Notes and Problems for Applied Mathematics I

Charles L. Byrne

December 7, 2005

Contents

Chapter 1

The Fourier Transform

We begin with exercises that treat basic properties of the Fourier transform and then introduce several examples of Fourier-transform pairs. The (possibly complex-valued) function $f(x)$ of the real variable x has for its Fourier transform (FT) the (possibly complex-valued) function $F(\omega)$ of the real variable ω given by

$$F(\omega) = \int f(x)e^{ix\omega}dx. \tag{1.1}$$

From $F(\omega)$ we can find $f(x)$ from the Fourier inversion formula

$$f(x) = \frac{1}{2\pi}\int F(\omega)e^{-ix\omega}d\omega. \tag{1.2}$$

1.1 Basic Properties

Exercise 1.1 *Let $F(\omega)$ be the FT of the function $f(x)$. Use the definitions of the FT and IFT given in Equations (1.1) and (1.2) to establish the following basic properties of the Fourier transform operation:*

Symmetry: The FT of the function $F(x)$ is $2\pi f(-\omega)$.

Conjugation: The FT of $\overline{f(x)}$ is $\overline{F(-\omega)}$.

Scaling: The FT of $f(ax)$ is $\frac{1}{|a|}F(\frac{\omega}{a})$ for any nonzero constant a.

Shifting: The FT of $f(x-a)$ is $e^{ia\omega}F(\omega)$.

Modulation: The FT of $f(x)\cos(\omega_0 x)$ is $\frac{1}{2}[F(\omega+\omega_0)+F(\omega-\omega_0)]$.

Differentiation: The FT of the nth derivative, $f^{(n)}(x)$ is $(-i\omega)^n F(\omega)$. The IFT of $F^{(n)}(\omega)$ is $(ix)^n f(x)$.

Convolution in x: Let f, F, g, G and h, H be FT pairs, with

$$h(x) = \int f(y)g(x-y)dy,$$

so that $h(x) = (f * g)(x)$ is the convolution of $f(x)$ and $g(x)$. Then $H(\omega) = F(\omega)G(\omega)$. For example, if we take $g(x) = \overline{f(-x)}$, then

$$h(x) = \int f(x+y)\overline{f(y)}dy = \int f(y)\overline{f(y-x)}dy = r_f(x)$$

is the *autocorrelation function* associated with $f(x)$ and

$$H(\omega) = |F(\omega)|^2 = R_f(\omega) \geq 0$$

is the *power spectrum* of $f(x)$.

Convolution in ω: Let f, F, g, G and h, H be FT pairs, with $h(x) = f(x)g(x)$. Then $H(\omega) = \frac{1}{2\pi}(F * G)(\omega)$.

1.2 Examples

Exercise 1.2 *Show that the Fourier transform of $f(x) = e^{-\alpha^2 x^2}$ is $F(\omega) = \frac{\sqrt{\pi}}{\alpha} e^{-(\frac{\omega}{2\alpha})^2}$. Hint: Calculate the derivative $F'(\omega)$ by differentiating under the integral sign in the definition of F and integrating by parts. Then solve the resulting differential equation.*

Let $u(x)$ be the *Heaviside function* that is $+1$ if $x \geq 0$ and 0 otherwise. Let $\chi_X(x)$ be the *characteristic function* of the interval $[-X, X]$ that is $+1$ for x in $[-X, X]$ and 0 otherwise. Let $\text{sgn}(x)$ be the *sign function* that is $+1$ if $x > 0$, -1 if $x < 0$ and zero for $x = 0$.

Exercise 1.3 *Show that the FT of the function $f(x) = u(x)e^{-ax}$ is $F(\omega) = \frac{1}{a - i\omega}$, for every positive constant a.*

Exercise 1.4 *Show that the FT of $f(x) = \chi_X(x)$ is $F(\omega) = 2\frac{\sin(X\omega)}{\omega}$.*

Exercise 1.5 *Show that the IFT of the function $F(\omega) = 2i/\omega$ is $f(x) = \text{sgn}(x)$.*

Hints: Write the formula for the inverse Fourier transform of $F(\omega)$ as

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{2i}{\omega}\cos\omega x d\omega - \frac{i}{2\pi}\int_{-\infty}^{+\infty} \frac{2i}{\omega}\sin\omega x d\omega,$$

which reduces to

$$f(x) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{1}{\omega}\sin\omega x d\omega,$$

since the integrand of the first integral is odd. For $x > 0$ consider the Fourier transform of the function $\chi_x(t)$. For $x < 0$ perform the change of variables $u = -x$.

We saw earlier that the $F(\omega) = \chi_\Omega(\omega)$ has for its inverse Fourier transform the function $f(x) = \frac{\sin\Omega x}{\pi x}$; note that $f(0) = \frac{\Omega}{\pi}$ and $f(x) = 0$ for the first time when $\Omega x = \pi$ or $x = \frac{\pi}{\Omega}$. For any Ω-band-limited function $g(x)$ we have $G(\omega) = G(\omega)\chi_\Omega(\omega)$, so that, for any x_0, we have

$$g(x_0) = \int_{-\infty}^{\infty} g(x)\frac{\sin\Omega(x-x_0)}{\pi(x-x_0)}dx.$$

We describe this by saying that the function $f(x) = \frac{\sin\Omega x}{\pi x}$ has the *sifting property* for all Ω-band-limited functions $g(x)$.

As Ω grows larger, $f(0)$ approaches $+\infty$, while $f(x)$ goes to zero for $x \neq 0$. The limit is therefore not a function; it is a *generalized function* called *the Dirac delta function at zero*, denoted $\delta(x)$. For this reason the function $f(x) = \frac{\sin\Omega x}{\pi x}$ is called an *approximate delta function*. The FT of $\delta(x)$ is the function $F(\omega) = 1$ for all ω. The Dirac delta function $\delta(x)$ enjoys the *sifting property* for all $g(x)$; that is,

$$g(x_0) = \int_{-\infty}^{\infty} g(x)\delta(x-x_0)dx.$$

It follows from the sifting and shifting properties that the FT of $\delta(x-x_0)$ is the function $e^{ix_0\omega}$.

The formula for the inverse FT now says

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\omega}d\omega. \tag{1.3}$$

If we try to make sense of this integral according to the rules of calculus we get stuck quickly. The problem is that the integral formula doesn't mean quite what it does ordinarily and the $\delta(x)$ is not really a function, but an operator on functions; it is sometimes called a *distribution*. The Dirac deltas are mathematical fictions, not in the bad sense of being lies or fakes, but in the sense of being made up for some purpose. They provide helpful descriptions of impulsive forces, probability densities in which a discrete point has nonzero probability, or, in array processing, objects far enough away to be viewed as occupying a discrete point in space.

We shall treat the relationship expressed by Equation (1.3) as a formal statement, rather than attempt to explain the use of the integral in what is surely an unconventional manner.

If we move the discussion into the ω domain and define the Dirac delta function $\delta(\omega)$ to be the FT of the function that has the value $\frac{1}{2\pi}$ for all x, then the FT of the complex exponential function $\frac{1}{2\pi}e^{-i\omega_0 x}$ is $\delta(\omega-\omega_0)$, visualized as a "spike" at ω_0, that is, a generalized function that has the value $+\infty$ at $\omega=\omega_0$ and zero elsewhere. This is a useful result, in that it provides the motivation for considering the Fourier transform of a signal $s(t)$ containing hidden periodicities. If $s(t)$ is a sum of complex exponentials with frequencies $-\omega_n$, then its Fourier transform will consist of Dirac delta functions $\delta(\omega-\omega_n)$. If we then estimate the Fourier transform of $s(t)$ from sampled data, we are looking for the peaks in the Fourier transform that approximate the infinitely high spikes of these delta functions.

Exercise 1.6 *Use the fact that* $\mathrm{sgn}(x)=2u(x)-1$ *and the previous exercise to show that* $f(x)=u(x)$ *has the FT* $F(\omega)=i/\omega+\pi\delta(\omega)$.

Generally, the functions $f(x)$ and $F(\omega)$ are complex-valued, so that we may speak about their real and imaginary parts. The next exercise explores the connections that hold among these real-valued functions.

Exercise 1.7 *Let* $f(x)$ *be arbitrary and* $F(\omega)$ *its Fourier transform. Let* $F(\omega)=R(\omega)+iX(\omega)$, *where* R *and* X *are real-valued functions, and similarly, let* $f(x)=f_1(x)+if_2(x)$, *where* f_1 *and* f_2 *are real-valued. Find relationships between the pairs* R,X *and* f_1,f_2.

Exercise 1.8 *Let* f,F *be a FT pair. Let* $g(x)=\int_{-\infty}^{x}f(y)dy$. *Show that the FT of* $g(x)$ *is* $G(\omega)=\pi F(0)\delta(\omega)+\frac{iF(\omega)}{\omega}$.

Hint: For $u(x)$ the Heaviside function we have

$$\int_{-\infty}^{x}f(y)dy=\int_{-\infty}^{\infty}f(y)u(x-y)dy.$$

We can use properties of the Dirac delta functions to extend the Parseval equation to Fourier transforms, where it is usually called the *Parseval-Plancherel* equation.

Exercise 1.9 *Let $f(x), F(\omega)$ and $g(x), G(\omega)$ be Fourier transform pairs. Use Equation (1.3) to establish the Parseval-Plancherel equation*

$$\langle f, g\rangle = \int f(x)\overline{g(x)}dx = \frac{1}{2\pi}\int F(\omega)\overline{G(\omega)}d\omega,$$

from which it follows that

$$||f||^2 = \langle f, f\rangle = \int |f(x)|^2 dx = \frac{1}{2\pi}\int |F(\omega)|^2 d\omega.$$

Exercise 1.10 *We define the even part of $f(x)$ to be the function*

$$f_e(x) = \frac{f(x) + f(-x)}{2},$$

and the odd part of $f(x)$ to be

$$f_o(x) = \frac{f(x) - f(-x)}{2};$$

define F_e and F_o similarly for F the FT of f. Let $F(\omega) = R(\omega) + iX(\omega)$ be the decomposition of F into its real and imaginary parts. We say that f is a causal function if $f(x) = 0$ for all $x < 0$. Show that, if f is causal, then R and X are related; specifically, show that X is the Hilbert transform of R, that is,

$$X(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{R(\alpha)}{\omega - \alpha} d\alpha.$$

Hint: If $f(x) = 0$ for $x < 0$ then $f(x)\mathrm{sgn}(x) = f(x)$. Apply the convolution theorem, then compare real and imaginary parts.

Exercise 1.11 *The one-sided Laplace transform (LT) of f is $\mathcal{F}$ given by*

$$\mathcal{F}(z) = \int_0^{\infty} f(x)e^{-zx}dx.$$

Compute $\mathcal{F}(z)$ for $f(x) = u(x)$, the Heaviside function. Compare $\mathcal{F}(-i\omega)$ with the FT of u.

1.3 The FT in Higher Dimensions

The Fourier transform is also defined for functions of several real variables $f(x_1, ..., x_N) = f(\mathbf{x})$. The multidimensional FT arises in image processing, scattering, transmission tomography, and many other areas.

We adopt the usual vector notation that ω and $\mathbf{x}$ are N-dimensional real vectors. We say that $F(\omega)$ is the *N-dimensional Fourier transform* of the possibly complex-valued function $f(\mathbf{x})$ if the following relation holds:

$$F(\omega) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(\mathbf{x}) e^{i\omega \cdot \mathbf{x}} d\mathbf{x},$$

where $\omega \cdot \mathbf{x}$ denotes the vector dot product and $d\mathbf{x} = dx_1 dx_2 ... dx_N$. In most cases we then have

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} F(\omega) e^{-i\omega \cdot \mathbf{x}} d\omega / (2\pi)^N;$$

we describe this by saying that $f(\mathbf{x})$ is the *inverse Fourier transform* of $F(\omega)$.

Consider the FT of a function of two variables $f(x, y)$:

$$F(\alpha, \beta) = \int \int f(x, y) e^{i(x\alpha + y\beta)} dx dy.$$

We convert to polar coordinates using $(x, y) = r(\cos\theta, \sin\theta)$ and $(\alpha, \beta) = \rho(\cos\omega, \sin\omega)$. Then

$$F(\rho, \omega) = \int_0^{\infty} \int_{-\pi}^{\pi} f(r, \theta) e^{ir\rho\cos(\theta - \omega)} r dr d\theta. \tag{1.4}$$

Say that a function $f(x, y)$ of two variables is a *radial* function if $x^2 + y^2 = x_1^2 + y_1^2$ implies $f(x, y) = f(x_1, y_1)$, for all points (x, y) and (x_1, y_1); that is, $f(x, y) = g(\sqrt{x^2 + y^2})$ for some function g of one variable.

Exercise 1.12 *Show that if f is radial then its FT F is also radial. Find the FT of the radial function $f(x, y) = \frac{1}{\sqrt{x^2 + y^2}}$.*

Hints: Insert $f(r, \theta) = g(r)$ in Equation (1.4) to obtain

$$F(\rho, \omega) = \int_0^{\infty} \int_{-\pi}^{\pi} g(r) e^{ir\rho\cos(\theta - \omega)} r dr d\theta$$

or

$$F(\rho, \omega) = \int_0^{\infty} r g(r) [\int_{-\pi}^{\pi} e^{ir\rho\cos(\theta - \omega)} d\theta] dr. \tag{1.5}$$

Show that the inner integral is independent of ω, and then use the fact that

$$\int_{-\pi}^{\pi} e^{ir\rho\cos\theta} d\theta = 2\pi J_0(r\rho),$$

with J_0 the 0th order Bessel function, to get

$$F(\rho, \omega) = H(\rho) = 2\pi \int_0^\infty r g(r) J_0(r\rho) dr. \tag{1.6}$$

The function $H(\rho)$ is called the *Hankel transform* of $g(r)$. Summarizing, we say that if $f(x, y)$ is a radial function obtained using g then its Fourier transform $F(\alpha, \beta)$ is also a radial function, obtained using the Hankel transform of g.

Chapter 2

Convolution

Convolution is an important concept in signal processing and occurs in several distinct contexts. The reader may recall an earlier encounter with convolution in a course on differential equations. The simplest example of convolution is the nonperiodic convolution of finite vectors, which is what we do to the coefficients when we multiply two polynomials together.

2.1 Nonperiodic Convolution

Recall the algebra problem of multiplying one polynomial by another. Suppose

$$A(x) = a_0 + a_1 x + ... + a_M x^M$$

and

$$B(x) = b_0 + b_1 x + ... + b_N x^N.$$

Let $C(x) = A(x)B(x)$. With

$$C(x) = c_0 + c_1 x + ... + c_{M+N} x^{M+N},$$

each of the coefficients c_j, $j = 0, ..., M+N$, can be expressed in terms of the a_m and b_n (an easy exercise!). The vector $c = (c_0, ..., c_{M+N})$ is called the *nonperiodic convolution* of the vectors $a = (a_0, ..., a_M)$ and $b = (b_0, ..., b_N)$. Nonperiodic convolution can be viewed as a particular case of periodic convolution, as we shall see.

2.2 The DFT and the Vector DFT

As we just discussed, nonperiodic convolution is another way of looking at the multiplication of two polynomials. This relationship between convolution on the one hand and multiplication on the other is a fundamental aspect of convolution. Whenever we have a convolution we should ask what related mathematical objects are being multiplied. We ask this question now with regard to periodic convolution; the answer turns out to be the *vector discrete Fourier transform*.

For a doubly infinite sequence $\{f_n | -\infty < n < \infty\}$, the function of $F(\omega)$ given by the infinite series

$$F(\omega) = \sum_{n=-\infty}^{\infty} f_n e^{in\omega} \tag{2.1}$$

is sometimes called the *discrete-time Fourier transform* (DTFT) of the sequence. Suppose now that the only nonzero entries of the sequence are $f_0, ..., f_{N-1}$ and denote by $\mathbf{f}$ the N by 1 vector with these N entries.Then the DTFT becomes what we shall call here the *discrete Fourier transform* (DFT) of the vector $\mathbf{f}$:

$$DFT_{\mathbf{f}}(\omega) = \sum_{n=0}^{N-1} f_n e^{in\omega}. \tag{2.2}$$

For $k = 0, ..., N-1$, we evaluate $DFT_{\mathbf{f}}(\omega)$ at the N equi-spaced points $2\pi k/N$ in the interval $[0, 2\pi)$ to obtain the entries

$$F_k = DFT_{\mathbf{f}}(2\pi k/N) = \sum_{n=0}^{N-1} f_n e^{2\pi i n k/N}$$

of the vector $\mathbf{F}$, which we then call the *vector DFT* (vDFT) of the vector $\mathbf{f}$ and write $\mathbf{F} = vDFT_{\mathbf{f}}$. The *fast Fourier transform* algorithm (FFT) to be discussed later gives a quick way to calculate the vector $\mathbf{F}$ from the vector $\mathbf{f}$.

In the signal processing literature no special name is given to what we call here $DFT_{\mathbf{f}}$, and the vector DFT of $\mathbf{f}$ is called the DFT of $\mathbf{f}$. This is unfortunate, because the function of the continuous variable given in Equation (2.2) is the more fundamental entity, the vector DFT being merely the evaluation of that function at N equispaced points. If we should wish to evaluate $DFT_{\mathbf{f}}(\omega)$ at $M > N$ equispaced points, say, for example, for the purpose of graphing the function, we would *zero-pad* the vector $\mathbf{f}$ by appending $M - N$ zero entries, to obtain an M by 1 vector $\mathbf{g}$ and then calculate $vDFT_{\mathbf{g}}$. The functions $DFT_{\mathbf{f}}(\omega)$ and $DFT_{\mathbf{g}}(\omega)$ are the same, while $vDFT_{\mathbf{f}}$ and $vDFT_{\mathbf{g}}$ are not. The FFT algorithm is most efficient when N is a power of two, so it is common practice to zero-pad $\mathbf{f}$ using as M the smallest power of two not less than N.

In many of the applications of signal processing, the function we wish to estimate is viewed as having the form of $F(\omega)$ in Equation (2.1). The values f_n must be obtained through measurements, so that, in practice, we know only finitely many of them. The function $F(\omega)$ must then be estimated; the function $DFT_{\mathbf{f}}(\omega)$ is one possible choice. Because no special name is given to what we call here $DFT_{\mathbf{f}}(\omega)$ and insufficient attention is paid to it, it is easy to mistake the entries F_k of $vDFT_{\mathbf{f}}$ for exact values of $F(\omega)$, rather than what they really are, exact values of $DFT_{\mathbf{f}}(\omega)$.

In the exercises that follow we investigate properties of the vector DFT and relate it to periodic convolution.

2.3 Periodic Convolution

Given the N by 1 vectors $\mathbf{f}$ and $\mathbf{d}$ with complex entries f_n and d_n, respectively, we define a third N by 1 vector $\mathbf{f} * \mathbf{d}$, the *periodic convolution* of $\mathbf{f}$ and $\mathbf{d}$, to have the entries

$$(\mathbf{f} * \mathbf{d})_n = f_0 d_n + f_1 d_{n-1} + ... + f_n d_0 + f_{n+1} d_{N-1} + ... + f_{N-1} d_{n+1}.$$

The first exercise relates the periodic convolution to the vector DFT.

Exercise 2.1 *Let* $\mathbf{F} = vDFT_{\mathbf{f}}$ *and* $\mathbf{D} = vDFT_{\mathbf{d}}$. *Define a third vector* $\mathbf{E}$ *having for its kth entry* $E_k = F_k D_k$, *for* $k = 0, ..., N-1$. *Show that* $\mathbf{E}$ *is the vDFT of the vector* $\mathbf{f} * \mathbf{d}$.

The vector $vDFT_{\mathbf{f}}$ can be obtained from the vector $\mathbf{f}$ by means of matrix multiplcation by a certain matrix G, called the *DFT matrix*. The matrix G has an inverse that is easily computed and can be used to go from $\mathbf{F} = vDFT_{\mathbf{f}}$ back to the original $\mathbf{f}$. The details are in Exercise 2.2.

Exercise 2.2 *Let* G *be the* N *by* N *matrix whose entries are* $G_{jk} = e^{i(j-1)(k-1)2\pi/N}$. *The matrix* G *is sometimes called the DFT matrix. Show that the inverse of* G *is* $G^{-1} = \frac{1}{N} G^{\dagger}$, *where* $G^{\dagger}$ *is the conjugate transpose of the matrix* G. *Then* $\mathbf{f} * \mathbf{d} = G^{-1}\mathbf{E} = \frac{1}{N} G^{\dagger}\mathbf{E}$.

As mentioned previously, nonperiodic convolution is really a special case of periodic convolution. Extend the $M+1$ by 1 vector a to an $M+N+1$ by 1 vector by appending N zero entries; similarly, extend the vector b to an $M+N+1$ by 1 vector by appending zeros. The vector c is now the periodic convolution of these extended vectors. Therefore, since we have an efficient algorithm for performing periodic convolution, namely the Fast Fourier Transform algorithm (FFT), we have a fast way to do the periodic (and thereby nonperiodic) convolution and polynomial multiplication.

2.4 Differential and Difference Equations

The ordinary first-order differential equation $y'(t) + ay(t) = f(t)$, with initial condition $y(0) = 0$ has for its solution

$$y(t) = e^{-at} \int_0^t e^{as} f(s) ds.$$

One way to look at such differential equations is to consider $f(t)$ to be the input to a system having $y(t)$ as its output. The system determines which terms will occur on the left side of the differential equation. Here we want to consider the discrete analog of such differential equations.

We replace the first derivative with the first difference, $y(n+1) - y(n)$ and we replace the input with the sequence $f = \{f(n)\}$, to obtain the difference equation

$$y(n+1) - y(n) + ay(n) = f(n). \tag{2.3}$$

With $b = 1 - a$ and assuming, for convenience, that $0 < b < 1$, we have

$$y(n+1) - by(n) = f(n). \tag{2.4}$$

The solution is $y = \{y(n)\}$ given by

$$y(n) = b^n \sum_{k=-\infty}^{n} b^{-k} f(k). \tag{2.5}$$

Comparing this with the solution of the differential equation, we see that the term b^n plays the role of $e^{-at} = (e^{-a})^t$, so that $b = 1 - a$ is substituting for e^{-a}. The infinite sum replaces the infinite integral, with $b^{-k} f(k)$ replacing the integrand $e^{as} f(s)$.

We can rewrite Equation (2.5) as

$$y(n) = \sum_{j=0}^{+\infty} h(j) f(n-j), \tag{2.6}$$

for $h(j) = b^j$, for $j = 0, 1, ...$ Therefore, the output sequence $\{y\}$ is the convolution of the input sequence $\{f\}$ with the sequence $\{h\}$. Since

$$b^n \sum_{k=-\infty}^{n} b^{-k} = 1 - b$$

the sequence $(1-b)^{-1} y(n)$ is an infinite *moving-average* sequence formed from the sequence f.

2.4.1 z-Transforms

We can derive the solution in Equation (2.5) using z-transforms. The sequence $w = \{w(n) = y(n) - by(n-1)\}$ can be viewed as the output of a convolution system g with input y and $g(0) = 1$, $g(1) = -b$ and $g(j) = 0$, otherwise; that is, w is the convolution of the sequences y and g. The z-transform of any sequence $\{h(j)\}$ is defined as

$$H(z) = \sum_{j=-\infty}^{+\infty} h(j) z^{-j}.$$

Therefore, the z-transform of the sequence g is

$$G(z) = 1 - bz^{-1} = (z-b)/z$$

and the inverse $G(z)^{-1} = z/(z-b)$ describes the inverse system.

Exercise 2.3 *Use $G(z)^{-1}$ and the fact that $Y(z) = G(z)F(z)$ to obtain the solution of the difference equation, which we know to be*

$$y(n) = 1f(n) + b^1 f(n-1) + b^2 f(n-2) + ... = b^n \sum_{k=-\infty}^{n} b^{-k} f(k).$$

2.4.2 Time-Invariant Systems

Note that in Equation (2.6) the $h(j)$ do not depend on the particular n for which $y(n)$ is being calculated. This is described by saying that the system having input $\{f(n)\}$ and output $\{y(n)\}$ is a *linear, time-invariant* system. This happens because the original differential equation has constant coefficients, which are independent of the time variable t. The *differential operator* D that transforms the function $y(t)$ into the function $y'(t) + ay(t)$ is a linear, time-invariant operator.

Exercise 2.4 *Explore the analogy between the use of the z-transform to solve the difference equation and the use of the Laplace transform to solve the differential equation.*

Chapter 3

Linear Algebra and Geometry in Finite-Dimensional Space

We consider now geometric and linear-algebraic aspects of finite-dimensional inner product spaces.

3.1 The Geometry of Euclidean Space

We denote by R^J the real Euclidean space consisting of all J-dimensional column vectors $x = (x_1, ..., x_J)^T$ with real entries x_j; here the superscript T denotes the transpose of the 1 by J matrix (or, row vector) $(x_1, ..., x_J)$. We denote by C^J the collection of all J-dimensional column vectors $x = (x_1, ..., x_J)^\dagger$ with complex entries x_j; here the superscript $\dagger$ denotes the conjugate transpose of the 1 by J matrix (or, row vector) $(x_1, ..., x_J)$. We let χ stand for either R^J or C^J.

3.2 Inner Products

For $x = (x_1, ..., x_J)^T$ and $y = (y_1, ..., y_J)^T$ in R^J, the dot product $x \cdot y$ is defined to be

$$x \cdot y = \sum_{j=1}^{J} x_j y_j.$$

Note that we can write

$$x \cdot y = y^T x = x^T y,$$

where juxtaposition indicates matrix multiplication. The norm, or Euclidean length, of x is

$$||x|| = \sqrt{x \cdot x} = \sqrt{x^T x}.$$

The Euclidean distance between two vectors x and y in R^J is $||x - y||$.

For $x = (x_1, ..., x_J)^T$ and $y = (y_1, ..., y_J)^T$ in C^J, the dot product $x \cdot y$ is defined to be

$$x \cdot y = \sum_{j=1}^{J} x_j \overline{y_j}.$$

Note that we can write

$$x \cdot y = y^\dagger x.$$

The norm, or Euclidean length, of x is

$$||x|| = \sqrt{x \cdot x} = \sqrt{x^\dagger x}.$$

As in the real case, the distance between vectors x and y is $||x - y||$.

Both of the spaces R^J and C^J, along with their dot products, are examples of finite-dimensional Hilbert space. Much of what follows in these notes applies to both R^J and C^J. In such cases, we shall simply refer to the underlying space as $\mathcal{X}$ and refer to the associated dot product using the *inner product* notation $\langle x, y \rangle$.

3.3 Cauchy's Inequality

Cauchy's inequality tells us that

$$|\langle x, y \rangle| \leq ||x|| ||y||,$$

with equality if and only if $y = \alpha x$, for some scalar α.

Proof of Cauchy's inequality: To prove Cauchy's inequality for the complex vector dot product, we write $x \cdot y = |x \cdot y| e^{i\theta}$. Let t be a real variable and consider

$$0 \leq ||e^{-i\theta}x - ty||^2 = (e^{-i\theta}x - ty) \cdot (e^{-i\theta}x - ty)$$
$$= ||x||^2 - t[(e^{-i\theta}x) \cdot y + y \cdot (e^{-i\theta}x)] + t^2||y||^2$$
$$= ||x||^2 - t[(e^{-i\theta}x) \cdot y + \overline{(e^{-i\theta}x) \cdot y}] + t^2||y||^2$$
$$= ||x||^2 - 2Re(te^{-i\theta}(x \cdot y)) + t^2||y||^2$$
$$= ||x||^2 - 2Re(t|x \cdot y|) + t^2||y||^2 = ||x||^2 - 2t|x \cdot y| + t^2||y||^2.$$

This is a nonnegative quadratic polynomial in the variable t, so it cannot have two distinct real roots. Therefore, the discriminant $4|x \cdot y|^2 - 4||y||^2||x||^2$ must be non-positive; that is, $|x \cdot y|^2 \leq ||x||^2||y||^2$. This is Cauchy's inequality. ∎

Exercise 3.1 *Use Cauchy's inequality to show that*

$$||x + y|| \leq ||x|| + ||y||;$$

this is called the triangle inequality.

We say that the vectors x and y are *mutually orthogonal* if $\langle x, y \rangle = 0$.

3.4 Hyperplanes in Euclidean Space

For a fixed column vector a with Euclidean length one and a fixed scalar γ the *hyperplane* determined by a and γ is the set $H(a, \gamma) = \{z | \langle a, z \rangle = \gamma\}$.

Exercise 3.2 *Show that the vector a is orthogonal to the hyperplane $H = H(a, \gamma)$; that is, if u and v are in H, then a is orthogonal to $u - v$.*

For an arbitrary vector x in $\mathcal{X}$ and arbitrary hyperplane $H = H(a, \gamma)$, the *orthogonal projection* of x onto H is the member $z = P_H x$ of H that is closest to x.

Exercise 3.3 *Show that, for $H = H(a, \gamma)$, $z = P_H x$ is the vector*

$$z = P_H x = x + (\gamma - \langle a, x \rangle) a. \tag{3.1}$$

For $\gamma = 0$, the hyperplane $H = H(a, 0)$ is also a *subspace* of $\mathcal{X}$, meaning that, for every x and y in H and scalars α and β, the linear combination $\alpha x + \beta y$ is again in H; in particular, the zero vector 0 is in $H(a, 0)$.

3.5 Convex Sets in Euclidean Space

A subset C of $\mathcal{X}$ is said to be *convex* if, for every pair of members x and y of C, and for every α in the open interval $(0, 1)$, the vector $\alpha x + (1 - \alpha) y$ is also in C.

Exercise 3.4 *Show that the unit ball U in $\mathcal{X}$, consisting of all x with $||x|| \leq 1$, is convex, while the surface of the ball, the set of all x with $||x|| = 1$, is not convex.*

A convex set C is said to be *closed* if it contains all the vectors that lie on its boundary. Given any nonempty closed convex set C and an arbitrary vector x in $\mathcal{X}$, there is a unique member of C closest to x, denoted $P_C x$, the orthogonal (or metric) projection of x onto C. For example, if $C = U$, the unit ball, then $P_C x = x / ||x||$, for all x such that $||x|| > 1$, and $P_C x = x$ otherwise. If C is R^J_+, the nonnegative cone of R^J, consisting of all vectors x with $x_j \geq 0$, for each j, then $P_C x = x_+$, the vector whose entries are $\max(x_j, 0)$.

3.6 Analysis in Euclidean Space

We say that an infinite sequence $\{x^k\}$ of vectors in $\mathcal{X}$ *converges* to the vector x if the limit of $||x - x^k||$ is zero, as $k \to +\infty$; then x is called the limit of the sequence. An infinite sequence $\{x^k\}$ is said to be *bounded* if there is a positive constant $b > 0$ such that $||x^k|| \leq b$, for all k.

Exercise 3.5 *Show that any convergent sequence is bounded. Find a bounded sequence of real numbers that is not convergent.*

For any bounded sequence $\{x^k\}$, there is at least one subsequence, often denoted $\{x^{k_n}\}$, that is convergent; the notation implies that the positive integers k_n are ordered, so that $k_1 < k_2 < ....$ The limit of such a subsequence is then said to be a *cluster point* of the original sequence.

Exercise 3.6 *Show that your bounded, but not convergent, sequence found in the previous exercise, has a cluster point.*

Exercise 3.7 *Show that, if x is a cluster point of the sequence $\{x^k\}$, and if $||x - x^k|| \geq ||x - x^{k+1}||$, for all k, then x is the limit of the sequence.*

A subset C of $\mathcal{X}$ is said to be *closed* if, for every convergent sequence $\{x^k\}$ of vectors in C, the limit point is again in C. For example, in $\mathcal{X} = \mathcal{R}$, the set $C = (0, 1]$ is not closed, because it does not contain the point $x = 0$, which is the limit of the sequence $\{x^k = \frac{1}{k}\}$; the set $[0, 1]$ is closed and is the *closure* of the set $(0, 1]$, that is, it is the smallest closed set containing $(0, 1]$.

When we investigate iterative algorithms, we will want to know if the sequence $\{x^k\}$ generated by the algorithm converges. As a first step, we will usually ask if the sequence is bounded? If it is bounded, then it will have at least one cluster point. We then try to discover if that cluster point is really the limit of the sequence.

3.7 Basic Linear Algebra

We begin with some definitions. Let S be a subspace of finite-dimensional Euclidean space R^J and Q a J by J Hermitian matrix, which means that Q is its own conjugate-transpose, $Q = Q^\dagger$, . We denote by $Q(S)$ the set

$$Q(S) = \{\mathbf{t} | \text{there exists}\, \mathbf{s} \in S \,\text{with}\, \mathbf{t} = Q\mathbf{s}\}$$

and by $Q^{-1}(S)$ the set

$$Q^{-1}(S) = \{\mathbf{u} | Q\mathbf{u} \in S\}.$$

Note that the set $Q^{-1}(S)$ is defined whether or not Q is invertible.

We denote by $S^\perp$ the set of vectors $\mathbf{u}$ that are orthogonal to every member of S; that is,

$$S^\perp = \{\mathbf{u}|\mathbf{u}^\dagger\mathbf{s} = 0, \text{for every}\, \mathbf{s} \in S\}.$$

Let H be a J by N matrix. Then $CS(H)$, the column space of H, is the subspace of R^J consisting of all the linear combinations of the columns of H. The null space of $H^\dagger$, denoted $NS(H^\dagger)$, is the subspace of R^J containing all the vectors $\mathbf{w}$ for which $H^\dagger\mathbf{w} = 0$.

Exercise 3.8 *Show that* $CS(H)^\perp = NS(H^\dagger)$.

Hint: If $\mathbf{v} \in CS(H)^\perp$, then $\mathbf{v}^\dagger H\mathbf{x} = 0$ for all $\mathbf{x}$, including $\mathbf{x} = H^\dagger\mathbf{v}$.

Exercise 3.9 *Show that* $CS(H) \cap NS(H^\dagger) = \{\mathbf{0}\}$.

Hint: If $\mathbf{y} = H\mathbf{x} \in NS(H^\dagger)$ consider $||\mathbf{y}||^2 = \mathbf{y}^\dagger\mathbf{y}$.

Exercise 3.10 *Let* S *be any subspace of* R^J. *Show that if* Q *is invertible and* $Q(S) = S$ *then* $Q^{-1}(S) = S$.

Hint: If $Q\mathbf{t} = Q\mathbf{s}$ then $\mathbf{t} = \mathbf{s}$.

Exercise 3.11 *Let* Q *be Hermitian. Show that* $Q(S)^\perp = Q^{-1}(S^\perp)$ *for every subspace* S. *If* Q *is also invertible then* $Q^{-1}(S)^\perp = Q(S^\perp)$. *Find an example of a non-invertible* Q *for which* $Q^{-1}(S)^\perp$ *and* $Q(S^\perp)$ *are different.*

We assume, now, that Q is Hermitian and invertible and that the matrix $H^\dagger H$ is invertible. Note that the matrix $H^\dagger Q^{-1}H$ need not be invertible under these assumptions. We shall denote by S an arbitrary subspace of R^J.

Exercise 3.12 *Show that* $Q(S) = S$ *if and only if* $Q(S^\perp) = S^\perp$.

Hint: Use Exercise 3.11.

Exercise 3.13 *Show that if* $Q(CS(H)) = CS(H)$ *then* $H^\dagger Q^{-1}H$ *is invertible.*

Hint: Show that $H^\dagger Q^{-1}H\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. Recall that $Q^{-1}H\mathbf{x} \in CS(H)$, by Exercise 3.11. Then use Exercise 3.9.

3.8 Linear and Nonlinear Operators

In our study of iterative algorithms we shall be concerned with sequences of vectors $\{x^k|k = 0, 1, ...\}$. The core of an iterative algorithm is the transition from the current vector x^k to the next one x^{k+1}. To understand the algorithm, we must understand the operation (or operator) T by which x^k is transformed into $x^{k+1} = Tx^k$. An *operator* is any function T defined on R^J or C^J with values again in the same space.

An operator T is *Lipschitz continuous* if there is a positive constant λ such that

$$||Tx - Ty|| \leq \lambda||x - y||,$$

for all x and y in $\mathcal{X}$.

Exercise 3.14 *Prove the following identity relating an arbitrary operator T to its complement $G = I - T$:*

$$||x - y||^2 - ||Tx - Ty||^2 = 2\langle Gx - Gy, x - y\rangle - ||Gx - Gy||^2. \qquad (3.2)$$

3.8.1 Linear and Affine-Linear Operators

For example, if $\mathcal{X} = C^J$ and A is a J by J complex matrix, then we can define an operator T by setting $Tx = Ax$, for each x in C^J; here Ax denotes the multiplicaton of the matrix A and the column vector x. Such operators are *linear operators*:

$$T(\alpha x + \beta y) = \alpha Tx + \beta Ty,$$

for each pair of vectors x and y and each pair of scalars α and β.

Exercise 3.15 *Show that, for $H = H(a, \gamma)$, $H_0 = H(a, 0)$, and any x and y in $\mathcal{X}$,*

$$P_H(x + y) = P_H x + P_H y - P_H 0,$$

so that

$$P_{H_0}(x + y) = P_{H_0}x + P_{H_0}y,$$

that is, the operator P_{H_0} is an additive operator. Also, show that

$$P_{H_0}(\alpha x) = \alpha P_{H_0}x,$$

so that P_{H_0} is a linear operator. Show that we can write P_{H_0} as a matrix multiplication:

$$P_{H_0}x = (I - aa^\dagger)x.$$

If d is a fixed nonzero vector in C^J, the operator defined by $Tx = Ax+d$ is not a linear operator; it is called an *affine linear operator*.

Exercise 3.16 *Show that, for any hyperplane* $H = H(a, \gamma)$ *and* $H_0 = H(a, 0)$,

$$P_H x = P_{H_0} x + P_H 0,$$

so P_H *is an affine linear operator.*

Exercise 3.17 *For* $i = 1, ..., I$ *let* H_i *be the hyperplane* $H_i = H(a^i, \gamma_i)$, $H_{i0} = H(a^i, 0)$, *and* P_i *and* P_{i0} *the orthogonal projections onto* H_i *and* H_{i0}, *respectively. Let* T *be the operator* $T = P_I P_{I-1} \cdots P_2 P_1$. *Show that* T *is an affine linear operator, that is,* T *has the form*

$$Tx = Bx + d,$$

for some matrix B *and some vector* d. *Hint: Use the previous exercise and the fact that* P_{i0} *is linear to show that*

$$B = (I - a^I (a^I)^\dagger) \cdots (I - a^1 (a^1)^\dagger).$$

3.8.2 Orthogonal Projection onto Convex Sets

For an arbitrary nonempty closed convex set C, the orthogonal projection $T = P_C$ is a nonlinear operator, unless, of course, $C = H(a, 0)$ for some vector a. We may not be able to describe $P_C x$ explicitly, but we do know a useful property of $P_C x$.

Proposition 3.1 *For a given* x, *the vector* z *is* $P_C x$ *if and only if*

$$\langle c - z, z - x \rangle \geq 0,$$

for all c *in the set* C.

Proof: For simplicity, we consider only the real case, $\mathcal{X} = R^J$. Let c be arbitrary in C and α in $(0, 1)$. Then

$$||x - P_C x||^2 \leq ||x - (1 - \alpha) P_C x - \alpha c||^2 = ||x - P_C x + \alpha (P_C x - c)||^2$$

$$= ||x - P_C x||^2 - 2\alpha \langle x - P_C x, c - P_C x \rangle + \alpha^2 ||P_C x - c||^2.$$

Therefore,

$$-2\alpha \langle x - P_C x, c - P_C x \rangle + \alpha^2 ||P_C x - c||^2 \geq 0,$$

so that

$$2\langle x - P_C x, c - P_C x \rangle \leq \alpha ||P_C x - c||^2$$

Taking the limit, as $\alpha \to 0$, we conclude that

$$\langle c - P_C x, P_C x - x \rangle \geq 0.$$

If z is a member of C that also has the property

$$\langle c - z, z - x \rangle \geq 0,$$

for all c in C, then we have both

$$\langle z - P_C x, P_C x - x \rangle \geq 0,$$

and

$$\langle z - P_C x, x - z \rangle \geq 0.$$

Adding on both sides of these two inequalities lead to

$$\langle z - P_C x, P_C x - z \rangle \geq 0.$$

But,

$$\langle z - P_C x, P_C x - z \rangle = -||z - P_C x||^2,$$

so it must be the case that $z = P_C x$. This completes the proof. ∎

3.8.3 Gradient Operators

Another important example of a nonlinear operator is the gradient of a real-valued function of several variables. Let $f(x) = f(x_i, ..., x_J)$ be a real number for each vector x in R^J. The *gradient* of f at the point x is the vector whose entries are the partial derivatives of f; that is,

$$\nabla f(x) = (\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_J}(x))^T.$$

The operator $Tx = \nabla f(x)$ is linear only if the function $f(x)$ is quadratic; that is, $f(x) = x^T A x$ for some square matrix x, in which case the gradient of f is $\nabla f(x) = \frac{1}{2}(A + A^T)x$.

3.9 Eigenvalues and Matrix Norms

Let S be a complex, square matrix. We say that λ is an eigenvalue of S if λ is a root of the complex polynomial $\det(\lambda I - S)$. Therefore, each S has as many (possibly complex) eigenvalues as it has rows or columns, although some of the eigenvalues may be repeated.

An equivalent definition is that λ is an eigenvalue of S if there is a non-zero vector x with $Sx = \lambda x$, in which case the vector x is called an *eigenvector* of S. From this definition we see that the matrix S is invertible if and only if zero is not one of its eigenvalues. The *spectral radius* of S, denoted $\rho(S)$, is the maximum of $|\lambda|$, over all eigenvalues λ of S.

Exercise 3.18 *Show that* $\rho(S^2) = \rho(S)^2$.

Exercise 3.19 *We say that S is Hermitian or self-adjoint if $S^\dagger = S$. Show that, if S is Hermitian, then every eigenvalue of S is real. Hint: suppose that $Sx = \lambda x$. Then consider $x^\dagger Sx$.*

A Hermitian matrix S is *positive-definite* if each of its eigenvalues is positive. If S is an I by I Hermitian matrix with (necessarily real) eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_I,$$

and associated (column) eigenvectors $\{u_i \,|i = 1, ..., I\}$ (which we may assume are mutually orthogonal), then S can be written as

$$S = \lambda_1 u_1 u_1^\dagger + \cdots + \lambda_I u_I u_I^\dagger.$$

This is the *eigenvalue/eigenvector decomposition* of S. The Hermitian matrix S is invertible if and only if all of its eigenvalues are non-zero, in which case we can write the inverse of S as

$$S^{-1} = \lambda_1^{-1} u_1 u_1^\dagger + \cdots + \lambda_I^{-1} u_I u_I^\dagger.$$

It follows from the eigenvector decomposition of S that $S = QQ^\dagger$ for some Hermitian, positive-definite matrix Q, called the *Hermitian square root* of S.

Let $||x||$ be any norm on C^J, not necessarily the Euclidean norm, $|||b|||$ any norm on C^I, and A a rectangular I by J matrix. The *matrix norm of* A, denoted $||A||$, derived from the two vectors norms, is the smallest positive constant c such that

$$|||Ax||| \leq c||x||,$$

for all x in C^J. If we choose the two vector norms carefully, then we can get an explicit description of $||A||$, but, in general, we cannot. We shall be particularly interested in the 2-norm of the square matrix A, denoted $||A||_2$, which is the matrix norm derived from the Euclidean vector norms. Unless otherwise stated, we shall understand $||A||$ to be the 2-norm of A.

From the definition of the 2-norm of A, we know that

$$||A|| = \max\{||Ax||/||x||\},$$

with the maximum over all nonzero vectors x. Since

$$||Ax||^2 = x^\dagger A^\dagger Ax,$$

we have

$$||A|| = \sqrt{\max\{\frac{x^\dagger A^\dagger Ax}{x^\dagger x}\}},$$

over all nonzero vectors x. The term inside the square root is also the largest eigenvalue of the Hermitian nonnegative-definite matrix $S = A^\dagger A$. Therefore, we can say

$$||A|| = \sqrt{\rho(A^\dagger A)}.$$

Exercise 3.20 *Show that, if S is Hermitian, then the 2-norm of S is $||S|| = \rho(S)$. Hint: use Exercise (3.18).*

Exercise 3.21 *Show that, for any square matrix S and any matrix norm $||S||$, we have $||S|| \geq \rho(S)$.*

3.9.1 Gerschgorin's Theorem

Gerschgorin's theorem gives us a way to estimate the eigenvalues of an arbitrary square matrix A.

Theorem 3.1 *Let A be J by J. For $j = 1, ..., J$, let C_j be the circle in the complex plane with center A_{jj} and radius $r_j = \sum_{m \neq j} |A_{jm}|$. Then every eigenvalue of A lies within one of the C_j.*

Proof: Let λ be an eigenvalue of A, with associated eigenvector u. Let u_j be the entry of the vector u having the largest absolute value. From $Au = \lambda u$, we have

$$(\lambda - A_{jj})u_j = \sum_{m \neq j} A_{jm} u_m,$$

so that

$$|\lambda - A_{jj}| \leq \sum_{m \neq j} |A_{jm}||u_m|/|u_j| \leq r_j.$$

This completes the proof. ∎

Index