to n
  p ← xi
  a ← a + aip

```

is usually not the best way of computing the value to be returned, since raising to a power is computationally expensive and may introduce numerical error. It is thus usually better to use the following algorithm, which requires an additional memory location.

```

a ← a0 + a1x
p ← x
for i ← 1 to n
  p ← p * x
  a ← a + aip

```

Even this code is not optimal, but Horner's rule

```

a ← an
for i ← n - 1 to 0
    a ← ax + ai

```

requiring only n multiplications and additions, is optimal.

Polynomials can be added, multiplied, and factored into simpler polynomials. The 'fundamental theorem of algebra' states that all polynomials with real coefficients can be uniquely factored into products of first- and second-degree polynomials with real coefficients, and into products of first-degree polynomials with complex coefficients. For example,

$$x^3 + x^2 + x + 1 = (x^2 + 1)(x + 1)$$

for real coefficients, while allowing complex coefficients we can factor further.

$$z^3 + z^2 + z + 1 = (z + i)(z - i)(z + 1)$$

A first-degree factor of $a(z)$ can always be written $z - \zeta$, in which case ζ is called a 'zero' (or 'root') of the polynomial. It is obvious that the polynomial as a whole returns zero at its zeros $a(\zeta) = 0$ and that the number of zeros is equal to the polynomial degree for complex polynomials (although some zeros may be identical), but may be less for real polynomials.

'Rational functions' are functions formed by dividing two polynomials.

$$r(x) = \frac{a(x)}{b(x)} = \frac{a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n}{b_0 + b_1x + b_2x^2 + b_3x^3 + \dots + b_mx^m} \quad (\text{A.11})$$

A 'zero' of a rational function is an argument for which the function returns zero, and is necessarily a zero of the numerator polynomial. A 'pole' of a rational function is a zero of the denominator polynomial and hence an argument for which the function as a whole is infinite.

EXERCISES

- A.6.1 The derivative of a polynomial $a(x)$ of degree n is a polynomial of degree $n - 1$ given by $a'(x) \equiv a_1 + 2a_2x + 3a_3x^2 + \dots + na_nx^{n-1}$. What is the most efficient method of simultaneously computing $a(x)$ and $a'(x)$?
- A.6.2 Horner's rule is not efficient for sparse polynomials which have many zero coefficients. For example, the best way to compute $p(x) = x^5$ is to compute $a_2 \leftarrow x^2, a_4 \leftarrow c^2, p(x) = a_4x$. What is the best way of computing $p(x) = x^n$ for general integer n ?

- A.6.3 Show that rational functions are uniquely determined by their zeros and poles (including multiplicities) and a single additional number.
- A.6.4 We define binary polynomials as polynomials for which each power of x is either present (i.e., its coefficient is 1) or absent (its coefficient is 0). How many different binary polynomials are there with degree up to m ? What is the connection between these polynomials and the nonnegative integers? The addition of two binary polynomials is defined by addition modulo 2 of the corresponding coefficients (note that each polynomial is its own additive inverse). To what operation of the integers does this correspond? How do you think polynomial multiplication should be defined?

A.7 Elementary Functions

In addition to polynomials there are several other functions with which the reader should feel comfortable. The natural logarithm $\ln(x)$ is one such function. It is defined for positive real numbers, and is uniquely determined by the properties

$$\begin{aligned}\ln(1) &= 0 \\ \ln(ab) &= \ln(a) + \ln(b) \\ \ln\left(\frac{a}{b}\right) &= \ln a - \ln b \\ \ln(a^b) &= b \ln(a)\end{aligned}\tag{A.12}$$

although it can also be defined by an integral representation.

$$\ln x = \int_1^x \frac{1}{t} dt\tag{A.13}$$

One can generalize the logarithm to complex numbers as well,

$$\ln(re^{i\theta}) = \ln r + i\theta\tag{A.14}$$

and one finds that $\ln(-1) = i\pi$, and $\ln(\pm i) = \pm i\frac{\pi}{2}$. Actually, this is only one possible value for the complex logarithm; any multiple of $2\pi i$ is just as good.

Logarithms transform multiplication into addition since they are the converse operation to raising to a power. The natural logarithms are logarithms to base e , that is, $y = \ln x$ means that $x = e^y$, or put in another way

$e^{\ln(a)} = a$ and $\ln(e^a) = a$. The function e^x will be discussed in a moment. It is often useful to know how to expand the natural logarithm around $x = 1$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots \quad (\text{A.15})$$

although this series converges very slowly.

Logarithms to other bases are related as follows.

$$\log_a x = \frac{\log_b x}{\log_b a} = \frac{\ln x}{\ln a}$$

The most important alternative bases are base 10 and base 2, and it is enough to remember $\ln 10 \approx 2.3$ and $\log 2 \approx 0.3$ to be able to mentally convert between them. Another logarithmic relation is the decibel (dB), being one-tenth of a Bel, which is simply the base 10 logarithm of a ratio.

$$r(\text{dB}) = 10 \log_{10} \frac{P_1}{P_2} \quad (\text{A.16})$$

Using one of the useful numbers we see that every factor of two contributes about 3 dB to the ratio (e.g., a ratio of two to one is about 3 dB, four to one is about 6 dB, eight to one about 9 dB, etc.). Of course a ratio of ten to one is precisely 10 dB.

The 'exponential function' e^x is simply the irrational number e raised to the x power. If x is not an integer, the idea of a power has to be generalized, and this can be done by requiring the following properties:

$$\begin{aligned} e^0 &= 1 \\ e^{a+b} &= e^a + e^b \\ e^{ab} &= (e^a)^b \end{aligned}$$

The solution turns out to be given by an infinite series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (\text{A.17})$$

and this same series can be used for complex numbers. To define noninteger powers of other numbers we can use

$$x^y \equiv e^{y \ln x} \quad (\text{A.18})$$

where \ln was defined above.

The Gaussian

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad (\text{A.19})$$

is another function based on the exponential. This function has a maximum at μ and a ‘width’ of σ , and is symmetric around μ . The peculiar constant is chosen so that its integral over all the argument axis is normalized to one.

$$\int_{-\infty}^{\infty} G(x) dx = 1$$

EXERCISES

- A.7.1 Generate three-dimensional plots of the complex exponential and the complex logarithm as surfaces over the complex plane.
- A.7.2 Derive the expansion (A.17) by requiring the derivative of the exponential function to equal itself.
- A.7.3 Prove the normalization of the Gaussian.

A.8 Trigonometric (and Similar) Functions

We assume that the reader is familiar with the basic trigonometric functions $\sin(x)$, $\cos(x)$, $\tan(x) = \frac{\sin(x)}{\cos(x)}$ and $\tan^{-1}(x)$, and their graphs, as well as the connection

$$\sin^2(x) + \cos^2(x) = 1 \quad (\text{A.20})$$

between these functions and the unit circle.

Perhaps their most fundamental property is *periodicity*

$$\begin{aligned} \sin(x + 2\pi n) &= \sin(x) \\ \cos(x + 2\pi n) &= \cos(x) \\ \tan(x + \pi n) &= \tan(x) \end{aligned} \quad (\text{A.21})$$

for all whole n , but there are various other symmetries as well.

$$\begin{aligned}
 \sin(-x) &= -\sin(x) \\
 \cos(-x) &= \cos(x) \\
 \sin\left(\frac{\pi}{2} + x\right) &= \cos(x) \\
 \cos\left(\frac{\pi}{2} + x\right) &= -\sin(x) \\
 \cos\left(\frac{\pi}{2} - x\right) &= \sin(x)
 \end{aligned}
 \tag{A.22}$$

In DSP we often need the ‘sum formulas’. The fundamental ones that we need quote are

$$\begin{aligned}
 \sin(a \pm b) &= \sin a \cos b \pm \cos a \sin b \\
 \cos(a \pm b) &= \cos a \cos b \mp \sin a \sin b
 \end{aligned}
 \tag{A.23}$$

from which we can derive ‘double angle formulas’

$$\begin{aligned}
 \sin(2a) &= 2 \sin a \cos a \\
 \cos(2a) &= \cos^2 a - \sin^2 a = 2 \cos^2 a - 1 = 1 - 2 \sin^2 a
 \end{aligned}
 \tag{A.24}$$

and the ‘square formulas’

$$\begin{aligned}
 \sin^2(a) &= \frac{1}{2} - \frac{1}{2} \cos(2a) \\
 \cos^2(a) &= \frac{1}{2} + \frac{1}{2} \cos(2a)
 \end{aligned}
 \tag{A.25}$$

a pair of identities that often come in handy. While not important for our purposes, for completeness we give

$$\tan(a \pm b) = \frac{\tan a \pm \tan b}{1 \mp \tan a \tan b}
 \tag{A.26}$$

We will also need another kind of sum formula.

$$\begin{aligned}
 \sin(a) + \sin(b) &= 2 \sin\left(\frac{1}{2}(a+b)\right) \cos\left(\frac{1}{2}(a-b)\right) \\
 \sin(a) - \sin(b) &= 2 \cos\left(\frac{1}{2}(a+b)\right) \sin\left(\frac{1}{2}(a-b)\right) \\
 \cos(a) + \cos(b) &= 2 \cos\left(\frac{1}{2}(a+b)\right) \cos\left(\frac{1}{2}(a-b)\right) \\
 \cos(a) - \cos(b) &= -2 \sin\left(\frac{1}{2}(a+b)\right) \sin\left(\frac{1}{2}(a-b)\right)
 \end{aligned}
 \tag{A.27}$$

Another relation derivable from the sum formulas that appears less frequently in trigonometry books but is very important in DSP is

$$a \sin(x) + b \cos(x) = A \sin(x + \theta)
 \tag{A.28}$$

which means that summing \sin and \cos of the same argument with any coefficients still leaves a simple \sin of the same argument. The desired relations are as follows.

$$\begin{aligned} a &= A \cos(\theta) & b &= A \sin(\theta) \\ A &= \sqrt{a^2 + b^2} & \theta &= \tan^{-1} \frac{b}{a} \end{aligned} \quad (\text{A.29})$$

On odd occasions it is useful to know other ‘multiple angle formulas’ such as

$$\begin{aligned} \sin(3a) &= 2 \sin(a) \cos(a) \\ \cos(3a) &= \cos^2(a) - \sin^2(a) = 2 \cos^2(a) - 1 = 1 - \sin^2(a) \\ \sin(4a) &= 2 \sin(a) \cos(a) \\ \cos(4a) &= \cos^2(a) - \sin^2(a) = 2 \cos^2(a) - 1 = 1 - \sin^2(a) \end{aligned} \quad (\text{A.30})$$

but these complex iterative forms can be replaced by simple two step recursions

$$\begin{aligned} \sin((k+1)a) &= 2 \cos(a) \sin(ka) - \sin((k-1)a) \\ \cos((k+1)a) &= 2 \cos(a) \cos(ka) - \cos((k-1)a) \end{aligned} \quad (\text{A.31})$$

the second of which is useful in deriving the Chebyshev polynomials $T_k(x)$, (see Appendix A.10).

From the sum and multiple angle formulas one can derive ‘product formulas’.

$$\begin{aligned} \sin(a) \sin(b) &= \frac{1}{2} (\cos(a-b) - \cos(a+b)) \\ \sin(a) \cos(b) &= \frac{1}{2} (\sin(a-b) + \sin(a+b)) \\ \cos(a) \cos(b) &= \frac{1}{2} (\cos(a-b) + \cos(a+b)) \end{aligned} \quad (\text{A.32})$$

Similarly, you may infrequently need to know further ‘power formulas’

$$\begin{aligned} \sin^3(a) &= \frac{3}{4} \sin(a) - \frac{1}{4} \sin(3a) \\ \cos^3(a) &= \frac{3}{4} \cos(a) + \frac{1}{4} \cos(3a) \\ \sin^4(a) &= \frac{3}{8} - \frac{1}{2} \cos(2a) + \frac{1}{8} \cos(4a) \\ \cos^4(a) &= \frac{3}{8} + \frac{1}{2} \cos(2a) + \frac{1}{8} \cos(4a) \end{aligned} \quad (\text{A.33})$$

(and more general formulas can be derived), but don’t bother trying to memorize these.

An important characteristic of the sines and cosines as functions is their mutual orthogonality

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(nt) \cos(mt) dt &= 0 \\ \int_{-\pi}^{\pi} \sin(nt) \sin(mt) dt &= \pi \delta_{n,m} \\ \int_{-\pi}^{\pi} \cos(nt) \cos(mt) dt &= \pi \delta_{n,m}\end{aligned}\tag{A.34}$$

as can be easily derived using the product formulas and direct integration (see Appendix A.9).

Sometimes it is useful to expand trigonometric functions in series. The two expansions

$$\begin{aligned}\sin(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \\ \cos(x) &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots\end{aligned}\tag{A.35}$$

are important and easy to remember. For really small x you can usually get away with the first x dependent term.

Using the trigonometric identities to simplify complex expressions is usually hard work. It's usually easier to replace real sinusoids with complex exponentials; use the simpler math of e^{ix} and take the real part at the end.

Just as the trigonometric functions are 'circular functions' in the sense that $x = \cos(\theta)$, $y = \sin(\theta)$ trace out a circle when θ goes from zero to 2π , so we can define the hyperbolic functions \sinh and \cosh that trace out a hyperbola with its vertex at $(0, 1)$. Similarly to equation (A.8), we define

$$\sinh(\theta) = \frac{e^{\theta} - e^{-\theta}}{2} \quad \cosh(\theta) = \frac{e^{\theta} + e^{-\theta}}{2} \quad \tanh(\theta) = \frac{\sinh \theta}{\cosh \theta}\tag{A.36}$$

and easily find the analog of equation (A.20),

$$\cosh^2(\theta) - \sinh^2(\theta) = 1\tag{A.37}$$

which proves that $x = \cosh(\theta)$, $y = \sinh(\theta)$ trace out a hyperbola. Unlike the circular functions, the hyperbolic functions are not periodic, but their expansions are similar to those of the circular functions.

$$\begin{aligned}\sinh(x) &= x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots \\ \cosh(x) &= 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots\end{aligned}\tag{A.38}$$

In addition to circular and hyperbolic functions, there are elliptical functions $\text{sn}(\varphi)$ and $\text{cn}(\varphi)$, which are defined in three steps. First we define the ‘Jacobian elliptical function’

$$u_k(\phi) \equiv \int_0^\phi \frac{dx}{\sqrt{1 - k^2 \sin^2(x)}} \quad (\text{A.39})$$

for real k in the range $0 \leq k \leq 1$ and nonnegative real φ . This integral arises in the determination of the length of an arc of an ellipse. There is a special case of the Jacobian elliptical function, called the ‘complete elliptical integral’

$$K_k \equiv u_k\left(\frac{\pi}{2}\right) = \int_0^{\frac{\pi}{2}} \frac{dx}{\sqrt{1 - k^2 \sin^2(x)}} \quad (\text{A.40})$$

that starts at $K_0 = \frac{\pi}{2}$ and increases monotonically, diverging as $k \rightarrow 1$. Second, we define the inverse Jacobian elliptical function $\varphi_k(u)$ as the inverse formula to equation (A.39). Finally, we can define ‘elliptical sine’ and ‘elliptical cosine’ functions.

$$\begin{aligned} \text{sn}_k(u) &\equiv \sin\left(\varphi_k(u)\right) \\ \text{cn}_k(u) &\equiv \cos\left(\varphi_k(u)\right) \end{aligned} \quad (\text{A.41})$$

It is obvious from the definitions that

$$\text{sn}_k^2(u) + \text{cn}_k^2(u) = 1$$

and that for $k = 0$ they are identical to the trigonometric sine and cosine. Much less obvious is that for all $k < 1$ they remain periodic, but with period $4K_k$, four times the complete elliptical integral. As $k \rightarrow 1$ the elliptical sine gets wider until at $k = 1$ where its period diverges, becoming equal to the hyperbolic tangent function. As k increases from zero the elliptical cosine at first becomes more like a triangle wave, but after $k = \frac{1}{\sqrt{2}}$ it develops an inflection, and at $k = 1$ it becomes $\frac{1}{\cosh(u)}$. We will return to the elliptical functions in Appendix A.10.

EXERCISES

A.8.1 Plot the circular, hyperbolic, and elliptical sines and cosines. Describe the similarities and differences.

A.8.2 Prove:

- $\sinh(-x) = -\sinh(x)$
- $\cosh(-x) = \cosh(x)$
- $(\cosh(x) + \sinh(x))^n = \cosh nx + \sinh nx$
- $\sinh(z) = -i \sin(iz)$
- $\cosh(z) = \cos(iz)$
- $\sinh(x + 2\pi ki) = \sinh(x)$
- $\cosh(x + 2\pi ki) = \cosh(x)$

A.8.3 Prove that the derivative of $\sinh(x)$ is $\cosh(x)$ and that of $\cosh(x)$ is $\sinh(x)$.

A.8.4 Derive half-angle formulas for sine and cosine.

A.8.5 Use the half-angle formulas and the fact that $\text{sinc}(0) = 1$, that is $\frac{\sin(x)}{x} \rightarrow 1$ when $x \rightarrow 0$, to numerically calculate π . (Hint: $\text{sinc}(\frac{\pi}{n}) \rightarrow 1$ when $n \rightarrow \infty$ so $n \sin(\frac{\pi}{n}) \rightarrow \pi$ in this same limit; start with known values for sine and cosine when $n = 4, 5$, or 6 and iteratively halve the argument.)

A.8.6 Prove equation (A.34).

A.8.7 Find sum and double angle formulas for the hyperbolic functions.

A.8.8 Derive expansions (A.35) and (A.38) from equation (A.17).

A.9 Analysis

We assume that the reader is familiar with the sigma notation for sums

$$\sum_{i=0}^N a_i = a_0 + a_1 + a_2 + \dots + a_N \quad (\text{A.42})$$

and knows its basic properties:

$$\begin{aligned} \sum_i (a_i + b_i) &= \sum_i a_i + \sum_i b_i \\ \sum_i ca_i &= c \sum_i a_i \\ \sum_i \sum_j a_{ij} &= \sum_j \sum_i a_{ij} \\ \left(\sum_i a_i \right)^2 &= \sum_i \sum_j a_i a_j = 2 \sum_{i < j} a_i a_j + \sum_i a_i^2 \end{aligned}$$

When dealing with sums a particularly useful notation is that of the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (\text{A.43})$$

which selects a particular term from a sum.

$$\sum_i a_i \delta_{ik} = a_k \quad (\text{A.44})$$

Certain sums can be carried out analytically. The sum of an n -term 'arithmetic series' $a_1 = a, a_2 = 2a, \dots, a_k = ka, \dots, a_n = na$ is n times average value.

$$a + 2a + 3a + \dots + na = \sum_{k=1}^n ka = n \frac{a_1 + a_n}{2} = \frac{1}{2}n(n+1)a \quad (\text{A.45})$$

The sum of a geometric series $a_0 = 1, a_1 = r, \dots, a_k = r^k \dots, a_{n-1} = r^{n-1}$ is

$$1 + r + r^2 + \dots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{(1 - r^n)}{1 - r} = \frac{a_0 - ra_n}{1 - r} \quad (\text{A.46})$$

and for $-1 < r < 1$ this sum converges when we go to infinity.

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1 - r} \quad (\text{A.47})$$

An important particular case of this sum is

$$\sum_{k=0}^{n-1} e^{ak} = \frac{(1 - e^{an})}{1 - e^a} \quad (\text{A.48})$$

and the infinite sum converges for negative a .

A key idea in mathematical analysis is that of continuity of a function. A real-valued function of a single variable is said to be 'continuous' if it has no jumps, (i.e., if when approaching an input to the function t_0 from below and above we arrive at the same output). For continuous functions we can 'interpolate' to find values between those already seen, and if these previously seen values are close enough, the interpolated value will not be far off.

We can define the 'derivative' of a function by considering how fast it changes when we change its inputs. The ratio of the output change to the

input change approaches the derivative when the input changes become very small. It is assumed that the reader knows how to differentiate basic functions. In particular we will need the following derivatives:

$$\begin{aligned}\frac{d}{dt}t^n &= nt^{n-1} \\ \frac{d}{dt}e^{at} &= ae^{at} \\ \frac{d}{dt}\sin(\omega t) &= \omega \cos(\omega t) \\ \frac{d}{dt}\cos(\omega t) &= -\omega \sin(\omega t)\end{aligned}\tag{A.49}$$

The ‘integral’ of a function is related to the area under its plot. As such integrals can be approximated by Riemann sums

$$\int f(t) dt \approx \sum_n f(t_n)\delta\tag{A.50}$$

where the summation is over rectangles of width δ approximating the curve. The ‘fundamental theorem of calculus’ states that integration is the inverse operation to differentiation. It is assumed that the reader can do basic integrals, and, for example, knows the following:

$$\begin{aligned}\int t^n dt &= \frac{1}{n+1}t^{n+1} \\ \int e^{at} dt &= \frac{1}{a}e^{at} \\ \int \sin(\omega t) dt &= -\frac{1}{\omega}\cos(\omega t) \\ \int \cos(\omega t) dt &= \frac{1}{\omega}\sin(\omega t)\end{aligned}\tag{A.51}$$

In certain contexts we call the derivative is called the ‘density’; let’s understand this terminology. When we say that the density of water is ρ we mean that the weight of a volume v of water is ρv . In order to discuss functions of a single variable consider a liquid in a long pipe with constant cross-section; we can now define a ‘linear density’ λ , and the weight of the water in a length L of water is λL . For an inhomogeneous liquid whose linear density varies from place to place, (e.g., the unlikely mixture of mercury, ketchup, water and oil in a long pipe) we must use a position dependent density $\lambda(x)$. The total weight is no longer simply the density times the total length, but if the density varies slowly then the weight of a small

length Δx in the vicinity of position x is approximately $\lambda(x)\Delta x$. If the density varies rapidly along the pipe's length, all we can say is that the weight of an infinitesimal length dx in the vicinity of position x is $\lambda(x) dx$ so that the total weight of the first L units of length is the integral.

$$W(L) = \int_0^L \lambda(x) dx$$

From the fundamental theorem of calculus it is clear that the density function $\lambda(x)$ is the derivative of the cumulative weight function $W(L)$.

EXERCISES

- A.9.1 Show that $1 + 2 + 3 + \dots + n = \frac{1}{2}n(n+1)$ and that $1 + 3 + 5 + \dots = n^2$ (i.e., that every triangular number is a perfect square).
- A.9.2 Show that $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$ diverges, but that $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 2$.
- A.9.3 What is the meaning of a continuous function of a complex variable? Of a differentiable function?
- A.9.4 The shortest way to get from point $(0, 0)$ to point $(1, 1)$ in the two-dimensional plane is the straight line of length $\sqrt{2}$. Another way is to go first along the straight lines connecting the points $(0, 0) - (1, 0) - (1, 1)$, traversing a path of length 2. Similarly, the paths $(0, 0) - (\frac{1}{2}, 0) - (\frac{1}{2}, \frac{1}{2}) - (\frac{1}{2}, 1) - (1, 1)$, $(0, 0) - (\frac{1}{4}, 0) - (\frac{1}{4}, \frac{1}{2}) - (\frac{3}{4}, \frac{1}{2}) - (\frac{3}{4}, 1) - (1, 1)$, and indeed any path with segments parallel to the axes have total path length 2. In the limit of an infinite number of segments our path is indistinguishable from the straight line and so we have proven that $\sqrt{2} = 2$. What's wrong with this 'proof'?

A.10 Differential Equations

Differential equations are equations in which functions and their derivatives appear. The solution of an algebraic equation is a *number*, but the solution of a differential equation is a *function*. For example, given

$$s(t) = -\lambda \frac{ds(t)}{dt} \tag{A.52}$$

we can immediately guess that $s(t) = e^{-\lambda t}$. So exponentials are solutions of differential equations of the first order. Similarly,

$$s(t) = -\omega^2 \frac{d^2s}{dt^2} \tag{A.53}$$

has the solution $s(t) = A \sin(\omega t + \phi)$, so sinusoids are the solutions of differential equations of the second order.

There are many other equations that give birth to other ‘named’ functions. For example, Legendre’s differential equation

$$(1 - t^2) \frac{d^2 s(t)}{dt^2} - 2t \frac{ds(t)}{dt} + n(n + 1)s(t) = 0 \quad (\text{A.54})$$

for nonnegative integer n has as solutions the Legendre polynomials $P_n(t)$, the first few of which are given here.

$$\begin{aligned} P_0(t) &= 1 \\ P_1(t) &= t \\ P_2(t) &= \frac{1}{2}(3t^2 - 1) \\ P_3(t) &= \frac{1}{2}(5t^3 - 3t) \end{aligned} \quad (\text{A.55})$$

The general form is

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2 - 1)^n$$

showing that they are indeed polynomials of degree n . We can efficiently compute the returned value for argument t using a recursion.

$$(n + 1)P_{n+1}(t) = (2n + 1)tP_n(t) - nP_{n-1}(t)$$

The Legendre polynomials are akin to the sinusoids in that the n polynomials are odd, the even n ones are even, and they obey orthogonality.

$$\int_{-1}^1 P_n(t)P_m(t)dt = \frac{2}{2n + 1} \delta_{n,m}$$

Hence any function on $[-1 \dots +1]$ can be expanded $s(t) = \sum a_n P_n(t)$ where

$$a_n = \frac{2n + 1}{2} \int_{-1}^1 s(t)P_n(t)dt$$

is the coefficient of $P_n(t)$ in the expansion.

Another named equation is Chebyshev’s differential equation

$$(1 - t^2) \frac{d^2 s}{dt^2} - t \frac{ds}{dt} + n^2 s = 0 \quad (\text{A.56})$$

the solutions for which are called the ‘Chebyshev polynomials’.

$$T_n(t) = \begin{cases} \cos(n \cos^{-1} t) & |t| \leq 1 \\ \cosh(n \cosh^{-1} t) & \text{else} \end{cases} \quad (\text{A.57})$$

The notation T_n derives from an alternative Latinization of their discoverer's name (Pafnuty Lvovich Tshebyshev).

We presented above somewhat complex formulas for the cosine of multiple angles $\cos(ka)$ in terms of $\cos(a)$ (A.31). Let's define a sequence of operators T_k that perform just that transformation

$$T_k(\cos a) \equiv \cos(ka) \quad (\text{A.58})$$

which you can think of as a sneaky way of defining functions in $x = \cos a$.

$$T_k(x) = \cos(k \cos^{-1} x) \quad (\text{A.59})$$

These functions are only defined for x in the domain $-1 \leq x \leq 1$, and their range is $-1 \leq T_k(x) \leq 1$, but they are exactly the functions defined above.

It can easily be seen from either definition that

$$\begin{aligned} T_0 &= 1 \\ T_1(t) &= t \end{aligned}$$

but it is painful to derive (e.g., by using (A.31)) even the next few:

$$\begin{aligned} T_2(t) &= 2t^2 - 1 \\ T_3(t) &= 4t^3 - 3t \\ T_4(t) &= 8t^4 - 8t^2 + 1 \end{aligned}$$

but the job is made manageable by a recursion that we shall derive below.

The functions $T_N(x)$ have a further interesting property. $T_0(t)$, being unity, attains its maximum absolute value for all t ; $T_1(t)$ starts at $|T_1(-1)| = 1$ and ends at $|T_1(+1)| = 1$; $|T_2(t)| = 1$ at the three values $t = -1, 0, +1$. In general, all T_N have N equally spaced zeros at positions

$$t = \cos\left(\frac{\pi(k - \frac{1}{2})}{N}\right) \quad k = 1, 2, \dots, N \quad (\text{A.60})$$

and $N + 1$ equally spaced extrema where $|T_N(t)| = \pm 1$ at

$$t = \cos\left(\frac{\pi k}{N}\right) \quad k = 0, 1, 2, \dots, N \quad (\text{A.61})$$

in the interval $[-1 \dots +1]$. This is not totally unexpected for a function that was defined in terms of $\cos a$, and is called the *equiripple* property. Equiripple means that the functions oscillate in roughly sinusoidal fashion

between extrema of the same absolute magnitude. This characteristic makes these functions useful in minimax function approximation.

The reader will note with surprise that in all of the examples given above T_N was actually a *polynomial* in t . We will now show something truly astounding, that for *all* N $T_N(t)$ is a polynomial in x of degree N . This is certainly unexpected for functions defined via trigonometric functions as in (A.59), and were just shown to be roughly sinusoidal. Nothing could be less polynomial than that! The trick is equation (A.31) which tells us that

$$T_{N+1}(t) = 2tT_N(t) - T_{N-1}(t)$$

which coupled with the explicit forms for $T_0(t)$ and $T_1(t)$ is a simple recursive scheme that only generates polynomials. We can see from the form of the recursion that the highest term is exactly N , and that its coefficient will be precisely 2^{N-1} (at least for $N > 0$).

The eminent German astronomer Friedrich Wilhelm Bessel was the first to measure distances to the stars. He was the first to notice that the brightest star in sky, Sirius, executes tiny oscillations disclosing the existence of an invisible partner (Sirius B was observed after his death). He also observed irregularities in the orbit of Uranus that later led to the discovery of Neptune. During his 1817 investigation of the gravitational three-body problem, he derived the differential equation

$$t^2 \frac{d^2 s}{dt^2} + t \frac{ds}{dt} + (t^2 - n^2)s = 0 \quad (\text{A.62})$$

which doesn't have polynomial solutions. One set of solutions are the Bessel functions of the first type $J_n(t)$, which look like damped sinusoids. The first few of these are plotted in Figure A.1.

Although we won't show this, these Bessel functions can be calculated using the following recursions.

$$\begin{aligned} J_0(t) &= 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 4^2} - \frac{x^6}{2^2 4^2 6^2} + \dots \\ J_1(t) &= \frac{x}{2} - \frac{x^3}{2^2 4} + \frac{x^5}{2^2 4^2 6} - \frac{x^7}{2^2 4^2 6^2 8} + \dots \\ J_{n+1}(t) &= \frac{2n}{t} J_n(t) - J_{n-1}(t) \end{aligned}$$

In addition to equation (A.53) the trigonometric functions obey an additional differential equation, namely

$$\left(\frac{ds(t)}{dt} \right)^2 = 1 - s^2(t) \quad (\text{A.63})$$

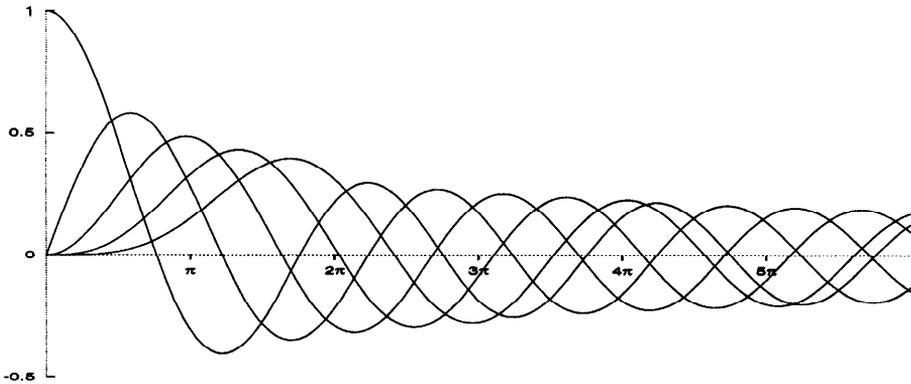


Figure A.1: Bessel functions of the first type $J_0(t), J_1(t), J_2(t), J_3(t),$ and $J_4(t)$

an equation that emphasizes the oscillatory behavior. Imposing the conditions that $s(0) = 0$ and $\frac{ds}{dt}|_0 = 1$ selects the sine while reversing the conditions selects the cosine. From this equation it is easy to deduce that sine and cosine are periodic with period

$$T = 2 \int_{-1}^1 \frac{ds}{\sqrt{1-s^2}} = 4 \int_0^1 \frac{ds}{\sqrt{1-s^2}} = 2\pi \quad (\text{A.64})$$

and that they are constrained to output values $-1 \leq s(t) \leq +1$.

We can generalize equation (A.63) to

$$\left(\frac{ds(t)}{dt} \right)^2 = (1 - s^2(t)) (1 - k^2 s^2(t)) \quad 0 \leq k \leq 1 \quad (\text{A.65})$$

where $k = 0$ reduces to the previous equation. The solutions to this equation are the elliptical functions $\text{sn}_k(t)$ and $\text{cn}_k(t)$ defined in Appendix A.8, and using logic similar to that preceding equation (A.64) we can prove that their period is $4K_k$.

EXERCISES

- A.10.1 What differential equation do the hyperbolic functions obey?
- A.10.2 Give an explicit formula for the k zeros and the $k - 1$ extrema of T_k .
- A.10.3 Write a program to expand functions in Chebyshev polynomials. Test it by approximating various polynomials. Expand $\cos(x)$ and $\tan(x)$ in Chebyshev polynomials. How many terms do you need for 1% accuracy?

A.10.4 Show that all the zeros of the Chebyshev polynomials are in the interval $-1 \leq t \leq +1$.

A.10.5 How can differential equations be solved numerically?

A.11 The Dirac Delta

The delta function is not a function, but a useful generalization of the concept of a function. It is defined by two requirements

$$\begin{aligned} \delta(t) &= 0 & \text{for all } t \neq 0 & & \text{(A.66)} \\ \int_{-\infty}^{\infty} \delta(t) dt &= 1 \end{aligned}$$

which obviously can't be fulfilled by any normal function.

From this definition it is obvious that the integral of Dirac's delta is Heaviside's step function

$$\Theta(t) = \int_{-\infty}^t \delta(\tau) d\tau \quad \text{(A.67)}$$

and conversely that the derivative of the unit step (which we would normally say doesn't have a derivative at zero) is the impulse.

$$\delta(t) = \left. \frac{d}{d\tau} \Theta(\tau) \right|_t \quad \text{(A.68)}$$

There are many useful integral relationships involving the delta. It can be used to select the value of a signal at a particular time,

$$\int_{-\infty}^{\infty} s(t) \delta(t - \tau) dt = s(\tau) \quad \text{(A.69)}$$

its 'derivative' selects the derivative of a signal,

$$\int_{-\infty}^{\infty} s(t) \frac{d}{dt} \delta(t - \tau) dt = \frac{d}{dt} s(\tau) \quad \text{(A.70)}$$

and you don't get anything if you don't catch the singularity.

$$\int_a^b \delta(t - \tau) = \begin{cases} 1 & a < \tau < b \\ 0 & \text{else} \end{cases} \quad \text{(A.71)}$$

As long as you only use them under integrals the following are true.

$$\begin{aligned}\delta(-t) &= \delta(t) \\ \delta(at) &= \frac{1}{|a|}\delta(t)\end{aligned}\tag{A.72}$$

The delta function has various ‘representations’ (i.e., disguises that it uses and that you need to recognize). The most important is the Fourier integral representation.

$$\delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} d\omega\tag{A.73}$$

EXERCISES

A.11.1 Prove

$$\delta(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \cos(\omega t) d\omega$$

A.11.2 Prove that $x\delta(x) = 0$.

A.11.3 Prove

$$\delta(h(t)) = \sum_n \frac{1}{|\dot{h}(t_n)|} \delta(t - t_n)$$

where the sum is over all times when $h(t_n) = 0$ but the derivative $\dot{h}(t_n) \neq 0$.

A.11.4 Can you think of a use for the n^{th} derivative of the delta?

A.11.5 Give an integral representation of Heaviside’s step function.

A.12 Approximation by Polynomials

We are often interested in approximating an arbitrary but smooth continuous function $f(x)$ by some other function $a(x)$ in some interval $a \leq x \leq b$. The error of this approximation at each point in the interval

$$\epsilon(x) = f(x) - a(x)$$

defines the faithfulness of the approximation at a particular x . The variable x will usually be either the time t or the frequency ω .

The approximating function $a(x)$ is always chosen from some family of functions. In this section we will concentrate on the polynomials

$$a(x) = \sum_{m=0}^M a_m x^m \quad (\text{A.74})$$

but weighted sums of sinusoids and many other sets of functions can be treated similarly. The important point is that the particular function in the family is specified by some parameter or parameters, and that these parameters are themselves continuous. For polynomials of degree up to M there are $M + 1$ parameters, namely the coefficients a_m for $m = 0 \dots M$. These parameters are continuous, and even a small change of a single coefficient results in a different polynomial. Our job is to find the polynomial in the family that best approximates the given function $f(x)$.

Comparison of the overall quality of two different approximations necessitates quantifying the accuracy of an approximation in the entire interval by a single value. Two reasonable candidates come to mind. The mean square error

$$\epsilon^2 = \frac{1}{b-a} \int_a^b \epsilon^2(x) dx = \frac{1}{b-a} \int_a^b (f(x) - a(x))^2 dx \quad (\text{A.75})$$

and the maximum error.

$$\epsilon_{max} = \max_{a \leq x \leq b} |\epsilon(x)| = \max_{a \leq x \leq b} |f(x) - a(x)| \quad (\text{A.76})$$

Although approximations with either low mean squared error or low maximum error are in some sense 'good' approximations, these criteria are fundamentally different. Requiring small maximum error ϵ^2 guarantees that the approximation error will be uniformly small; while with small mean squared error, the pointwise approximation error may be small over most of the interval but large at specific ω .

We can thus define two different types of approximation problems. The first is to find that function $a(x)$ in a family according to the Least Mean Squared (LMS) error criterion. The second is to find the function that has minimal maximum error, called the *minimax* criterion. Since the function $a(x)$ is specified by its parameters in the family, both the LMS and minimax problems reduce to finding the parameters that obey the respective criterion. In this section we will limit ourselves to the family of polynomials of degree M , as in equation (A.74); hence the question is simply how to find the best $M + 1$ coefficients a_m (i.e., those coefficients that minimize either the LMS or maximal error).

These two approximation types are not the only ones, but they are the important ones when the problem is to minimize the error *in an interval*. Were we to want the best polynomial approximation in the vicinity of a single point x_0 , the best polynomial approximation would be the truncated Taylor expansion. However, as we distance ourselves from x_0 the error increases, and so the Taylor expansion is not an appropriate approximation over an entire interval. Were we to want the best approximation at some finite number of points x_k for $k = 1 \dots K$, the best approximation would be Lagrange's collocating polynomial of degree $K - 1$.

$$\begin{aligned}
 a(x) = & \frac{(x-x_2)(x-x_3)\cdots(x-x_K)}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_K)} f(x_1) \\
 & + \frac{(x-x_1)(x-x_3)\cdots(x-x_K)}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_K)} f(x_2) \\
 & + \quad \dots \\
 & + \frac{(x-x_1)(x-x_2)\cdots(x-x_{K-1})}{(x_K-x_1)(x_K-x_2)\cdots(x_K-x_{K-1})} f(x_K)
 \end{aligned} \tag{A.77}$$

Although the collocating polynomial has zero error at the K points, we have no control over what happens in between these points, and in general it will oscillate wildly.

We will first consider the LMS approximation, where we are looking for the coefficients of (A.74) that minimize the mean squared error (A.75). Substituting, we can explicitly write the squared error (the normalization is irrelevant to the minimization) in terms of the coefficients a_m to be found.

$$\int_a^b \left(f(x) - \sum_{m=0}^M a_m x^m \right)^2 dx$$

Differentiating and setting equal to zero we obtain the 'normal equations' that can be solved for the coefficients.

$$\begin{aligned}
 \sum I_{m,l} x_l &= F_m \\
 I_{m,l} &\equiv \int_a^b x^{l+m} dx \\
 F_m &\equiv \int_a^b f(x) x^m dx
 \end{aligned} \tag{A.78}$$

These equations can be solved by any of the usual methods for solving equations with symmetric matrices, but unfortunately often turn out to

be very sensitive numerically; hence the SVD approach is recommended for large M . An alternative approach based on orthogonal polynomials is more stable numerically, and is based on giving a more sophisticated linear algebra interpretation to the normal equations. Think of the powers $1, x, x^2, x^3, \dots, x^M$ as a basis for a space of functions. Each element of the vector on the right F_m is the projection of $f(x)$ onto one of the basis functions, while the matrix $I_{m,l}$ contains the projections of the various nonorthogonal basis functions on each other. This is precisely the technique we use when finding Fourier components; we project the function onto the sinusoids, but don't need to solve equations because the $I_{m,l}$ matrix is diagonal due to the orthogonality of the sinusoids.

So what we need here is an orthogonal basis to replace the basis of powers. The Legendre polynomials of equation (A.55) are such a basis, and hence one can find their coefficients without solving equations, and then convert these to the coefficients of the powers by a linear transformation.

In DSP the squared error of equation (A.75) is replaced by a sum over a discrete time or frequency

$$\epsilon^2 = \frac{1}{N} \sum_{n=1}^N \epsilon^2(x_n) = \frac{1}{N} \sum_{n=1}^N \left(f(x_n) - a(x_n) \right)^2$$

and the normal equations are the same, but F_m and $I_{m,l}$ contain sums rather than integrals. The Legendre polynomials are not orthogonal when the inner product is a sum, but there are other polynomials, called the Szego polynomials, that are.

The finding of the minimax polynomial is in general a more difficult problem, since there is no simple error expression to be differentiated. Chebyshev proved a useful theorem, called the 'alternation theorem', that makes minimax polynomial approximation tractable. To understand the alternation theorem, consider first the following simpler result. If a polynomial $a(x)$ is the minimax approximation to a function $f(x)$ in the interval $[a \dots b]$, and the minimax error is ϵ_{max} , then there are two points x_1 and x_2 in the interval such that $\epsilon(x_1) = -\epsilon_{max}$ and $\epsilon(x_2) = +\epsilon_{max}$. Why is this true? By the definition of ϵ_{max} , the pointwise error is constrained to lie between two parallel lines $-\epsilon_{max} \leq \epsilon(x) \leq +\epsilon_{max}$, and it must touch at least one of these lines. In addition, were it not to touch the other we would be able to shift the supposed minimax polynomial by a constant, thereby decreasing ϵ_{max} .

What Chebyshev proved is that the pointwise error of the true minimax polynomial touches the bounding lines many more times, alternating between the lower bound and the upper one. Once again, were it not to do so

there would be a way of reducing the maximum error without increasing the degree. Therefore the minimax error is 'equiripple', i.e., oscillates between lower and upper bounds touching first one and then the other.

Theorem: The Alternation Theorem

A necessary and sufficient condition for the polynomial $a(x)$ of degree M to be the minimax approximation in an interval is for the error function to have *at least* $M + 2$ extrema in the interval, and for the error to alternate between $-\epsilon_{max}$ and $+\epsilon_{max}$ at these extrema. ■

The equiripple property led Chebyshev to seek a family of polynomials that oscillate between ± 1 in the interval $-1 \leq x \leq +1$ (it is easy to modify these to arbitrary bounds and intervals). He discovered, of course, the Chebyshev polynomials of equation (A.57). These polynomials are optimal for the purpose since they oscillate precisely as required and furthermore 'use up' all their oscillatory behavior in the interval of interest (once outside they diverge to infinity as fast as a polynomial can). In particular, the error of the M^{th} degree minimax approximation to x^{M+1} in the interval $[-1 \dots +1]$ is precisely $2^{-M}T_{M+1}(x)$. The search for minimax polynomials (combinations of powers of x) is thus more conveniently replaced by the search for combinations of Chebyshev polynomials

$$a(x) = \sum_{m=0}^M b_m T_m(x) = \sum_{m=0}^M b_m \cos(m \cos^{-1} x)$$

or using a change of variables,

$$c(x) = \sum_{m=0}^M b_m \cos(mx) = \sum_{m=0}^M c_m \cos^m(x) \quad (\text{A.79})$$

where we have implicitly used the general multiple angle formula of equation (A.31). In particular, the alternation theorem still holds in terms of this new representation in terms of trigonometric polynomials.

The Russian mathematician Evgeny Yakovlevich Remez enhanced the practice of approximation by trigonometric polynomials, and rational functions of cosines. His 'exchange algorithm' is a practical method for finding the coefficients in equation (A.79), based on the alternation theorem. The idea is simple. We know that the error has $M + 2$ extrema and that the error is maximal there. Were we to know the precise positions of the extrema ξ_i , the following $M + 2$ equations would hold

$$\epsilon(\xi_i) = f(\xi_i) - \sum_{m=0}^M b_m \cos(m\xi_i) = (-1)^i \epsilon_0 \quad \text{for } i = 1 \dots M+2$$

and could be solved for the $M + 1$ coefficients b_m and the maximal error ϵ_0 . Don't be confused; since the ξ_i are assumed to be known, $F_i = f(\xi_i)$ and $C_{i,m} = \cos(m\xi_i)$ are constants, and the equations to be solved are linear.

$$\sum_{m=0}^M C_{i,m} b_m - (-1)^i \epsilon_0 = F_i$$

Unfortunately we do not really know where the extrema are, so we make some initial guess and solve. This results in a polynomial approximation to $f(x)$, but usually not a minimax one. The problem is that we forced the error to be $\pm\epsilon_0$ at the specified points, but these points were arbitrarily chosen and the error may be larger than ϵ_0 at other points in the interval. To fix this we pick the $M + 2$ extrema with the highest error and exchange our original extrema with these new ξ_i and solve for b_m and ϵ_0 once again. We continue to iterate until the actual maximal error is smaller than the desired error. McClellan, Parks, and Rabiner found a faster way to perform the iterations by using the Lagrange's collocating polynomial (equation (A.77)) instead of directly solving the linear equations.

EXERCISES

- A.12.1 Approximate the function $f(x) = e^x$ on the interval $[-1 \leq x \leq +1]$ by a polynomial of degree 4 using a Taylor expansion at $x = 0$, collocating polynomials that touch the function at $\pm \frac{\pi}{n}$, LMS, and minimax polynomials. Determine the maximum error for all the above methods.
- A.12.2 Give explicit formulas for the slope and zero crossing of the line that LMS approximates N empirical data points. What are the expected errors for these parameters?
- A.12.3 How can we match $y = Ae^{\alpha x}$ to empirical data using techniques of this section? Does this technique truly find the minimum error?
- A.12.4 Show that the normal equations for polynomial approximation become ill-conditioned for high polynomial degree and large number of data points.
- A.12.5 Find a set of digital orthogonal polynomials, $p^{[m]}(t)$, such that for $m_1 \neq m_2$: $\sum_{n=0}^N p^{[m_1]}(t_n) p^{[m_2]}(t_n) = 0$. How can these polynomials be used for LMS polynomial approximation?

A.13 Probability Theory

We will not need much probability theory in this book, although we assume that the reader has had some exposure to the subject. The probability of some nondeterministic event occurring is defined by considering the event to be a particular realization of an ensemble of similar events. The fraction of times the event occurs in the ensemble is the probability. For example, if we throw a cubical die, the ensemble consists of six types of events, namely throwing a 1, or a 2, or a 3, or a 4, or a 5, or a 6. For a fair die these events are equally probable and the probability of throwing a 1 is thus $P(1) = \frac{1}{6}$. If we were informed that the die came up odd, but not the exact number, the ensemble shrinks to three possibilities, and the probability of a 1 given that it is odd is $P(1|\text{odd}) = \frac{1}{3}$.

A ‘random variable’ is a mapping from the set of all possible outcomes of some experiment into the real numbers. The idea is to change events into numbers in order to be able to treat them numerically. For example, the experiment might be observing the output of a black box and the random variable the value of the signal observed. The random variable will have some ‘distribution’, representing the probability it will take on a given value. The ‘law of large numbers’ states (roughly) that the distribution of the sum of a large number of independent random variables will always be approximately Gaussian. For this reason random variables with Gaussian distribution are called ‘normal’.

When the possible outcomes are from a continuous set the probability of the random number being any particular real number is usually zero; so we are more interested in the probability that the outcome is approximately x . We thus define the ‘probability density’ $p(x)$ such that the probability of the random variable being between $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$ is $p(x) dx$. Since the probability of any x is unity, probability densities are always normalized.

$$\int p(x) dx = 1$$

For example, if the event is the marking of a test, the mark’s probability density will be approximately Gaussian, with its peak at the average mark.

The most important single piece of information about any random variable is its ‘expected value’

$$\langle x \rangle = \mu_1 = \sum_n x_n p(x_n) \quad \Bigg| \quad \langle x \rangle = \mu_1 = \int x p(x) dx \quad (\text{A.80})$$

the left form being used for discrete variables and the right form for continuous ones. The terms ‘expectation’, ‘average’, and ‘mean’ are also commonly applied to this same quantity. For the simple case of N discrete equally-probable values the expectation is precisely the arithmetic mean; for N nonequally-probable values it is the weighted average. Even for the most general case, if you have to make a single guess as to the value a random variable you should probably pick its expectation. Such a guess will be unbiased—half the time it will be too low and half the time too high.

Although the average is definitely important information, it doesn’t tell the full story; in particular we would like to know how ‘wide’ the distribution is. You may propose to compute the average deviation from the average value,

$$\langle x - \langle x \rangle \rangle = 0$$

but as we have just mentioned this is always zero. A better proposal is the ‘variance’

$$\text{Var} = \langle (x - \langle x \rangle)^2 \rangle \quad (\text{A.81})$$

which is always positive. If the expectation is zero then the variance is simply $\langle x^2 \rangle$, but even in general it is related to this quantity.

$$\text{Var} = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - 2 \langle x \rangle \langle x \rangle + \langle x \rangle^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Since the units of variance are not those of length it is often more convenient to define the ‘standard deviation’.

$$\sigma = \sqrt{\text{Var}}$$

The distribution of a normal (Gaussian) random variable is completely determined given its expectation and variance; for random variables with other distributions we need further information. From the distribution of a random variable x we can determine its ‘moments’,

$$\mu_k \equiv \int x^k p(x) dx \quad \Bigg| \quad \mu_k \equiv \sum_n x_n^k p(x_n) \quad (\text{A.82})$$

and conversely the distribution is uniquely determined from the set of all moments. The zeroth moment is unity by definition (normalization) and the first moment is the expectation. From the second moment and higher we will assume that the mean is zero; if it isn’t for your distribution simply define a new variable $x - \langle x \rangle$. The second moment is precisely the variance.

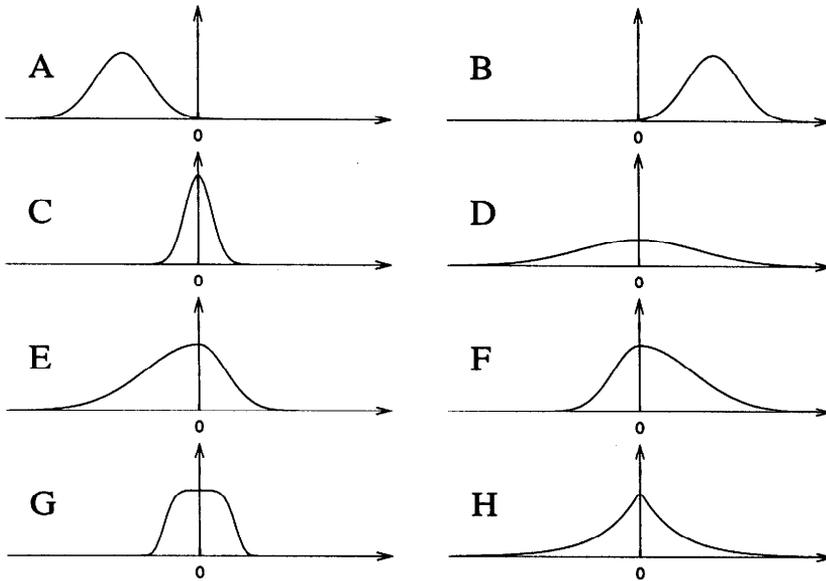


Figure A.2: Moments of probability distributions. In (A) is a distribution with negative first moment (expectation, mean, average) while in (B) is a distribution with positive first moment. In (C) is a distribution with zero mean but smaller second moment (variance) while in (D) is a distribution with larger variance. In (E) is a distribution with zero mean and negative third moment (skew) while in (F) is a distribution with positive skew. In (G) is a distribution with negative kurtosis while in (H) is a distribution with positive kurtosis.

The third moment divided by the standard deviation raised to the third power is called the 'skew';

$$\text{skew} \equiv \frac{\mu_3}{\sigma^3}$$

it measures deviation from symmetry around zero. Normal random variables have zero skew.

For the Gaussian distribution the fourth moment divided by the standard deviation raised to the third power equals three; so to measure deviation from normality we define the 'kurtosis' as follows.

$$\text{kurtosis} \equiv \frac{\mu_4}{\sigma^4} - 3$$

Distributions with positive kurtosis have narrower main lobes but higher tails than the Gaussian. The meaning of the first few moments is depicted graphically in Figure A.2.

Frequently real-world objects have more than one characteristic; for example, people have both height h and weight w . The obvious extension of

the above concepts is to define the ‘joint probability’ $p(h, w) dh dw$ meaning the probability of the person having height in the vicinity of h and simultaneously weight about w . For such joint probability distributions we have the so-called ‘marginals’,

$$p(h) = \int p(h, w) dw \qquad p(w) = \int p(h, w) dh$$

where the integrations are over the entire range of possible heights and weights, $p(h)dh$ is the percentage of people with height between h and $h + dh$ regardless of weight, and $p(w)dw$ is the percentage of people with weight between w and $w + dw$ regardless of height. The integration over both height and weight must give one.

Two random variables are said to be ‘statistically independent’ if knowledge of the value of one does not affect the knowledge of the other. For example, we can usually assume that consecutive throws of a fair coin are independent, and knowing what happened on the first 100 throws does not help us to predict what will happen on the next. Two random variables are said to be uncorrelated if their crosscorrelation (defined as the expectation of their product) is zero. Statistically independent random variables are necessarily uncorrelated, but the converse need not be true.

EXERCISES

- A.13.1 Define $p(B|A)$ to be the probability of event B occurring given that event A occurred. Prove that the probability of both events A and B occurring is $p(A \wedge B) = p(A)p(B|A)$; and if A and B are independent events that $p(A \wedge B) = p(A)p(B)$.
- A.13.2 Prove that the probability of either of two events occurring is $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$, and that if A and B are mutually exclusive events that $p(A \vee B) = p(A) + p(B)$.
- A.13.3 Prove Bayes’ theorem $p(B|A) = p(A|B)p(B)/p(A)$ and explain how this enables defining probabilities that can not be defined by our original definition.
- A.13.4 Let the probability of an experiment succeeding be p . Show that the probability of exactly m successes out of n identical independent experiments is given by the binomial distribution.

$$p(m) = \binom{n}{m} p^m (1 - p)^{n-m}$$

Show that the binomial distribution approaches the normal distribution for large n . (Hint: Use Stirling’s approximation for the factorials in the binary coefficient).

A.14 Linear Algebra

Linear algebra is the study of vectors. ‘Vector’ actually means several radically different things that turn out, almost by accident, to be connected. If your background is science, the word ‘vector’ probably triggers the geometric meaning, while computer scientists always think of n -tuples of numbers. The technical mathematical meaning is more general than either of these, and allows such entities as the set of all analog signals, or of all digital signals, or of all periodic signals, to be vector spaces as well.

The abstract mathematical definition of ‘vector’ is an element of a vector space, a concept that we introduced in Appendix A.5. Compiling all the requirements set forth there, a vector space must obey all of the following rules.

Addition: For every two vectors \underline{x} and \underline{y} , there is a unique vector \underline{z} such that $\underline{z} = \underline{x} + \underline{y}$; this addition is commutative and associative,

Zero: There is a ‘zero vector’ $\underline{0}$, such that $\underline{x} + \underline{0} = \underline{x}$ for every vector \underline{x} ,

Inverse: Every vector \underline{x} has an inverse vector $-\underline{x}$ such that $\underline{x} + -\underline{x} = \underline{0}$,

Multiplication: For every vector \underline{x} and number a there is a vector \underline{ax} .

In addition some vector spaces have further properties.

Inner Product: For every two vectors \underline{x} and \underline{y} , there is a unique number a such that $a = \underline{x} \cdot \underline{y}$,

Norm: For every vector \underline{x} there is a unique nonnegative real number r such that $r = |\underline{x}|$; $r = 0$ if and only if $\underline{x} = \underline{0}$,

Metric: For every two vectors \underline{x} and \underline{y} , there is a unique nonnegative real number d such that $d = D(\underline{x}, \underline{y})$; $d = 0$ if and only if $\underline{x} = \underline{y}$.

From these basic definitions many interesting concepts and theorems can be derived. We can make general ‘linear combinations’ of vectors

$$\sum_{i=1}^N s_i \underline{V}_i = s_1 \underline{V}_1 + s_2 \underline{V}_2 + \dots + s_N \underline{V}_N \quad (\text{A.83})$$

which must return a vector in the space. The set of all vectors that can be so formed are called the ‘span’ of $\underline{V}_1, \underline{V}_2, \dots, \underline{V}_N$. The span is itself a subspace of the original space. It is not difficult to prove this directly from the axioms. For example, we must be able to create the zero vector, which can be done by

choosing $s_1 = s_2 = \dots = s_N = 0$. If this is the *only* way of creating the zero vector, then we say that the vectors $\underline{V}_1, \underline{V}_2, \dots, \underline{V}_N$ are 'linearly independent'.

If the vectors $\underline{V}_1, \underline{V}_2, \dots, \underline{V}_N$ are linearly independent and span the entire vector space, we say that they are a 'basis' for the space. Given a basis, any vector in the space may be created in a unique way; were there to be two representations

$$\begin{aligned}\underline{X} &= \sum_{i=1}^N r_i \underline{V}_i = r_1 \underline{V}_1 + r_2 \underline{V}_2 + \dots + r_N \underline{V}_N \\ \underline{X} &= \sum_{i=1}^N s_i \underline{V}_i = s_1 \underline{V}_1 + s_2 \underline{V}_2 + \dots + s_N \underline{V}_N\end{aligned}$$

then by subtracting the equations we would find

$$\underline{0} = \sum_{i=1}^N (r_i - s_i) \underline{V}_i = (r_1 - s_1) \underline{V}_1 + (r_2 - s_2) \underline{V}_2 + \dots + (r_N - s_N) \underline{V}_N$$

which by linear independence of the basis requires all the respective scalar coefficients to be equal.

There are many different bases (for example, in two dimensions we can take any two noncolinear vectors) but all have the same number N of vectors, which is called the 'dimension' of the space. The dimension may be finite (such as for two- or three-dimensional vectors), denumerably infinite (digital signals), or nondenumerably infinite (analog signals).

We defined two vectors to be orthogonal if their dot product is zero. A set of three or more vectors can also be orthogonal, the requirement being that every pair is orthogonal. If a set of unit-length vectors are orthogonal, we call them 'orthonormal'.

$$\underline{V}_i \cdot \underline{V}_j = \delta_{i,j} \tag{A.84}$$

It is not hard to show that any finite number of orthonormal vectors are linearly independent, and that if given a basis we can create from it an orthonormal basis.

Given a vector and a basis how do we find the expansion coefficients? By dotting the vector with *every* basis vector we obtain a set of equations, called 'normal equations', that can be solved for the coefficients.

$$\begin{aligned}\underline{X} \cdot \underline{V}_1 &= X_1 \underline{V}_1 \cdot \underline{V}_1 + X_2 \underline{V}_2 \cdot \underline{V}_1 + \dots + X_N \underline{V}_N \cdot \underline{V}_1 \\ \underline{X} \cdot \underline{V}_2 &= X_1 \underline{V}_1 \cdot \underline{V}_2 + X_2 \underline{V}_2 \cdot \underline{V}_2 + \dots + X_N \underline{V}_N \cdot \underline{V}_2 \\ &\vdots \\ \underline{X} \cdot \underline{V}_N &= X_1 \underline{V}_1 \cdot \underline{V}_N + X_2 \underline{V}_2 \cdot \underline{V}_N + \dots + X_N \underline{V}_N \cdot \underline{V}_N\end{aligned}$$

Now we see a useful characteristic of orthonormal bases; only for these can we find the i^{th} coefficient by dotting with \underline{V}_i alone.

$$\underline{X} \cdot \underline{V}_i = \sum_{j=1}^N X_j \underline{V}_j \cdot \underline{V}_i = \sum_{j=1}^N X_j \delta_{i,j} = X_i \quad (\text{A.85})$$

A.15 Matrices

A ‘matrix’ is a rectangularly shaped array of numbers, and thus specification of a particular ‘matrix element’ A_{ij} requires two indices, i specifying the ‘row’ and j specifying the ‘column’. In this book we symbolize matrices by $\underline{\underline{A}}$, the double underline alluding to the two-dimensionality of the array, just as the single underline indicated that vectors are one-dimensional arrays. When actually specifying a matrix we write it like this

$$\begin{pmatrix} 11 & 12 & 13 & 14 \\ 21 & 22 & 23 & 24 \\ 31 & 32 & 33 & 34 \\ 41 & 42 & 43 & 44 \end{pmatrix}$$

this being a 4-by-4 matrix with the numbers 11, 12, 13, 14 residing on the first row, 11, 21, 31, 41 being in the first column, and 11, 22, 33, 44 comprising the ‘diagonal’.

The ‘transpose’ $\underline{\underline{A}}^t$ of a matrix is obtained by interchanging the rows and columns $A_{ij}^t = A_{ji}$. If $\underline{\underline{A}}$ is N -by- M then $\underline{\underline{A}}^t$ will be M by N . For matrices with complex elements the corresponding concept is the ‘Hermitian transpose’ $\underline{\underline{A}}^H$, where $A_{ij}^H = A_{ji}^*$.

Actually $\underline{\underline{A}}$ vectors can be considered to be special cases of matrices with either a single row or a single column. A ‘row vector’ is thus a horizontal array

$$\left(11 \quad 12 \quad 13 \quad 14 \right)$$

and a ‘column vector’ a vertical array.

$$\begin{pmatrix} 11 \\ 21 \\ 31 \\ 41 \end{pmatrix}$$

'Square matrices' have the same number of rows as they have columns. If a square matrix is equal to its transpose (i.e., $a_{ij} = a_{ji}$), then we say that the matrix is 'symmetric'.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1N} \\ a_{12} & a_{22} & a_{23} & \dots & a_{2N} \\ a_{13} & a_{23} & a_{33} & \dots & a_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1N} & a_{2N} & a_{3N} & \dots & a_{NN} \end{pmatrix} \quad (\text{A.86})$$

If a complex-valued matrix obeys $a_{ij} = a_{ji}^*$, then we say that it is 'Hermitian'. The elements of a square matrix with constant difference between their indices are said to reside on the same diagonal, and the elements a_{ii} of a square matrix are called its 'main diagonal'. A matrix with no nonzero elements off the main diagonal is said to be 'diagonal'; a matrix with nonzero elements on or below (above) the main diagonal is called 'lower (upper) triangular'. If all the diagonals have all their elements equal,

$$\begin{pmatrix} a & b & c & \dots \\ x & a & b & c & \dots \\ y & x & a & b & c & \dots \\ z & y & x & a & b & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & a & b \\ \dots & \dots & \dots & \dots & x & a \end{pmatrix} \quad (\text{A.87})$$

the matrix is called 'Toeplitz'. A matrix can be both symmetric and Toeplitz.

Matrices can be multiplied by scalars (real or complex numbers) by multiplying every element in the array. Matrices of the same shape can be added by adding their corresponding elements $C_{ij} = A_{ij} + B_{ij}$, but the multiplication is somewhat less obvious.

$$C_{ik} = \sum_j A_{ij} B_{jk} \quad (\text{A.88})$$

Matrix multiplication is not generally commutative, and doesn't even have to be between similarly shaped matrices. The requirement is that the matrix on the right have the same number of rows as the matrix on the left has columns. If the left matrix is L -by- M and the right is M -by- N then the product matrix will be L -by- N . In particular the product of two N -by- N square matrices is itself N -by- N square. Also, the inner (dot) product of

two vectors is automatically obtained if we represent one of the vectors as a row vector and the other as a column vector.

$$\begin{pmatrix} u_1 & u_2 & \dots & u_N \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix} = u_1v_1 + u_2v_2 + \dots + u_Nv_N \quad (\text{A.89})$$

If we place the vectors in the opposite order we obtain the 'outer product', which is a N -by- N matrix.

$$\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \begin{pmatrix} v_1 & v_2 & \dots & v_N \end{pmatrix} = \begin{pmatrix} u_1v_1 & u_1v_2 & \dots & u_1v_N \\ u_2v_1 & u_2v_2 & \dots & u_2v_N \\ \vdots & \vdots & \ddots & \vdots \\ u_Nv_1 & u_Nv_2 & \dots & u_Nv_N \end{pmatrix} \quad (\text{A.90})$$

The N -by- M 'zero matrix' $\underline{\underline{0}}$ is the matrix with all elements equal to zero. It is obvious from the definitions that $\underline{\underline{0}} + \underline{\underline{A}} = \underline{\underline{A}} + \underline{\underline{0}} = \underline{\underline{A}}$. The set of all N -by- M matrices with real elements is a field over the reals with respect to matrix addition and multiplication using this zero element.

The N -by- N square 'identity matrix' $\underline{\underline{I}}$ is given by $I_{ij} = \delta_{i,j}$

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

and as its name implies $\underline{\underline{IA}} = \underline{\underline{A}}$ and $\underline{\underline{AI}} = \underline{\underline{A}}$ whenever the multiplication is legal.

A square matrix is called orthogonal if $\underline{\underline{AA^t}} = \underline{\underline{I}}$, i.e., if the rows (or columns) when viewed as vectors are orthonormal. For complex matrices, a matrix for which $\underline{\underline{AA^H}} = \underline{\underline{I}}$ is called 'unitary'.

One of the reasons that matrices are so important is that they perform transformations on vectors. For example, vectors in the two-dimensional plane are rotated by θ by multiplying them by a rotation matrix $\underline{\underline{R}}_\theta$.

$$\underline{\underline{R}}_\theta = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \quad (\text{A.91})$$

It is easy to see that rotation matrices are orthogonal, and hence $\underline{R}_\theta \underline{x}$ has the same length as \underline{x} . It is also not difficult to prove that $\underline{R}_{\alpha+\beta} = \underline{R}_\alpha \underline{R}_\beta$ (i.e., that rotations can be performed in steps).

If we perform an orthogonal transformation \underline{R} on a vector space, the particular representation of a vector \underline{x} changes to $\underline{x}' = \underline{R}\underline{x}$, but we can think of the abstract vector itself as being unchanged. For instance, rotation of the axes change the vector representation, but the vectors themselves have a deeper meaning. Similarly, a matrix \underline{M} that performed some operation on vectors is changed by such changes of axes. The matrix in the new axes that performs the same function is

$$\underline{M}' = \underline{R} \underline{M} \underline{R}^{-1} = \underline{R} \underline{M} \underline{R}^t \quad (\text{A.92})$$

as can be easily seen. If the original effect of the matrix was $\underline{y} = \underline{M}\underline{x}$ then in the new representation we have

$$\underline{y}' = \underline{M}' \underline{x}' = \underline{R} \underline{M} \underline{R}^{-1} \underline{R} \underline{x} = \underline{R} \underline{y}$$

as expected. Two matrices that are related by $\underline{B} = \underline{R} \underline{A} \underline{R}^t$ where \underline{R} is orthogonal, are said to be 'similar'.

There are four common tasks relating to matrices: inversion, diagonalization, Cholesky decomposition, and singular value decomposition (SVD). 'Inversion' of \underline{A} is the finding of a matrix \underline{A}^{-1} such that $\underline{A} \underline{A}^{-1} = \underline{I}$. This is closely related to the task of equation solving that is discussed in the next section. 'Diagonalization' of \underline{A} means finding a diagonal matrix \underline{D} that is similar to the original matrix.

$$\underline{A} = \underline{R} \underline{D} \underline{R}' \quad (\text{A.93})$$

Expressed another way, given a matrix \underline{A} if we have $\underline{A}\underline{x} = \lambda\underline{x}$ we say that λ is an 'eigenvalue' of \underline{A} and \underline{x} an 'eigenvector'. Placing all the eigenvalues on the main diagonal of a diagonal matrix results in the diagonal matrix to which \underline{A} is similar. The orthogonal matrix can be constructed from the eigenvectors. The Cholesky (also called LDU) decomposition of a square matrix \underline{A} is a representation

$$\underline{A} = \underline{L} \underline{D} \underline{U} \quad (\text{A.94})$$

where \underline{L} (\underline{U}) is lower (upper) diagonal with ones on the main diagonal, and \underline{D} is diagonal. The singular value decomposition (SVD) of a (not necessarily

square) $\underline{\underline{A}}$ is a representation

$$\underline{\underline{A}} = \underline{\underline{U}} \underline{\underline{D}} \underline{\underline{V}} \quad (\text{A.95})$$

where $\underline{\underline{U}}$ and $\underline{\underline{V}}$ are orthogonal (by column and by row respectively), and $\underline{\underline{D}}$ is diagonal with nonnegative elements.

There are many relationships between the above tasks. For example, given either the diagonal, Cholesky, or SVD representations, it is simple to invert the matrix by finding the reciprocals of the diagonal elements. Indeed the Cholesky decomposition is the fastest, and the SVD is the numerically safest, method for inverting a general square matrix. Numeric linear algebra has a rich literature to which the reader is referred for further detail.

EXERCISES

A.15.1 Show that $\underline{\underline{(AB)}}^{-1} = \underline{\underline{B}}^{-1} \underline{\underline{A}}^{-1}$.

A.15.2 The 2-by-2 Pauli spin matrices are defined as follows.

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Show that these matrices are Hermitian and unitary. Find σ_i^2 and $\sigma_i \sigma_j$.

A.15.3 The commutator, defined as $\underline{\underline{[A, B]}} = \underline{\underline{AB}} - \underline{\underline{BA}}$, can be nonzero since matrix multiplication needn't be commutative. Find the commutators for the Pauli matrices. Define the anticommutator as the above but with a plus sign. Show that the Pauli matrices anticommute.

A.15.4 Find the Crout (LU) and Cholesky (LDU) decompositions of

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 8 & 8 \\ 3 & 8 & 26 \end{pmatrix}$$

by setting it equal to

$$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix} \begin{pmatrix} a & b & d \\ 0 & c & e \\ 0 & 0 & f \end{pmatrix}$$

and to

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix} \begin{pmatrix} d & 0 & 0 \\ 0 & e & 0 \\ 0 & 0 & f \end{pmatrix} \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}$$

multiplying out and solving the equations. How many operations are needed? Why is the Cholesky method better? Now solve the equations with right-hand side (2, 8, 20).

A.16 Solution of Linear Algebraic Equations

A common problem in algebra is the solution of sets of linear equations

$$\underline{\underline{A}}x = \underline{b} \quad (\text{A.96})$$

where $\underline{\underline{A}}$ is a known $N * N$ matrix, \underline{b} is a known N -dimensional vector, and \underline{x} is the N -dimensional vector we want to find. Writing this out in full,

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2N} \\ A_{31} & A_{32} & A_{33} & \dots & A_{3N} \\ & & \vdots & & \\ A_{N1} & A_{N2} & A_{N3} & \dots & A_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_N \end{pmatrix} \quad (\text{A.97})$$

which means

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 \dots A_{1N}x_N &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 \dots A_{2N}x_N &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 \dots A_{3N}x_N &= b_3 \\ &\vdots \\ A_{N1}x_1 + A_{N2}x_2 + A_{N3}x_3 \dots A_{NN}x_N &= b_N \end{aligned}$$

and we see that this is actually N equations in N variables.

If we know how to invert the matrix $\underline{\underline{A}}$ the solution to the equations is immediate: $\underline{x} = \underline{\underline{A}}^{-1}\underline{b}$. This method of equation solving is especially effective if we have to solve many sets of equations with the same matrix but different right-hand sides. However, if we need to solve only a single instance it will usually be more efficient to directly solve the equations without inverting the matrix.

If $\underline{\underline{A}}$ happens to be lower (or upper) triangular then equation A.97 has the special form

$$\begin{pmatrix} A_{11} & 0 & 0 & \dots & 0 \\ A_{21} & A_{22} & 0 & \dots & 0 \\ A_{31} & A_{32} & A_{33} & \dots & 0 \\ & & \vdots & & \\ A_{N1} & A_{N2} & A_{N3} & \dots & A_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_N \end{pmatrix} \quad (\text{A.98})$$

and the solution to these equations is simple to find.

The first equation is

$$A_{11}x_1 = b_1$$

which is immediately solvable.

$$x_1 = \frac{b_1}{A_{11}}$$

With x_1 known we can solve the second equation as well.

$$A_{21}x_1 + A_{22}x_2 = b_2 \quad \implies \quad x_2 = \frac{b_2 - b_1 \frac{A_{21}}{A_{11}}}{a_{22}}$$

We can continue with this process, known as ‘back-substitution’, until all the unknowns have been found.

Back-substitution is only directly applicable to equations containing upper or lower triangular matrices, but we will now show how to transform more general sets of linear equations into just that form. First note that adding the multiple of one equation to another equation (i.e., adding the multiple of one row of $\underline{\underline{A}}$ to another row *and* the corresponding elements of $\underline{\underline{b}}$) does not change the solution vector $\underline{\underline{x}}$. Even more obviously, interchanging the order of two equations (i.e., interchanging two rows of $\underline{\underline{A}}$ *and* the corresponding elements of $\underline{\underline{b}}$) does not change the solution. Using just these two tricks we can magically transform arbitrary sets of equations (A.97) into the triangular form of equation (A.98).

The basic strategy was invented by Gauss and therefore called ‘Gaussian elimination’ and it can be extended to a method for finding the inverse of a matrix. However, if we really need to invert a matrix, there may be better methods. A matrix that has some special form may have an efficient inversion algorithm. For example, Toeplitz matrices can be inverted in $O(N^2)$ time by the Levinson-Durbin recursion discussed in Section 9.10. In addition, if numerical accuracy problems arise when using one of the standard algorithms, there are iterative algorithms to improve solutions.

Sometimes we know the inverse of matrix $\underline{\underline{A}}$, and need the inverse of another related matrix. If we are interested in the inverse of the matrix $\underline{\underline{A}} + \underline{\underline{B}}$, the following lemma is of use

$$\left(\underline{\underline{A}} + \underline{\underline{B}}\right)^{-1} = \underline{\underline{A}}^{-1} - \underline{\underline{A}}^{-1} \left(\underline{\underline{A}}^{-1} + \underline{\underline{B}}^{-1}\right)^{-1} \underline{\underline{A}}^{-1} \quad (\text{A.99})$$

and a somewhat more general form is often called the ‘matrix inversion lemma’.

$$\left(\underline{\underline{A}} + \underline{\underline{B}}\underline{\underline{C}}\underline{\underline{D}}\right)^{-1} = \underline{\underline{A}}^{-1} - \underline{\underline{A}}^{-1} \underline{\underline{B}} \left(\underline{\underline{D}}\underline{\underline{A}}^{-1}\underline{\underline{B}} + \underline{\underline{C}}^{-1}\right)^{-1} \underline{\underline{D}}\underline{\underline{A}}^{-1} \quad (\text{A.100})$$

Let's prove this last lemma by multiplying the supposed inverse by the matrix,

$$\begin{aligned}
 (\underline{\underline{A}} + \underline{\underline{BCD}})^{-1}(\underline{\underline{A}} + \underline{\underline{BCD}}) &= \\
 \left(\underline{\underline{A}}^{-1} - \underline{\underline{A}}^{-1}\underline{\underline{B}}(\underline{\underline{DA}}^{-1}\underline{\underline{B}} + \underline{\underline{C}}^{-1})^{-1}\underline{\underline{DA}}^{-1} \right) (\underline{\underline{A}} + \underline{\underline{BCD}}) &= \\
 \underline{\underline{I}} + \underline{\underline{A}}^{-1}\underline{\underline{BCD}} - \underline{\underline{A}}^{-1}\underline{\underline{B}}(\underline{\underline{DA}}^{-1}\underline{\underline{B}} + \underline{\underline{C}}^{-1})^{-1}\underline{\underline{D}}(\underline{\underline{I}} + \underline{\underline{A}}^{-1}\underline{\underline{BCD}}) &= \\
 \underline{\underline{I}} + \underline{\underline{A}}^{-1}\underline{\underline{B}}(\underline{\underline{C}} - \underline{\underline{X}})\underline{\underline{D}} &
 \end{aligned}$$

where

$$\begin{aligned}
 \underline{\underline{X}} &= (\underline{\underline{DA}}^{-1}\underline{\underline{B}} + \underline{\underline{C}}^{-1})^{-1}(\underline{\underline{I}} + \underline{\underline{DA}}^{-1}\underline{\underline{BC}}) \\
 &= (\underline{\underline{DA}}^{-1}\underline{\underline{B}} + \underline{\underline{C}}^{-1})^{-1}(\underline{\underline{C}}^{-1} + \underline{\underline{DA}}^{-1}\underline{\underline{B}})\underline{\underline{C}} = \underline{\underline{C}}
 \end{aligned}$$

which completes the proof.

EXERCISES

- A.16.1 Assume that $\underline{\underline{B}}$ is an approximation to $\underline{\underline{A}}^{-1}$, with error $\underline{\underline{R}} = \underline{\underline{I}} - \underline{\underline{BA}}$. Show that $\underline{\underline{A}}^{-1} = (\underline{\underline{I}} + \underline{\underline{R}} + \underline{\underline{R}}^2 + \underline{\underline{R}}^3 + \dots)\underline{\underline{B}}$ and that this can be used to iteratively improve the inverse.
- A.16.2 You know that x and y obey the equation $x + 3y = 8$ and determine numerically that they also obey $2x + 6.00001y = 8.00001$. What are x and y ? Suppose that the numerically determined equation is $2x + 5.99999y = 8.00002$. What are x and y now? Explain the discrepancy.