# Selected Topics in Applied Mathematics

**Charles L. Byrne**
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854

August 1, 2014

(Supplementary readings for 92.530–531
Applied Mathematics I and II)
(The most recent version is available as a pdf file at
http://faculty.uml.edu/cbyrne/cbyrne.html)

# Contents

# Chapter 1

# Preface

These are notes on various topics in applied mathematics, designed to supplement the text for the courses 92.530 Applied Mathematics I and 92.531 Applied Mathematics II. The text for these courses is *Advanced Mathematics for Engineers and Scientists*, M. Spiegel, McGraw-Hill Schaum's Outline Series, ISBN 978-0-07-163540-0. Chapter references in the notes are to chapters in this text.

For extra credit work there is one chapter containing well known problems in applied mathematics and other exercises scattered throughout these notes.

# Part I

# Readings for Applied Mathematics I

# Chapter 2

# More Fundamentals(Chapter 1)

## 2.1 The Dot Product

Let $\mathbb{R}^N$ denote the collection of all $N$-dimensional column vectors of real numbers; for example,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \end{bmatrix},$$

where each of the $x_n$, $n = 1, 2, ..., N$ is some real number. When $N = 1$ we write $\mathbb{R}^1 = \mathbb{R}$, the collection of all real numbers. For notational convenience, we sometimes write

$$x^T = (x_1, x_2, ..., x_N),$$

which is the *transpose* of the column vector $x$.

If $x$ and $y$ are members of $\mathbb{R}^N$, then the *dot product* of $x$ and $y$ is the real number

$$x \cdot y = x_1 y_1 + x_2 y_2 + ... + x_N y_N.$$

The *magnitude* or *size* of a vector $x$ is

$$\|x\| = \sqrt{x_1^2 + x_2^2 + ... + x_N^2} = \sqrt{x \cdot x}.$$

When $N = 2$ or $N = 3$ we can give more meaning to the dot product.

For $N = 2$ or $N = 3$ we have

$$x \cdot y = \|x\|\|y\| \cos\theta,$$

where $\theta$ is the angle between $x$ and $y$, when they are viewed as directed line segments in a plane, emerging from a common base point.

In general, when $N$ is larger, the angle between $x$ and $y$ no longer makes sense, but we still have a useful inequality, called *Cauchy's Inequality*:

$$|x \cdot y| \leq \|x\|\|y\|,$$

and

$$|x \cdot y| = \|x\|\|y\|$$

precisely when, or *if and only if*, as mathematicians say, $x$ and $y$ are parallel, that is, there is a real number $\alpha$ with

$$y = \alpha x.$$

## 2.2 The Gradient and Directional Derivatives

Let $f(x_1, x_2, ..., x_N)$ be a real-valued function of $N$ real variables, which we denote by $f : \mathbb{R}^N \to \mathbb{R}$. For such functions we are interested in their first partial derivatives. The first partial derivative of $f$, at the point $(x_1, x_2, ..., x_N)$, in the direction of $x_n$ is defined to be

$$\lim_{h \to 0} \frac{f(x_1, x_2, ..., x_{n-1}, x_n + h, x_{n+1}, ..., x_N) - f(x_1, x_2, ..., x_{n-1}, x_n, x_{n+1}, ..., x_N)}{h},$$

provided that this limit exists. We denote this limit as $f_n(x_1, ..., x_N)$, or $\frac{\partial f}{\partial x_n}(x_1, ..., x_N)$. When all the first partial derivatives of $f$ exist at a point we say that $f$ is differentiable at that point.

When we are dealing with small values of $N$, such as $N = 3$, it is common to write $f(x, y, z)$, where now $x$, $y$, and $z$ are real variables, not vectors. Then the first partial derivatives can be denoted $f_x$, $f_y$, and $f_z$.

The *gradient* of the function $f : \mathbb{R}^N \to \mathbb{R}$ at the point $(x_1, x_2, ..., x_N)$, written $\nabla f(x_1, ..., x_N)$, is the column vector whose entries are the first partial derivatives of $f$ at that point.

Let $d$ be a member of $\mathbb{R}^N$ with $\|d\| = 1$; then $d$ is called a *direction vector*. The *directional derivative* of $f$, at the point $(x_1, ..., x_N)$, in the direction of $d$, is

$$\nabla f(x_1, ..., x_N) \cdot d.$$

From Cauchy's Inequality we see that the absolute value of the directional derivative at a given point is at most the magnitude of the gradient at that point, and is equal to that magnitude precisely when $d$ is parallel to

the gradient. It follows that the direction in which the gradient points is the direction of greatest increase in $f$, and the opposite direction is the direction of greatest decrease. The gradient, therefore, is perpendicular to the tangent plane to the surface of constant value, the level surface, passing through this point. These facts are important in optimization, when we try to find the largest and smallest values of $f$.

## 2.3 Optimization

If $f : \mathbb{R} \to \mathbb{R}$ is a differentiable real-valued function of a real variable, and we want to find its local maxima and minima, we take the derivative and set it to zero. When $f : \mathbb{R}^N \to \mathbb{R}$ is a differentiable real-valued function of $N$ real variables, we find local maxima and minima by calculating the gradient and finding out where the gradient is zero, that is, where all the first partial derivatives are zero.

## 2.4 Lagrange Multipliers

If $f : \mathbb{R} \to \mathbb{R}$ is differentiable and we want to maximize or minimize $f$ for $x$ in the interval $[a, b]$, that is, we want to solve a constrained optimization problem, we must not look only at the places where the derivative is zero, but we must check the endpoints also. For functions of more than one variable, constrained optimization problems are more difficult. Consider the following example.

Let $f(x, y) = x^2 + y^2$ and suppose we want to minimize $f$, but only for those points $(x, y)$ with $\frac{x}{2} + \frac{y}{3} - 1 = 0$. One way is to solve for $y$, getting $y = \frac{-3}{2}x + 3$, putting this into $x^2 + y^2$ to get a function of $x$ alone, and then minimizing that function of $x$. This does not always work, though. Lagrange multipliers can help in more complicated cases.

Suppose that we want to minimize a differentiable function $f(x_1, ..., x_N)$, subject to $g(x_1, ..., x_N) = 0$, where $g$ is another differentiable real-valued function. The function $f$ determines *level surfaces*, which are the sets of all points in $\mathbb{R}^N$ on which $f$ has the same value; think of elevation lines on a map. Similarly, $g$ determines its own set of level surfaces. Our constraint is that we must consider only those points in $\mathbb{R}^N$ on the level surface where $g = 0$. At the solution point $(x_1^*, ..., x_N^*)$, the level surface for $g = 0$ must be tangent to a level surface of $f$, which says that the gradient of $g$ must be parallel to the gradient of $f$ at that point; in other words, there is a real number $\alpha$ such that

$$\nabla g(x_1^*, ..., x_N^*) = \alpha \nabla f(x_1^*, ..., x_N^*),$$

which we can write in the more traditional way as

$$\nabla f(x_1^*, ..., x_N^*) + \lambda \nabla g(x_1^*, ..., x_N^*) = 0.$$

Suppose then that we form the function

$$h(x_1, ..., x_N) = f(x_1, ..., x_N) + \lambda g(x_1, ..., x_N).$$

Then we want $(x_1^*, ..., x_N^*)$ such that

$$\nabla h(x_1^*, ..., x_N^*) = 0.$$

This is the *Lagrange multiplier* approach.

Let's return to the problem of minimizing $f(x, y) = x^2 + y^2$, subject to the constraint $g(x, y) = \frac{x}{2} + \frac{y}{3} - 1 = 0$. Then

$$h(x, y) = x^2 + y^2 + \lambda(\frac{x}{2} + \frac{y}{3} - 1).$$

Setting $h_x = 0$ and $h_y = 0$ we find that we need

$$2x + \frac{\lambda}{2} = 0,$$

and

$$2y + \frac{\lambda}{3} = 0.$$

We don't know what $\lambda$ is, but we don't care, because we can write

$$\lambda = -4x,$$

and

$$\lambda = -6y,$$

from which we conclude that $y = \frac{2}{3}x$. This is a second relationship between $x$ and $y$ and now we can find the answer.

## 2.5 Richardson's Method

We know that if $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

Suppose that we want to approximate $f'(x)$ numerically, by taking a small value of $h$. Eventually, if $h$ is too small, we run into trouble, because both the top and the bottom of the difference quotient go to zero as $h$ goes to zero. So we are dividing a small number by a small number, which is to

be avoided, according to the rules of numerical analysis. The download file on Richardson's method that is available on the website shows what can happen, as we try to calculate $f'(3)$ for $f(x) = \log x$.

We know from the Taylor expansion that, if $f$ is a nice function, then

$$\frac{f(x+h) - f(x)}{h} = f'(x) + \frac{1}{2!}f''(x)h + \frac{1}{3!}f'''(x)h^2 + ... \tag{2.1}$$

Similarly,

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{1}{3!}f'''(x)h^2 + ... \tag{2.2}$$

Therefore, the left side of Equation (2.1) goes to $f'(x)$ on the order of $h$, while the left side of Equation (2.2) goes to $f'(x)$ on the order of $h^2$. This tells us that if we use Equation (2.2) to estimate $f'(x)$ we won't need to take $h$ as small to get a good answer. This is the basic idea of Richardson's method, which can be applied to other types of problems, such as approximating integrals.

## 2.6 Leibnitz's Rule and Distributions

Let $f(x,t)$ be a real-valued function of two real variables. Then

$$\int_a^b f(x,t)dx = F(t) \tag{2.3}$$

is a function of $t$ alone. The derivative of $F(t)$, if it exists, can be expressed in terms of the partial derivative, with respect to $t$, of $f(x,t)$:

$$F'(t) = \int_a^b \frac{\partial f}{\partial t}(x,t)dx. \tag{2.4}$$

Leibnitz's Rule extends this formula to the case in which $a$ and $b$ are also allowed to depend on $t$.

Let $h(t)$ and $g(t)$ be real-valued functions of $t$. For convenience, we assume that $g(t) < h(t)$ for all $t$. Let

$$F(t) = \int_{g(t)}^{h(t)} f(x,t)dx. \tag{2.5}$$

Leibnitz's Rule then states that

$$F'(t) = \int_{g(t)}^{h(t)} \frac{\partial f}{\partial t}(x,t)dx + f(h(t),t)h'(t) - f(g(t),t)g'(t). \tag{2.6}$$

We can use distributions to see why this is plausible.

Distribution theory allows us to extend the notion of derivative to functions that do not possess derivatives in the ordinary sense, such as the Heaviside function $U(x)$, which equals one for $x \geq 0$ and zero for $x < 0$. Integration by parts is the key here.

Suppose that $v(x)$ is differentiable and goes to zero as $|x|$ approaches $+\infty$. Then integration by parts tells us that

$$\int_{-\infty}^{+\infty} u'(x)v(x)dx = -\int_{-\infty}^{+\infty} u(x)v'(x)dx. \tag{2.7}$$

If $u(x)$ doesn't have a derivative in the usual sense, we define $u'(x)$ as the *generalized* function $u'(x)$ that has the property described in Equation (2.7).

For example, let $u(x) = U(x)$, the Heaviside function. Then $U'(x)$ has the property that, for all $v(x)$ as above,

$$\int_{-\infty}^{+\infty} U'(x)v(x)dx = -\int_{0}^{+\infty} v'(x)dx = v(0). \tag{2.8}$$

Therefore, $U'(x)$ can be defined by the property

$$\int_{-\infty}^{+\infty} U'(x)v(x)dx = v(0). \tag{2.9}$$

But Equation (2.9) is also the definition of the generalized function (or distribution) called the Dirac delta function, denoted $\delta(x)$. So $U'(x) = \delta(x)$. We can now use this to motivate Leibnitz's Rule.

Denote by $\chi_{[a,b]}(x)$ the function that is one for $a \leq x \leq b$ and zero otherwise; note that

$$\chi_{[a,b]}(x) = U(x-a) - U(x-b), \tag{2.10}$$

so that the derivative of $\chi_{[a,b]}(x)$, in the distributional sense, is

$$\chi'_{[a,b]}(x) = \delta(x-a) - \delta(x-b). \tag{2.11}$$

Then we can write

$$F(t) = \int_{g(t)}^{h(t)} f(x,t)dx = \int_{-\infty}^{+\infty} \chi_{[g(t),h(t)]}(x)f(x,t)dx. \tag{2.12}$$

The function $c(x,t) = \chi_{[g(t),h(t)]}(x)$ has the distributional partial derivative, with respect to $t$, of

$$\frac{\partial c}{\partial t}(x,t) = -g'(t)\delta(x-g(t)) + h'(t)\delta(x-h(t)). \tag{2.13}$$

Using the product rule and differentiating under the integral sign, we get

$$F'(t) = \int_{g(t)}^{h(t)} \frac{\partial f}{\partial t}(x,t)dx + h'(t)f(h(t),t) - g'(t)f(g(t),t). \tag{2.14}$$

## 2.7 The Complex Exponential Function

The most important function in signal processing is the complex-valued function of the real variable $x$ defined by

$$h(x) = \cos(x) + i\sin(x). \tag{2.15}$$

For reasons that will become clear shortly, this function is called the *complex exponential function*. Notice that the magnitude of the complex number $h(x)$ is always equal to one, since $\cos^2(x) + \sin^2(x) = 1$ for all real $x$. Since the functions $\cos(x)$ and $\sin(x)$ are $2\pi$-periodic, that is, $\cos(x+2\pi) = \cos(x)$ and $\sin(x+2\pi) = \sin(x)$ for all $x$, the complex exponential function $h(x)$ is also $2\pi$-periodic.

### 2.7.1 Real Exponential Functions

In calculus we encounter functions of the form $g(x) = a^x$, where $a > 0$ is an arbitrary constant. These functions are the *exponential* functions, the most well-known of which is the function $g(x) = e^x$. Exponential functions are those with the property

$$g(u+v) = g(u)g(v) \tag{2.16}$$

for every $u$ and $v$. Recall from calculus that for exponential functions $g(x) = a^x$ with $a > 0$ the derivative $g'(x)$ is

$$g'(x) = a^x \ln(a) = g(x)\ln(a). \tag{2.17}$$

Now we consider the function $h(x)$ in light of these ideas.

### 2.7.2 Why is $h(x)$ an Exponential Function?

We show now that the function $h(x)$ in Equation (2.15) has the property given in Equation (2.16), so we have a right to call it an exponential function; that is, $h(x) = c^x$ for some constant $c$. Since $h(x)$ has complex values, the constant $c$ cannot be a real number, however.

Calculating $h(u)h(v)$, we find

$$h(u)h(v) = (\cos(u)\cos(v) - \sin(u)\sin(v)) + i(\cos(u)\sin(v) + \sin(u)\cos(v))$$

$$= \cos(u+v) + i\sin(u+v) = h(u+v).$$

So $h(x)$ is an exponential function; $h(x) = c^x$ for some complex constant $c$. Inserting $x = 1$, we find that $c$ is

$$c = \cos(1) + i\sin(1).$$

Let's find another way to express $c$, using Equation (2.17). Since

$$h'(x) = -\sin(x) + i\cos(x) = i(\cos(x) + i\sin(x)) = ih(x),$$

we conjecture that $\ln(c) = i$; but what does this mean?

For $a > 0$ we know that $b = \ln(a)$ means that $a = e^b$. Therefore, we say that $\ln(c) = i$ means $c = e^i$; but what does it mean to take $e$ to a complex power? To define $e^i$ we turn to the Taylor series representation for the exponential function $g(x) = e^x$, defined for real $x$:

$$e^x = 1 + x + x^2/2! + x^3/3! + ....$$

Inserting $i$ in place of $x$ and using the fact that $i^2 = -1$, we find that

$$e^i = (1 - 1/2! + 1/4! - ...) + i(1 - 1/3! + 1/5! - ...);$$

note that the two series are the Taylor series for $\cos(1)$ and $\sin(1)$, respectively, so $e^i = \cos(1) + i\sin(1)$. Then the complex exponential function in Equation (2.15) is

$$h(x) = (e^i)^x = e^{ix}.$$

Inserting $x = \pi$, we get

$$h(\pi) = e^{i\pi} = \cos(\pi) + i\sin(\pi) = -1$$

or

$$e^{i\pi} + 1 = 0,$$

which is the remarkable relation discovered by Euler that combines the five most important constants in mathematics, $e$, $\pi$, $i$, 1, and 0, in a single equation.

Note that $e^{2\pi i} = e^{0i} = e^0 = 1$, so

$$e^{(2\pi+x)i} = e^{2\pi i}e^{ix} = e^{ix}$$

for all $x$.

### 2.7.3 What is $e^z$, for $z$ complex?

We know from calculus what $e^x$ means for real $x$, and now we also know what $e^{ix}$ means. Using these we can define $e^z$ for any complex number $z = a + ib$ by $e^z = e^{a+ib} = e^a e^{ib}$.

We know from calculus how to define $\ln(x)$ for $x > 0$, and we have just defined $\ln(c) = i$ to mean $c = e^i$. But we could also say that $\ln(c) = i(1 + 2\pi k)$ for any integer $k$; that is, the periodicity of the complex exponential function forces the function $\ln(x)$ to be multi-valued.

For any nonzero complex number $z = |z|e^{i\theta(z)}$, we have

$$\ln(z) = \ln(|z|) + \ln(e^{i\theta(z)}) = \ln(|z|) + i(\theta(z) + 2\pi k),$$

for any integer $k$. If $z = a > 0$ then $\theta(z) = 0$ and $\ln(z) = \ln(a) + i(k\pi)$ for any even integer $k$; in calculus class we just take the value associated with $k = 0$. If $z = a < 0$ then $\theta(z) = \pi$ and $\ln(z) = \ln(-a) + i(k\pi)$ for any odd integer $k$. So we can define the logarithm of a negative number; it just turns out not to be a real number. If $z = ib$ with $b > 0$, then $\theta(z) = \frac{\pi}{2}$ and $\ln(z) = \ln(b) + i(\frac{\pi}{2} + 2\pi k)$ for any integer $k$; if $z = ib$ with $b < 0$, then $\theta(z) = \frac{3\pi}{2}$ and $\ln(z) = \ln(-b) + i(\frac{3\pi}{2} + 2\pi k)$ for any integer $k$.

Adding $e^{-ix} = \cos(x) - i\sin(x)$ to $e^{ix}$ given by Equation (2.15), we get

$$\cos(x) = \frac{1}{2}(e^{ix} + e^{-ix});$$

subtracting, we obtain

$$\sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix}).$$

These formulas allow us to extend the definition of cos and sin to complex arguments $z$:

$$\cos(z) = \frac{1}{2}(e^{iz} + e^{-iz})$$

and

$$\sin(z) = \frac{1}{2i}(e^{iz} - e^{-iz}).$$

In signal processing the complex exponential function is often used to describe functions of time that exhibit periodic behavior:

$$h(\omega t + \theta) = e^{i(\omega t+\theta)} = \cos(\omega t + \theta) + i\sin(\omega t + \theta),$$

where the *frequency* $\omega$ and *phase angle* $\theta$ are real constants and $t$ denotes time. We can alter the magnitude by multiplying $h(\omega t + \theta)$ by a positive constant $|A|$, called the *amplitude*, to get $|A|h(\omega t + \theta)$. More generally, we can combine the amplitude and the phase, writing

$$|A|h(\omega t + \theta) = |A|e^{i\theta}e^{i\omega t} = Ae^{i\omega t},$$

where $A$ is the complex amplitude $A = |A|e^{i\theta}$. Many of the functions encountered in signal processing can be modeled as linear combinations of such complex exponential functions or *sinusoids*, as they are often called.

## 2.8 Complex Exponential Signal Models

In a later chapter we consider signal models $f(x)$ that are sums of trigonometric functions;

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{L} \Big( a_k \cos(\omega_k x) + b_k \sin(\omega_k x) \Big), \tag{2.18}$$

where the $\omega_k$ are known, but the $a_k$ and $b_k$ are not. Now that we see how to convert sines and cosines to complex exponential functions, using

$$\cos(\omega_k x) = \frac{1}{2} \Big( \exp(i\omega_k x) + \exp(-i\omega_k x) \Big) \tag{2.19}$$

and

$$\sin(\omega_k x) = \frac{1}{2i} \Big( \exp(i\omega_k x) - \exp(-i\omega_k x) \Big), \tag{2.20}$$

we can write $f(x)$ as

$$f(x) = \sum_{m=-L}^{L} c_m \exp(i\omega_m x), \tag{2.21}$$

where $c_0 = \frac{1}{2}a_0$,

$$c_k = \frac{1}{2}(a_k - ib_k), \tag{2.22}$$

and

$$c_{-k} = \frac{1}{2}(a_k + ib_k), \tag{2.23}$$

for $k = 1, ..., L$. The complex notation is more commonly used in signal processing. Note that if the original coefficients $a_k$ and $b_k$ are real numbers, then $c_{-m} = \overline{c_m}$.

# Chapter 3

# Differential Equations (Chapters 2,3)

## 3.1 Second-Order Linear ODE

The most general form of the second-order linear homogeneous ordinary differential equation with variable coefficients is

$$R(x)y''(x) + P(x)y'(x) + Q(x)y(x) = 0. \tag{3.1}$$

Many differential equations of this type arise when we employ the technique of separating the variables to solve a partial differential equation. We shall consider several equivalent forms of Equation (3.1).

### 3.1.1 The Standard Form

Of course, dividing through by the function $R(x)$ and renaming the coefficient functions, we can also write Equation (3.1) in the *standard* form as

$$y''(x) + P(x)y'(x) + Q(x)y(x) = 0. \tag{3.2}$$

There are other equivalent forms of Equation (3.1).

### 3.1.2 The Sturm-Liouville Form

Let $S(x) = \exp(-F(x))$, where $F'(x) = (R'(x) - P(x))/R(x)$. Then we have

$$\frac{d}{dx}(S(x)R(x)) = S(x)P(x).$$

From Equation (3.1) we obtain

$$S(x)R(x)y''(x) + S(x)P(x)y'(x) + S(x)Q(x)y(x) = 0,$$

so that

$$\frac{d}{dx}(S(x)R(x)y'(x)) + S(x)Q(x)y(x) = 0,$$

which then has the form

$$\frac{d}{dx}(p(x)y'(x)) + g(x)y(x) = 0. \tag{3.3}$$

We shall be particularly interested in special cases having the form

$$\frac{d}{dx}(p(x)y'(x)) - w(x)q(x)y(x) + \lambda w(x)y(x) = 0, \tag{3.4}$$

where $w(x) > 0$ and $\lambda$ is a constant. Rewriting Equation (3.4) as

$$-\frac{1}{w(x)}\frac{d}{dx}(p(x)y'(x)) + q(x)y(x) = \lambda y(x), \tag{3.5}$$

we are reminded of eigenvector problems in linear algebra,

$$Ax = \lambda x, \tag{3.6}$$

where $A$ is a square matrix, $\lambda$ is an eigenvalue of $A$, and $x \neq 0$ is an associated eigenvector. What is now playing the role of $A$ is the *linear differential operator* $L$ that operates on a function $y$ to produce the function $Ly$ given by

$$(Ly)(x) = -\frac{1}{w(x)}\Big(\frac{d}{dx}(p(x)y'(x))\Big) + q(x)y(x). \tag{3.7}$$

If $y(x)$ satisfies the equation

$$Ly = \lambda y,$$

then $y(x)$ is said to be an *eigenfunction* of $L$, with associated eigenvalue $\lambda$.

### 3.1.3 The Normal Form

We start now with the differential equation as given by Equation (3.2). This differential equation can be written in the equivalent *normal* form

$$u''(x) + q(x)u(x) = 0, \tag{3.8}$$

where

$$y(x) = u(x)v(x),$$

$$v(x) = -\exp\Big(-\frac{1}{2}\int P dx\Big),$$

and

$$q(x) = Q(x) - \frac{1}{4}P(x)^2 - \frac{1}{2}P'(x).$$

One reason for wanting to put the differential equation into normal form is to relate the properties of its solutions to the properties of $q(x)$. For example, we are interested in the location of zeros of the solutions of Equation (3.8), as compared with the zeros of the solutions of

$$u''(x) + r(x)u(x) = 0. \tag{3.9}$$

In particular, we want to compare the spacing of zeros of solutions of Equation (3.8) to that of the known solutions of the equation

$$u''(x) + u(x) = 0.$$

If $q(x) < 0$, then any non-trivial solution of Equation (3.8) has at most one zero; think of the equation

$$u''(x) - u(x) = 0,$$

with solutions $u(x) = e^x$ and $u(x) = e^{-x}$. Therefore, when we study an equation in normal form, we shall always assume that $q(x) > 0$.

Determining important properties of the solutions of a differential equation without actually finding those solutions is called *qualitative analysis*. We shall have more to say about qualitative analysis later in these notes.

## 3.2 Recalling the Wave Equation

The one-dimensional wave equation is

$$\phi_{tt}(x,t) = c^2\phi_{xx}(x,t), \tag{3.10}$$

where $c > 0$ is the propagation speed. Separating variables, we seek a solution of the form $\phi(x,t) = f(t)y(x)$. Inserting this into Equation (3.10), we get

$$f''(t)y(x) = c^2 f(t)y''(x),$$

or

$$f''(t)/f(t) = c^2 y''(x)/y(x) = -\omega^2,$$

where $\omega > 0$ is the separation constant. We then have the separated differential equations

$$f''(t) + \omega^2 f(t) = 0, \tag{3.11}$$

and

$$y''(x) + \frac{\omega^2}{c^2} y(x) = 0. \tag{3.12}$$

Equation (3.12) can be written as an eigenvalue problem:

$$-y''(x) = \frac{\omega^2}{c^2} y(x) = \lambda y(x), \tag{3.13}$$

where, for the moment, the $\lambda$ is unrestricted.

The solutions to Equation (3.12) are

$$y(x) = \alpha \sin\left(\frac{\omega}{c} x\right) + \beta \cos\left(\frac{\omega}{c} x\right).$$

For each arbitrary $\omega$, the corresponding solutions of Equation (3.11) are

$$f(t) = \gamma \sin(\omega t) + \delta \cos(\omega t).$$

In the vibrating string problem, the string is fixed at both ends, $x = 0$ and $x = L$, so that

$$\phi(0, t) = \phi(L, t) = 0,$$

for all $t$. Therefore, we must have $y(0) = y(L) = 0$, so that the solutions must have the form

$$y(x) = A_m \sin\left(\frac{\omega_m}{c} x\right) = A_m \sin\left(\frac{\pi m}{L} x\right),$$

where $\omega_m = \frac{\pi c m}{L}$, for any positive integer $m$. Therefore, the boundary conditions limit the choices for the separation constant $\omega$, and thereby the choices for $\lambda$. In addition, if the string is not moving at time $t = 0$, then

$$f(t) = \delta \cos(\omega_m t).$$

We want to focus on Equation (3.12).

What we have just seen is that the boundary conditions $y(0) = y(L) = 0$ limit the possible values of $\lambda$ for which there can be solutions: we must have

$$\lambda = \lambda_m = \left(\frac{\omega_m}{c}\right)^2 = \left(\frac{\pi m}{L}\right)^2,$$

for some positive integer $m$. The corresponding solutions

$$y_m(x) = \sin\left(\frac{\pi m}{L} x\right)$$

are the *eigenfunctions*. This is analogous to the linear algebra case, in which $Ax = \lambda x$, with $x$ non-zero, only holds for special choices of $\lambda$.

In the vibrating string problem, we typically have the condition $\phi(x,0) = h(x)$, where $h(x)$ is some function that describes the initial position of the string. The problem that remains is to find a linear combination of the eigenfunctions that satisfies this additional initial condition. Therefore, we need to find coefficients $A_m$ so that

$$h(x) = \sum_{m=1}^{\infty} A_m \sin\Big(\frac{\pi m}{L}x\Big). \tag{3.14}$$

This again reminds us of finding a basis for a finite-dimensional vector space consisting of eigenvectors of a given matrix. As we discuss in the chapter **Some Linear Algebra**, this can be done only for a certain special kind of matrices, the *normal* matrices, which includes the *self-adjoint* ones. The property of a matrix being self-adjoint is one that we shall usefully extend later to linear differential operators.

Orthogonality of the eigenfunctions $y_m(x)$ will help us find the coefficients $A_m$. In this case we know the $y_m(x)$ and can demonstrate their orthogonality directly, using trigonometric identities. But we can also demonstrate their orthogonality using only the fact that each $y_m$ solves the eigenvalue problem for $\lambda_m$; this sort of approach is what is done in qualitative analysis. We multiply the equation

$$y_m'' = -\lambda_m y_m$$

by $y_n$ and the equation

$$y_n'' = -\lambda_n y_n$$

by $y_m$ and subtract, to get

$$y_m'' y_n - y_n'' y_m = (\lambda_n - \lambda_m)(y_m y_n).$$

Using

$$y_m'' y_n - y_n'' y_m = (y_n y_m' - y_m y_n')',$$

and integrating, we get

$$\begin{aligned} 0 &= y_n(L)y_m'(L) - y_m(L)y_n'(L) - y_n(0)y_m'(0) + y_m(0)y_n'(0) \\ &= (\lambda_n - \lambda_m)\int_0^L y_m(x)y_n(x)dx, \end{aligned}$$

so that

$$\int_0^L y_m(x)y_n(x)dx = 0,$$

for $m \neq n$. Using this orthogonality of the $y_m(x)$, we can easily find the coefficients $A_m$.

## 3.3 A Brief Discussion of Some Linear Algebra

In this section we review briefly some notions from linear algebra that we shall need shortly. For more detail, see the chapter *Some Linear Algebra*.

Suppose that $V$ is an $N$-dimensional complex vector space on which there is defined an *inner product*, with the inner product of members $a$ and $b$ denoted $\langle a, b\rangle$. For example, we take $V = \mathbb{C}^N$, the space of all $N$-dimensional complex column vectors, with the inner product defined by the complex dot product

$$a \cdot b = b^\dagger a = \sum_{n=1}^{N} a_n \overline{b_n}, \tag{3.15}$$

where $b^\dagger$ is the row vector with entries $\overline{b_n}$.

For any linear operator $T : V \to V$ we define the *adjoint* of $T$ to be the linear operator $T^*$ satisfying $\langle Ta, b\rangle = \langle a, T^*b\rangle$, for all $a$ and $b$ in $V$. A word of warning: If we change the inner product, the adjoint changes. We consider two examples.

- **Example 1:** Let $A$ be any $N$ by $N$ complex matrix, $V = \mathbb{C}^N$, and define the linear operator $T$ on $V$ to be multiplication on the left by $A$; that is,
$$Tx = Ax,$$
for any vector $x$ in $V$. If the inner product on $V$ is the usual one coming from the dot product, as in Equation (3.15), then $T^*$ is the operator defined by
$$T^*x = A^\dagger x,$$
where $A^\dagger$ is the conjugate transpose of the matrix $A$.

- **Example 2:** If, on the other hand, we define an inner product on $V = \mathbb{C}^N$ by
$$\langle a, b\rangle = b^\dagger Q a,$$
where $Q$ is a positive-definite Hermitian matrix, then $T^*$ is the linear operator defined by multiplication on the left by the matrix $Q^{-1}A^\dagger Q$.

**Definition 3.1** *Given $V$ and the inner product, we say that a linear operator $T$ on $V$ is self-adjoint if $T^* = T$.*

For Example 1, $T$ is self-adjoint if the associated matrix $A$ is Hermitian, that is, $A^\dagger = A$. For Example 2, $T$ is self-adjoint if the associated matrix $A$ satisfies $QA = A^\dagger Q$.

A non-zero vector $u$ in $V$ is said to be an eigenvector of $T$ if there is a constant $\lambda$ such that $Tu = \lambda u$; then $\lambda$ is called the eigenvalue of $T$ associated with the eigenvector $u$. We have the following important results concerning self-adjoint linear operators.

**Theorem 3.1** *If $T$ is self-adjoint on the inner product space $V$, then all its eigenvalues are real numbers.*

**Proof:** By the defining properties of an inner product, we have

$$\langle \lambda x, y \rangle = \lambda \langle x, y \rangle,$$

and

$$\langle x, \lambda y \rangle = \overline{\lambda} \langle x, y \rangle.$$

Since $T = T^*$, we have

$$\lambda \langle u, u \rangle = \langle Tu, u \rangle = \langle u, Tu \rangle = \langle u, \lambda u \rangle = \overline{\lambda} \langle u, u \rangle.$$

Therefore, $\lambda = \overline{\lambda}$. ∎

**Theorem 3.2** *If $\lambda_m \neq \lambda_n$ are two eigenvalues of the self-adjoint linear operator $T$ associated with eigenvectors $u^m$ and $u^n$, respectively, then $\langle u^m, u^n \rangle = 0$.*

**Proof:** We have

$$\lambda_m \langle u^m, u^n \rangle = \langle Tu^m, u^n \rangle = \langle u^m, Tu^n \rangle = \langle u^m, \lambda_n u^n \rangle = \lambda_n \langle u^m, u^n \rangle.$$

∎

## 3.4 Preview of Coming Attractions

We have seen that the differential equation (3.1) can be reformulated in several equivalent forms. The equation

$$y''(x) + y(x) = \frac{d}{dx}(y'(x)) + y(x) = 0$$

is simultaneously in the standard form, the Sturm-Liouville form, and the normal form. As we shall see, considering its normal form will help us uncover properties of solutions of Equation (3.8) for other functions $q(x)$. Under certain boundary conditions, the linear differential operator $L$ given by Equation (3.7) will be *self-adjoint*, its eigenvalues will then be real numbers, and solutions to the eigenfunction problem will enjoy orthogonality properties similar to those we just presented for the $y_m(x)$ solving $y''(x) + y(x) = 0$.

Some of our discussion of these subjects in later chapters is taken from the book by Simmons [42]. Another good source that combines the mathematics with the history of the subject is the book by Gonzalez-Velasco [19].

# Chapter 4

# Extra Credit Problems (Chapters 2,3)

## 4.1 The Problems

- **Chemical Reaction:** Suppose that two chemical substances in solution react together to form a compound. If the reaction occurs by the collision and interaction of the molecules of the substances, we expect the rate of formation of the compound to be proportional to the number of collisions per unit of time, which in turn is jointly proportional to the amounts of the substances that are untransformed. A chemical reaction that proceeds in this manner is called a *second-order reaction*, and this law of reaction is often referred to as the *law of mass action*. Consider a second-order reaction in which $x$ grams of the compound contain $ax$ grams of the first substance and $bx$ grams of the second, where $a + b = 1$. If there are $aA$ grams of the first substance and $bB$ grams of the second present initially, and if $x = 0$ when $t = 0$, find $x$ as a function of $t$. ([42], p. 18)

- **Retarded Fall:** If we assume that air exerts a resisting force proportional to the velocity of a falling body, then the differential equation of the motion is
$$\frac{d^2y}{dt^2} = g - c\frac{dy}{dt},$$
where $c > 0$ is some constant. If the velocity $v = \frac{dy}{dt}$ is zero when $t = 0$, find the limiting or (*terminal*) velocity as $t \to +\infty$. If the retarding force is proportional to the *square* of the velocity, then the differential equation becomes
$$\frac{d^2y}{dt^2} = g - c(\frac{dy}{dt})^2.$$

Find the terminal velocity in this case. ([42], pp. 20, 24.)

- **Escape Velocity:** The force that gravity exerts on a body of mass $m$ at the surface of the earth is $mg$. In space, however, Newton's law of gravitation asserts that this force varies inversely as a square of the distance to the earth's center. If a projectile fired upward from the surface is to keep traveling indefinitely, show that its initial velocity must be at least $\sqrt{2gR}$, where $R$ is the radius of the earth (about 4000 miles). This *escape velocity* is approximately 7 miles/second or about 25,000 miles/hour. Hint: If $x$ is the distance from the center of the earth to the projectile and $v = \frac{dx}{dt}$ is its velocity, then

$$\frac{d^2x}{dt^2} = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = v\frac{dv}{dx}.$$

  ([42], p. 24)

  Another way to view this problem is to consider an object falling to earth from space. Calculate its velocity upon impact, as a function of the distance to the center of the earth at the beginning of its fall, neglecting all but gravity. Then calculate the upper limit of the impact velocity as the distance goes to infinity.

- **The Snowplow Problem:** It began snowing on a certain morning and the snow continued to fall steadily throughout the day. At noon a snowplow started to clear a road, at a constant rate, in terms of the volume of snow removed per hour. The snowplow cleared 2 miles by 2 p.m. and 1 more mile by 4 p.m. When did it start snowing? ([42], p. 31)

- **Torricelli's Law:** According to Torricelli's Law, water in an open tank will flow out through a small hole in the bottom with the speed it would acquire in falling freely from the water level to the hole. A hemispherical bowl of radius $R$ is initially filled with water, and a small circular hole of radius $r$ is punched in the bottom at time $t = 0$. How long does it take for the bowl to empty itself? ([42], p. 32)

- **The Coffee and Cream Problem:** The President and the Prime Minister order coffee and receive cups of equal temperature at the same time. The President adds a small amount of cool cream immediately, but does not drink his coffee until 10 minutes later. The Prime Minister waits ten minutes and then adds the same amount of cool cream and begins to drink. Who drinks the hotter coffee? ([42], p. 33)

- **The Two Tanks Problem:** A tank contains 50 gallons of brine in which 25 pounds of salt are dissolved. Beginning at time $t = 0$, water

runs into this tank at the rate of 2 gallons/minute, and the mixture flows out at the same rate through a second tank initially containing 50 gallons of pure water. When will the second tank contain the greatest amount of salt? ([42], p. 62)

- **Torricelli Again:** A cylindrical tank is filled with water to a height of $D$ feet. At height $h < D$ feet a small hole is drilled into the side of the tank. According to Torricelli's Law, the horizontal velocity with which the water spurts from the side of the tank is $v = \sqrt{2g(D-h)}$. What is the distance $d$ from the base of the tank to where the water hits the ground? For fixed $D$, what are the possible values of $d$ as $h$ varies? Given $D$ and $d$, can we find $h$? This last question is an example of an *inverse problem* ([24], pp. 26-27). We shall consider more inverse problems below.

- **The Well Problem:** A rock is dropped into a well in which the unknown water level is $d$ feet below the top of the well. If we measure the time lapse from the dropping of the rock until the hearing of the splash, can we use this to determine $d$? ([24], p. 40)

- **The Pool Table Problem:** Suppose our 'pool table' is the unit square $\{(x,y)|0 \leq x \leq 1, 0 \leq y \leq 1\}$. Suppose the cue ball is at $(x_1, y_1)$ and the target ball is at $(x_2, y_2)$. In how many ways can we hit the target ball with the cue ball using a 'bank shot', in which the cue ball rebounds off the side of the table once before striking the target ball? Now for a harder problem: there is no pool table now. The cue ball is launched from the origin into the first quadrant at an angle $\theta > 0$ with the positive $x$-axis. It bounces off a straight line and returns to the positive $x$-axis at the point $r(\theta)$, making an angle $\psi(\theta) > 0$. Can we determine the equation of the straight line from this information? What if we do not know $r(\theta)$? ([24], pp. 41-44)

- **Torricelli, Yet Again!:** A container is formed by revolving the curve $x = f(y)$ around the (vertical) $y$-axis. The container is filled to a height of $y$ and the water is allowed to run out through a hole of cross-sectional area $a$ in the bottom. The time it takes to drain is $T(y)$. How does the drain-time function $T$ depend on the shape function $f$? Can we determine $f$ if we know $T$? How could we approximate $f$ from the values $T(y_n)$, $n = 1, ..., N$? ([24], pp. 59–66)

- **Mixing Problems:** Let $q(t)$ denote the quantity of a pollutant in a container at time $t$. Then the rate at which $q(t)$ changes with time is the difference between the rate at which the pollutant enters the container and the rate at which it is removed. Suppose the container has volume $V$, water with a concentration $a$ of pollutant enters the container at a rate $r$ and the well-stirred mixture leaves the container

again at the rate $r$. Write the differential equation governing the behavior of the function $q(t)$. Suppose now that $q(0) = 0$ and $a$ and $r$ are unknown. Show that they can be determined from two measurements of $q(t)$. If, instead, $q(0)$ is also unknown, show that all three parameters can be determined from three measurements of $q(t)$. ([24], pp. 92–96)

- **Frictionless Sliding:** A half-line begins at the origin and continues into the fourth quadrant, making an angle of $\alpha$ with the positive $x$-axis. A particle descends from the origin along this half-line, under the influence of gravity and without resistance. Let $C$ be a circle contained in the third and fourth quadrants, passing through the origin and tangent to the $x$-axis. Let $T$ be the time required for the particle to reach the point where the half-line again intersects $C$. Show that $T$ is independent of $\alpha$ and depends only on the radius of the circle. ([24], pp. 96–102)

# Chapter 5

# Qualitative Analysis of ODEs (Chapter 2,3)

We are interested in second-order linear differential equations with possibly varying coefficients, as given in Equation (3.1), which we can also write as

$$y'' + P(x)y' + Q(x)y = 0. \tag{5.1}$$

Although we can find explicit solutions of Equation (5.1) in special cases, such as

$$y'' + y = 0, \tag{5.2}$$

generally, we will not be able to do this. Instead, we can try to answer certain questions about the behavior of the solution, without actually finding the solution; such an approach is called *qualitative analysis*. The discussion here is based on that in Simmons [42].

## 5.1 Existence and Uniqueness

We begin with the fundamental existence and uniqueness theorem for solutions of Equation (5.1).

**Theorem 5.1** *Let $P(x)$ and $Q(x)$ be continuous functions on the interval $[a, b]$. If $x_0$ is any point in $[a, b]$ and $y_0$ and $y_0'$ any real numbers, then there is a unique solution of Equation (5.1) satisfying the conditions $y(x_0) = y_0$ and $y'(x_0) = y_0'$.*

The proof of this theorem is somewhat lengthy and we shall omit the proof here.

## 5.2 A Simple Example

We know that the solution to Equation (5.2) satisfying $y(0) = 0$, and $y'(0) = 1$ is $y(x) = \sin x$; with $y(0) = 1$ and $y'(0) = 0$, the solution is $y(x) = \cos x$. But, suppose that we did not know these solutions; what could we find out without solving for them?

Suppose that $y(x) = s(x)$ satisfies Equation (5.2), with $s(0) = 0$, $s(\pi) = 0$, and $s'(0) = 1$. As the graph of $s(x)$ leaves the point $(0, 0)$ with $x$ increasing, the slope is initially $s'(0) = 1$, so the graph climbs above the $x$-axis. But since $y''(x) = -y(x)$, the second derivative is negative for $y(x) > 0$, and becomes increasingly so as $y(x)$ climbs higher; therefore, the derivative is decreasing from $s'(0) = 1$, eventually equaling zero, at say $x = m$, and continuing to become negative. The function $s(x)$ will be zero again at $x = \pi$, and, by symmetry, we have $m = \frac{\pi}{2}$.

Now let $y(x) = c(x)$ solve Equation (5.2), but with $c(0) = 1$, and $c'(0) = 0$. Since $y(x) = s(x)$ satisfies Equation (5.2), so does $y(x) = s'(x)$, with $s'(0) = 1$ and $s''(0) = 0$. Therefore, $c(x) = s'(x)$. Since the derivative of the function $s(x)^2 + c(x)^2$ is zero, this function must be equal to one for all $x$. In the section that follows, we shall investigate the zeros of solutions.

## 5.3 The Sturm Separation Theorem

**Theorem 5.2** *Let $y_1(x)$ and $y_2(x)$ be linearly independent solutions of Equation (5.1). Then their zeros are distinct and occur alternately.*

**Proof:** We know that solutions $y_1(x)$ and $y_2(x)$ are linearly independent if and only if the Wronskian

$$W(x, y_1, y_2) = y_1(x)y_2'(x) - y_2(x)y_1'(x),$$

is different from zero for all $x$ in the interval $[a, b]$. Therefore, when the two functions are linearly independent, the function $W(x, y_1, y_2)$ must have constant sign on the interval $[a, b]$. Therefore, the two functions $y_1(x)$ and $y_2(x)$ have no common zero. Suppose that $y_2(x_1) = y_2(x_2) = 0$, with $x_1 < x_2$ successive zeros of $y_2(x)$. Suppose, in addition, that $y_2(x) > 0$ in the interval $(x_1, x_2)$. Therefore, we have $y_2'(x_1) > 0$ and $y_2'(x_2) < 0$. It follows that $y_1(x_1)$ and $y_1(x_2)$ have opposite signs, and there must be a zero between $x_1$ and $x_2$. ∎

## 5.4 From Standard to Normal Form

Equation (5.1) is called the *standard form* of the differential equation. To put the equation into *normal form*, by which we mean an equation of the

form

$$u''(x) + q(x)u(x) = 0, \tag{5.3}$$

we write $y(x) = u(x)v(x)$. Inserting this product into Equation (5.1), we obtain

$$vu'' + (2v' + Pv)u' + (v'' + Pv' + Qv)u = 0.$$

With

$$v = \exp\Big(-\frac{1}{2}\int P dx\Big),$$

the coefficient of $u'$ becomes zero. Now we set

$$q(x) = Q(x) - \frac{1}{4}P(x)^2 - \frac{1}{2}P'(x),$$

to get

$$u''(x) + q(x)u(x) = 0.$$

## 5.5 On the Zeros of Solutions

We assume now that $u(x)$ is a non-trivial solution of Equation (5.3). As we shall show shortly, if $q(x) < 0$ and $u(x)$ satisfies Equation (5.3), then $u(x)$ has at most one zero; for example, the equation

$$u''(x) - u(x) = 0$$

has $e^x$ and $e^{-x}$ for solutions. Since we are interested in oscillatory solutions, we restrict $q(x)$ to be (eventually) positive. With $q(x) > 0$ and

$$\int_1^\infty q(x)dx = \infty,$$

the solution $u(x)$ will have infinitely many zeros, but only finitely many on any bounded interval.

**Theorem 5.3** *If $q(x) < 0$ for all $x$, then $u(x)$ has at most one zero.*

**Proof:** Let $u(x_0) = 0$. Since $u(x)$ is not identically zero, we must have $u'(x_0) \neq 0$, by Theorem 5.1. Therefore, assume that $u'(x) > 0$ for $x$ in the interval $[x_0, x_0 + \epsilon]$, where $\epsilon$ is some positive number. Since $u''(x) = -q(x)u(x)$, we know that $u''(x) > 0$ also, for $x$ in the interval $[x_0, x_0 + \epsilon]$. So the slope of $u(x)$ is increasing to the right of $x_0$, and so there can be no zero of $u(x)$ to the right of $x_0$. A similar argument shows that there can be no zeros of $u(x)$ to the left of $x_0$. ∎

**Theorem 5.4** *If $q(x) > 0$ for all $x > 0$ and $\int_1^\infty q(x)dx = \infty$, then $u(x)$ has infinitely many positive zeros.*

**Proof:** Assume, to the contrary, that $u(x)$ has only finitely many positive zeros, and that there are no positive zeros to the right of the positive number $x_0$. Assume also that $u(x_0) > 0$. From $u''(x) = -q(x)u(x)$ we know that the slope of $u(x)$ is decreasing to the right of $x_0$, so long as $u(x)$ remains above the $x$-axis. If the slope ever becomes negative, the graph of $u(x)$ will continue to drop at an ever increasing rate and will have to cross the $x$-axis at some point to the right of $x_0$. Therefore, to avoid having a root beyond $x_0$, the slope must remain positive. We prove the theorem by showing that the slope eventually becomes negative.

Let $v(x) = -u'(x)/u(x)$, for $x \geq x_0$. Then $v'(x) = q(x) + v^2(x)$, and

$$v(x) - v(x_0) = \int_{x_0}^{x} q(x)dx + \int_{x_0}^{x} v^2(x)dx.$$

Since

$$\int_1^\infty q(x)dx = \infty,$$

we see that $v(x)$ must eventually become positive, as $x \to \infty$. Therefore, $u'(x)$ and $u(x)$ eventually have opposite signs. Since we are assuming that $u(x)$ remains positive to the right of $x_0$, it follows that $u'(x)$ becomes negative somewhere to the right of $x_0$. ∎

## 5.6 Sturm Comparison Theorem

Solutions to

$$y'' + 4y = 0$$

oscillate faster than solutions of Equation (5.2). This leads to the Sturm Comparison Theorem.

**Theorem 5.5** *Let $y''+q(x)y = 0$ and $z''+r(x)z = 0$, with $0 < r(x) < q(x)$, for all $x$. Then between any two zeros of $z(x)$ is a zero of $y(x)$.*

### 5.6.1 Bessel's Equation

Bessel's Equation is

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0. \tag{5.4}$$

In normal form, it becomes

$$u'' + \Big(1 + \frac{1 - 4\nu^2}{4x^2}\Big)u = 0. \tag{5.5}$$

Information about the zeros of solutions of Bessel's Equation can be obtained by using Sturm's Comparison Theorem and comparing with solutions of Equation (5.2).

## 5.7 Analysis of $y'' + q(x)y = 0$

Using the Sturm Comparison Theorem, we can prove the following lemma.

**Lemma 5.1** *Let $y'' + q(x)y = 0$, and $z'' + r(x)z = 0$, with $0 < r(x) < q(x)$. Let $y(b_0) = z(b_0) = 0$ and $z(b_j) = 0$, and $b_j < b_{j+1}$, for $j = 1, 2, ....$ Then, $y$ has at least as many zeros as $z$ in $[b_0, b_n]$. If $y(a_j) = 0$, for $b_0 < a_1 < a_2 < ...$, then $a_n < b_n$.*

**Lemma 5.2** *Suppose that $0 < m^2 < q(x) < M^2$ on $[a, b]$, and $y(x)$ solves $y'' + q(x)y = 0$ on $[a, b]$. If $x_1$ and $x_2$ are successive zeros of $y(x)$ then*

$$\frac{\pi}{M} < x_2 - x_1 < \frac{\pi}{m}.$$

*If $y(a) = y(b) = 0$ and $y(x) = 0$ for $n - 1$ other points in $(a, b)$, then*

$$\frac{m(b-a)}{\pi} < n < \frac{M(b-a)}{\pi}.$$

**Lemma 5.3** *Let $y_\lambda$ solve*

$$y'' + \lambda q(x)y = 0,$$

*with $y_\lambda(a) = 0$, and $y'_\lambda(a) = 1$. Then, there exist $\lambda_1 < \lambda_2 < ...$, converging to $+\infty$, such that $y_\lambda(b) = 0$ if and only if $\lambda = \lambda_n$, for some $n$. The solution $y_{\lambda_n}$ has exactly $n - 1$ roots in $(a, b)$.*

## 5.8 Toward the 20th Century

The goal in qualitative analysis is to learn something about the solutions of a differential equation by examining its form, rather than actually finding solutions. We do this by exploiting similarities between one equation and another; in other words, we study classes of differential equations all at once. This is what we did earlier, when we studied problems of the Sturm-Liouville type. The simplest boundary-value problem,

$$y''(x) + \lambda y(x) = 0,$$

with $y(0) = y(L) = 0$, can be solved explicitly. Its eigenfunction solutions are $y_m(x) = \sin(m\pi x/L)$, which are orthogonal over the interval $[0, L]$, with respect to the inner product defined by

$$\langle f, g \rangle = \int_0^L f(x)g(x)dx.$$

This suggests that other differential equations that can be written in Sturm-Liouville form may have eigenfunction solutions that are also orthogonal, with respect to some appropriate inner product. As we have seen, this program works out beautifully. What is happening here is a transition from classical applied mathematics, with its emphasis on particular problems and equations, to a more modern, 20th century style mathematics, with an emphasis on families of functions or even more abstract *inner-product spaces*, Hilbert spaces, Banach spaces, and so on.

# Chapter 6

# The Trans-Atlantic Cable (Chapters 4,12)

## 6.1 Introduction

In 1815, at the end of the war with England, the US was a developing country, with most people living on small farms, eating whatever they could grow themselves. Only those living near navigable water could market their crops. Poor transportation and communication kept them isolated. By 1848, at the end of the next war, this time with Mexico, things were different. The US was a transcontinental power, integrated by railroads, telegraph, steamboats, the Erie Canal, and innovations in mass production and agriculture. In 1828, the newly elected President, Andrew Jackson, arrived in Washington by horse-drawn carriage; he left in 1837 by train. The most revolutionary change was in communication, where the recent advances in understanding electromagnetism produced the telegraph. It wasn't long before efforts began to lay a telegraph cable under the Atlantic Ocean, even though some wondered what England and the US could possibly have to say to one another.

The laying of the trans-Atlantic cable was, in many ways, the 19th century equivalent of landing a man on the moon, involving, as it did, considerable expense, too frequent failure, and a level of precision in engineering design and manufacturing never before attempted. From a scientific perspective, it was probably more difficult, given that the study of electromagnetism was in its infancy at the time.

Early on, Faraday and others worried that sending a message across a vast distance would take a long time, but they reasoned, incorrectly, that this would be similar to filling a very long hose with water. What they did not realize initially was that, as William Thomson was to discover,

the transmission of a pulse through an undersea cable was described more by a heat equation than a wave equation. This meant that a signal that started out as a sharp pulse would be spread out as time went on, making communication extremely slow. The problem was the increased capacitance with the ground.

Somewhat later, Oliver Heaviside realized that, when all four of the basic elements of the electrical circuit, the inductance, the resistance, the conductance to the ground and the capacitance to the ground, were considered together, it might be possible to adjust these parameters, in particular, to increase the inductance, so as to produce undistorted signals. Heaviside died in poverty, but his ideas eventually were adopted.

In 1859 Queen Victoria sent President Buchanan a 99 word greeting using an early version of the cable, but the message took over sixteen hours to be received. By 1866 one could transmit eight words a minute along a cable that stretched from Ireland to Newfoundland, at a cost of about 1500 dollars per word in today's money. With improvements in insulation, using gutta percha, a gum from a tropical tree also used to make golf balls, and the development of magnetic alloys that increased the inductance of the cable, messages could be sent faster and more cheaply.

In this chapter we survey the development of the mathematics of the problem. We focus, in particular, on the partial differential equations that were used to describe the transmission problem. What we give here is a brief glimpse; more detailed discussion of this problem is found in the books by Körner [32], Gonzalez-Velasco [19], and Wylie [47].

## 6.2 The Electrical Circuit ODE

We begin with the ordinary differential equation that describes the horizontal motion of a block of wood attached to a spring. We let $x(t)$ be the position of the block relative to the equilibrium position $x = 0$, with $x(0)$ and $x'(0)$ denoting the initial position and velocity of the block. When an external force $f(t)$ is imposed, a portion of this force is devoted to overcoming the inertia of the block, a portion to compressing or stretching the spring, and the remaining portion to resisting friction. Therefore, the differential equation describing the motion is

$$mx''(t) + ax'(t) + kx(t) = f(t), \tag{6.1}$$

where $m$ is the mass of the block, $a$ the coefficient of friction, and $k$ the spring constant.

The charge $Q(t)$ deposited on a capacitor in an electrical circuit due to an imposed electromotive force $E(t)$ is similarly described by the ordinary

differential equation

$$LQ''(t) + RQ'(t) + \frac{1}{C}Q(t) = E(t). \tag{6.2}$$

The first term, containing the inductance coefficient $L$, describes the portion of the force $E(t)$ devoted to overcoming the effect of a change in the current $I(t) = Q'(t)$; here $L$ is analogous to the mass $m$. The second term, containing the resistance coefficient $R$, describes that portion of the force $E(t)$ needed to overcome resistance to the current $I(t)$; now $R$ is analogous to the friction coefficient $a$. Finally, the third term, containing the reciprocal of the capacitance $C$, describes the portion of $E(t)$ used to store charge on the capacitor; now $\frac{1}{C}$ is analogous to $k$, the spring constant.

## 6.3 The Telegraph Equation

The objective here is to describe the behavior of $u(x,t)$, the voltage at location $x$ along the cable, at time $t$. In the beginning, it was believed that the partial differential equation describing the voltage would be the wave equation

$$u_{xx} = \alpha^2 u_{tt}.$$

If this were the case, an initial pulse

$$E(t) = H(t) - H(t-T)$$

would move along the cable undistorted; here $H(t)$ is the Heaviside function that is zero for $t < 0$ and one for $t \geq 0$. Thomson (later Sir William Thomson, and even later, Lord Kelvin) thought otherwise.

Thomson argued that there would be a voltage drop over an interval $[x, x+\Delta x]$ due to resistance to the current $i(x,t)$ passing through the cable, so that

$$u(x+\Delta x, t) - u(x,t) = -Ri(x,t)\Delta x,$$

and so

$$\frac{\partial u}{\partial x} = -Ri.$$

He also argued that there would be capacitance to the ground, made more significant under water. Since the apparent change in current due to the changing voltage across the capacitor is

$$i(x+\Delta x, t) - i(x,t) = -Cu_t(x,t)\Delta x,$$

we have

$$\frac{\partial i}{\partial x} = -C\frac{\partial u}{\partial t}.$$

Eliminating the $i(x,t)$, we can write

$$u_{xx}(x,t) = CRu_t(x,t), \tag{6.3}$$

which is the heat equation, not the wave equation.

## 6.4 Consequences of Thomson's Model

To see what Thomson's model predicts, we consider the following problem. Suppose we have a semi-infinite cable, that the voltage is $u(x,t)$ for $x \geq 0$, and $t \geq 0$, and that $u(0,t) = E(t)$. Let $U(x,s)$ be the Laplace transform of $u(x,t)$, viewed as a function of $t$. Then, from Thomson's model we have

$$U(x,s) = \mathcal{L}(E)(s)e^{-\sqrt{CRs}x},$$

where $\mathcal{L}(E)(s)$ denotes the Laplace transform of $E(t)$. Since $U(x,s)$ is the product of two functions of $s$, the convolution theorem applies. But first, it is helpful to find out which function has for its Laplace transform the function $e^{-\alpha x\sqrt{s}}$. The answer comes from the following fact: the function

$$be^{-b^2/4t}/2\sqrt{\pi}t^{3/2}$$

has for its Laplace transform the function $e^{-b\sqrt{s}}$. Therefore, we can write

$$u(x,t) = \frac{\sqrt{CR}x}{2\sqrt{\pi}} \int_0^t E(t-\tau)\frac{e^{-CRx^2/4\tau}}{\tau\sqrt{\tau}}d\tau.$$

Now we consider two special cases.

### 6.4.1 Special Case 1: $E(t) = H(t)$

Suppose now that $E(t) = H(t)$, the Heaviside function. Using the substitution

$$z = CRx^2/4\tau,$$

we find that

$$u(x,t) = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{CR}x/2\sqrt{\pi}} e^{-z^2}dz. \tag{6.4}$$

The function

$$\text{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r e^{-z^2}dz$$

is the well known *error function*, so we can write

$$u(x,t) = 1 - \text{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t}}\Big). \tag{6.5}$$

### 6.4.2 Special Case 2: $E(t) = H(t) - H(t-T)$

Now suppose that $E(t)$ is the pulse $H(t) - H(t-T)$. Using the results from the previous subsection, we find that, for $t > T$,

$$u(x,t) = \operatorname{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t-T}}\Big) - \operatorname{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t}}\Big). \tag{6.6}$$

For fixed $x$, $u(x,t)$ is proportional to the area under the function $e^{-z^2}$, over an interval that, as time goes on, moves steadily to the left and decreases in length. For small $t$ the interval involves only large $z$, where the function $e^{-z^2}$ is nearly zero and the integral is nearly zero. As $t$ increases, the interval of integration moves to the left, so that the integrand grows larger, but the length of the interval grows smaller. The net effect is that the voltage at $x$ increases gradually over time, and then decreases gradually; the sharp initial pulse is smoothed out in time.

## 6.5 Heaviside to the Rescue

It seemed that Thomson had solved the mathematical problem and discovered why the behavior was not wave-like. Since it is not really possible to reduce the resistance along the cable, and capacitance to the ground would probably remain a serious issue, particularly under water, it appeared that little could be done to improve the situation. But Heaviside had a solution.

Heaviside argued that Thomson had ignored two other circuit components, the leakage of current to the ground, and the self-inductance of the cable. He revised Thomson's equations, obtaining

$$u_x = -Li_t - Ri,$$

and

$$i_x = -Cu_t - Gu,$$

where $L$ is the inductance and $G$ is the coefficient of leakage of current to the ground. The partial differential equation governing $u(x,t)$ now becomes

$$u_{xx} = LCu_{tt} + (LG + RC)u_t + RGu, \tag{6.7}$$

which is the formulation used by Kirchhoff. As Körner remarks, never before had so much money been riding on the solution of one partial differential equation.

### 6.5.1 A Special Case: $G = 0$

If we take $G = 0$, thereby assuming that no current passes into the ground, the partial differential equation becomes

$$u_{xx} = LCu_{tt} + RCu_t, \tag{6.8}$$

or

$$\frac{1}{CL}u_{xx} = u_{tt} + \frac{R}{L}u_t. \tag{6.9}$$

If $R/L$ could be made small, we would have a wave equation again, but with a propagation speed of $1/\sqrt{CL}$. This suggested to Heaviside that one way to obtain undistorted signaling would be to increase $L$, since we cannot realistically hope to change $R$. He argued for years for the use of cables with higher inductance, which eventually became the practice, helped along by the invention of new materials, such as magnetic alloys, that could be incorporated into the cables.

### 6.5.2 Another Special Case

Assume now that $E(t)$ is the pulse. Applying the Laplace transform method described earlier to Equation (6.7), we obtain

$$U_{xx}(x,s) = (Cs+G)(Ls+R)U(x,s) = \lambda^2 U(x,s),$$

from which we get

$$U(x,s) = A(s)e^{\lambda x} + \Big(\frac{1}{s}(1-e^{-Ts}) - A(s)\Big)e^{-\lambda x}.$$

If it happens that $GL = CR$, we can solve easily for $\lambda$:

$$\lambda = \sqrt{CL}s + \sqrt{GR}.$$

Then we have

$$U(x,s) = e^{-\sqrt{GR}x}\frac{1}{s}(1-e^{-Ts})e^{-\sqrt{CL}xs},$$

so that

$$u(x,t) = e^{-\sqrt{GR}x}\Big(H(t-x\sqrt{CL}) - H(t-T-x\sqrt{CL})\Big). \tag{6.10}$$

This tells us that we have an undistorted pulse that arrives at the point $x$ at the time $t = x\sqrt{CL}$.

In order to have $GL = CR$, we need $L = CR/G$. Since $C$ and $R$ are more or less fixed, and $G$ is typically reduced by insulation, $L$ will need to be large. Again, this argues for increasing the inductance in the cable.

# Chapter 7

# The Laplace Transform and the Ozone Layer (Chapter 4)

In farfield propagation problems, we often find the measured data to be related to the desired object function by a Fourier transformation. The image reconstruction problem then becomes one of estimating a function from finitely many noisy values of its Fourier transform. In this chapter we consider an inverse problem involving the Laplace transform. The example is taken from Twomey's book [44].

## 7.1 The Laplace Transform

The Laplace transform of the function $f(x)$ defined for $0 \leq x < +\infty$ is the function

$$\mathcal{F}(s) = \int_0^{+\infty} f(x)e^{-sx}dx.$$

## 7.2 Scattering of Ultraviolet Radiation

The sun emits ultraviolet (UV) radiation that enters the Earth's atmosphere at an angle $\theta_0$ that depends on the sun's position, and with intensity $I(0)$. Let the $x$-axis be vertical, with $x = 0$ at the top of the atmosphere and $x$ increasing as we move down to the Earth's surface, at $x = X$. The intensity at $x$ is given by

$$I(x) = I(0)e^{-kx/\cos\theta_0}.$$

Within the ozone layer, the amount of UV radiation scattered in the direction $\theta$ is given by

$$S(\theta, \theta_0) I(0) e^{-kx/\cos\theta_0} \Delta p,$$

where $S(\theta, \theta_0)$ is a known parameter, and $\Delta p$ is the change in the pressure of the ozone within the infinitesimal layer $[x, x+\Delta x]$, and so is proportional to the concentration of ozone within that layer.

## 7.3 Measuring the Scattered Intensity

The radiation scattered at the angle $\theta$ then travels to the ground, a distance of $X - x$, weakened along the way, and reaches the ground with intensity

$$S(\theta, \theta_0) I(0) e^{-kx/\cos\theta_0} e^{-k(X-x)/\cos\theta} \Delta p.$$

The total scattered intensity at angle $\theta$ is then a superposition of the intensities due to scattering at each of the thin layers, and is then

$$S(\theta, \theta_0) I(0) e^{-kX/\cos\theta_0} \int_0^X e^{-x\beta} dp,$$

where

$$\beta = k[\frac{1}{\cos\theta_0} - \frac{1}{\cos\theta}].$$

This superposition of intensity can then be written as

$$S(\theta, \theta_0) I(0) e^{-kX/\cos\theta_0} \int_0^X e^{-x\beta} p'(x) dx.$$

## 7.4 The Laplace Transform Data

Using integration by parts, we get

$$\int_0^X e^{-x\beta} p'(x) dx = p(X) e^{-\beta X} - p(0) + \beta \int_0^X e^{-\beta x} p(x) dx.$$

Since $p(0) = 0$ and $p(X)$ can be measured, our data is then the Laplace transform value

$$\int_0^{+\infty} e^{-\beta x} p(x) dx;$$

note that we can replace the upper limit $X$ with $+\infty$ if we extend $p(x)$ as zero beyond $x = X$.

The variable $\beta$ depends on the two angles $\theta$ and $\theta_0$. We can alter $\theta$ as we measure and $\theta_0$ changes as the sun moves relative to the earth. In this way we get values of the Laplace transform of $p(x)$ for various values of $\beta$.

The problem then is to recover $p(x)$ from these values. Because the Laplace transform involves a smoothing of the function $p(x)$, recovering $p(x)$ from its Laplace transform is more ill-conditioned than is the Fourier transform inversion problem.

# Chapter 8

# The Finite Fourier Transform (Chapter 7)

## 8.1 Fourier Series

Suppose that $f(x)$ is a real or complex function defined for $0 \leq x \leq 2A$, with Fourier series representation

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k \cos(\frac{\pi}{A}kx) + b_k \sin(\frac{\pi}{A}kx). \tag{8.1}$$

Then the Fourier coefficients $a_k$ and $b_k$ are

$$a_k = \frac{1}{A}\int_0^{2A} f(x)\cos(\frac{\pi}{A}kx)dx, \tag{8.2}$$

and

$$b_k = \frac{1}{A}\int_0^{2A} f(x)\sin(\frac{\pi}{A}kx)dx. \tag{8.3}$$

To obtain the Fourier coefficients we need to know $f(x)$ for all $x$ in the interval $[0, 2A]$. In a number of applications, we do not have complete knowledge of the function $f(x)$, but rather, we have measurements of $f(x)$ taken at a finite number of values of the variable $x$. In such circumstances, the *finite Fourier transform* can be used in place of Fourier series.

## 8.2 Linear Trigonometric Models

A popular finite-parameter model is to consider $f(x)$ as a finite sum of trigonometric functions. For example, we may assume that $f(x)$ is a func-

tion of the form

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{M}\Big(a_k\cos(\omega_k x) + b_k\sin(\omega_k x)\Big), \tag{8.4}$$

where the $\omega_k$ are known, but the $a_k$ and $b_k$ are not. We find the unknown $a_k$ and $b_k$ by fitting the model to the data. We obtain data $f(x_n)$ corresponding to the $N$ points $x_n$, for $n = 0, 1, ..., N-1$, where $N = 2M+1$, and we solve the system

$$f(x_n) = \frac{1}{2}a_0 + \sum_{k=1}^{M}\Big(a_k\cos(\omega_k x_n) + b_k\sin(\omega_k x_n)\Big),$$

for $n = 0, ..., N-1$, to find the $a_k$ and $b_k$.

When $M$ is large, calculating the coefficients can be time-consuming. One particular choice for the $x_n$ and $\omega_k$ reduces the computation time significantly.

### 8.2.1 Equi-Spaced Frequencies

It is often the case that we can choose the $x_n$ at which we evaluate or measure the function $f(x)$. We suppose now that we have selected $N = 2M+1$ evaluation points equi-spaced from $x = 0$ to $x = 2A$; that is, $x_n = \frac{2An}{N}$, for $n = 0, ..., N-1$. Now let us select $\omega_k = \frac{\pi}{A}k$, for $k = 1, ..., M$. These are $M$ values of the variable $\omega$, equi-spaced within the interval $(0, \frac{M\pi}{A}]$. Our model for the function $f(x)$ is now

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{M}\Big(a_k\cos(\frac{\pi}{A}kx) + b_k\sin(\frac{\pi}{A}kx)\Big). \tag{8.5}$$

In keeping with the common notation, we write $f_n = f(\frac{2An}{N})$ for $n = 0, ..., N-1$. Then we have to solve the system

$$f_n = \frac{1}{2}a_0 + \sum_{k=1}^{M}\Big(a_k\cos(\frac{2\pi}{N}kn) + b_k\sin(\frac{2\pi}{N}kn)\Big), \tag{8.6}$$

for $n = 0, ..., N-1$, to find the $N$ coefficients $a_0$ and $a_k$ and $b_k$, for $k = 1, ..., M$. These $N$ coefficients are known, collectively, as the *finite Fourier transform* of the data.

### 8.2.2 Simplifying the Calculations

Calculating the solution of a system of $N$ linear equations in $N$ unknowns generally requires the number of multiplications to be on the order of $N^3$.

As we shall see in this subsection, choosing $\omega_k = \frac{\pi}{A}k$ leads to a form of orthogonality that will allow us to calculate the parameters in a relatively simple manner, with the number of multiplications on the order of $N^2$. Later, we shall see how to use the *fast Fourier transform* algorithm to reduce the number of computations even more.

For fixed $j = 1, ..., M$ consider the sums

$$\sum_{n=0}^{N-1} f_n \cos(\frac{2\pi}{N}jn) = \frac{1}{2}a_0 \sum_{n=0}^{N-1} \cos(\frac{2\pi}{N}jn)$$

$$+\sum_{k=1}^{M} \Bigg( a_k \Big( \sum_{n=0}^{N-1} \cos(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn)\Big)$$

$$+b_k \Big( \sum_{n=0}^{N-1} \sin(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn)\Big)\Bigg), \tag{8.7}$$

and

$$\sum_{n=0}^{N-1} f_n \sin(\frac{2\pi}{N}jn) = \frac{1}{2}a_0 \sum_{n=0}^{N-1} \sin(\frac{2\pi}{N}jn)$$

$$+\sum_{k=1}^{M} \Bigg( a_k \Big( \sum_{n=0}^{N-1} \cos(\frac{2\pi}{N}kn)\sin(\frac{2\pi}{N}jn)\Big)$$

$$+b_k \Big( \sum_{n=0}^{N-1} \sin(\frac{2\pi}{N}kn)\sin(\frac{2\pi}{N}jn)\Big)\Bigg). \tag{8.8}$$

We want to obtain the following:

**Lemma 8.1** *For $N = 2M + 1$ and $j, k = 1, 2, ..., M$, we have*

$$\sum_{n=0}^{N-1} \sin(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn) = 0,$$

$$\sum_{n=0}^{N-1} \cos(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn) = \begin{cases} 0, & \text{if } j \neq k; \\ \frac{N}{2}, & \text{if } j = k \neq 0; \\ N, & \text{if } j = k = 0; \end{cases}$$

*and*

$$\sum_{n=0}^{N-1} \sin(\frac{2\pi}{N}kn)\sin(\frac{2\pi}{N}jn) = \begin{cases} 0, & \text{if } j \neq k, \text{ or } j = k = 0; \\ \frac{N}{2}, & \text{if } j = k \neq 0. \end{cases}$$

**Exercise 8.1** *Using trigonometric identities, show that*

$$\cos(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn) = \frac{1}{2}\Big(\cos(\frac{2\pi}{N}(k+j)n) + \cos(\frac{2\pi}{N}(k-j)n)\Big),$$

$$\sin(\frac{2\pi}{N}kn)\cos(\frac{2\pi}{N}jn) = \frac{1}{2}\Big(\sin(\frac{2\pi}{N}(k+j)n) + \sin(\frac{2\pi}{N}(k-j)n)\Big),$$

*and*

$$\sin(\frac{2\pi}{N}kn)\sin(\frac{2\pi}{N}jn) = -\frac{1}{2}\Big(\cos(\frac{2\pi}{N}(k+j)n) - \cos(\frac{2\pi}{N}(k-j)n)\Big).$$

**Exercise 8.2** *Use trigonometric identities to show that*

$$\sin((n+\frac{1}{2})x) - \sin((n-\frac{1}{2})x) = 2\sin(\frac{x}{2})\cos(nx),$$

*and*

$$\cos((n+\frac{1}{2})x) - \cos((n-\frac{1}{2})x) = -2\sin(\frac{x}{2})\sin(nx).$$

**Exercise 8.3** *Use the previous exercise to show that*

$$2\sin(\frac{x}{2})\sum_{n=0}^{N-1}\cos(nx) = \sin((N-\frac{1}{2})x) + \sin(\frac{x}{2}),$$

*and*

$$2\sin(\frac{x}{2})\sum_{n=0}^{N-1}\sin(nx) = \cos(\frac{x}{2}) - \cos((N-\frac{1}{2})x).$$

*Hints: sum over* $n = 0, 1, ..., N-1$ *on both sides and note that*

$$\sin(\frac{x}{2}) = -\sin(-\frac{x}{2}).$$

**Exercise 8.4** *Use trigonometric identities to show that*

$$\sin((N-\frac{1}{2})x) + \sin(\frac{x}{2}) = 2\cos(\frac{N-1}{2}x)\sin(\frac{N}{2}x),$$

*and*

$$\cos\frac{x}{2} - \cos((N-\frac{1}{2})x) = 2\sin(\frac{N}{2}x)\sin(\frac{N-1}{2}x).$$

*Hints: Use*

$$N - \frac{1}{2} = \frac{N}{2} + \frac{N-1}{2},$$

*and*

$$\frac{1}{2} = \frac{N}{2} - \frac{N-1}{2}.$$

**Exercise 8.5** *Use the previous exercises to show that*

$$\sin(\frac{x}{2})\sum_{n=0}^{N-1}\cos(nx) = \sin(\frac{N}{2}x)\cos(\frac{N-1}{2}x),$$

*and*

$$\sin(\frac{x}{2})\sum_{n=0}^{N-1}\sin(nx) = \sin(\frac{N}{2}x)\sin(\frac{N-1}{2}x).$$

Let $m$ be any integer. Substituting $x = \frac{2\pi m}{N}$ in the equations in the previous exercise, we obtain

$$\sin(\frac{\pi}{N}m)\sum_{n=0}^{N-1}\cos(\frac{2\pi mn}{N}) = \sin(\pi m)\cos(\frac{N-1}{N}\pi m), \tag{8.9}$$

and

$$\sin(\frac{\pi}{N}m)\sum_{n=0}^{N-1}\sin(\frac{2\pi mn}{N}) = \sin(\pi m)\sin(\frac{N-1}{N}\pi m). \tag{8.10}$$

With $m = k + j$, we have

$$\sin(\frac{\pi}{N}(k+j))\sum_{n=0}^{N-1}\cos(\frac{2\pi(k+j)n}{N}) = \sin(\pi(k+j))\cos(\frac{N-1}{N}\pi(k+j)) \tag{8.11}$$

and

$$\sin(\frac{\pi}{N}(k+j))\sum_{n=0}^{N-1}\sin(\frac{2\pi(k+j)n}{N}) = \sin(\pi(k+j))\sin(\frac{N-1}{N}\pi(k+j)) \tag{8.12}$$

Similarly, with $m = k - j$, we obtain

$$\sin(\frac{\pi}{N}(k-j))\sum_{n=0}^{N-1}\cos(\frac{2\pi(k-j)n}{N}) = \sin(\pi(k-j))\cos(\frac{N-1}{N}\pi(k-j)) \tag{8.13}$$

and

$$\sin(\frac{\pi}{N}(k-j))\sum_{n=0}^{N-1}\sin(\frac{2\pi(k-j)n}{N}) = \sin(\pi(k-j))\sin(\frac{N-1}{N}\pi(k-j)) \tag{8.14}$$

**Exercise 8.6** *Prove Lemma 8.1.*

It follows immediately from Lemma 8.1 that

$$\sum_{n=0}^{N-1} f_n = Na_0,$$

and that

$$\sum_{n=0}^{N-1} f_n \cos(\frac{2\pi}{N}jn) = \frac{N}{2}a_j,$$

and

$$\sum_{n=0}^{N-1} f_n \sin(\frac{2\pi}{N}jn) = \frac{N}{2}b_j,$$

for $j = 1, ..., M$.

## 8.3 From Real to Complex

Throughout these notes we have limited the discussion to real data and models involving only real coefficients and real-valued functions. It is more common to use complex data and complex-valued models. Limiting the discussion to the real numbers comes at a price. Although complex variables may not be as familiar to the reader as real variables, there is some advantage in allowing the data and the models to be complex, as is the common practice in signal processing.

Suppose now that $f(x)$ is complex, for $0 \leq x \leq 2A$, and, as before, we have evaluated $f(x)$ at the $N = 2M+1$ points $x = \frac{2A}{N}n$, $n = 0, 1, ..., N-1$. Now we have the $N$ complex numbers $f_n = f(\frac{2A}{N}n)$.

In the model for the real-valued $f(x)$ given by Equation (8.5) it appeared that we used only $M + 1$ values of $\omega_k$, including $\omega_0 = 0$ for the constant term. In fact, though, if we were to express the sine and cosine functions in terms of complex exponential functions, we would see that we have used the frequencies $\frac{\pi}{A}j$, for $j = -M, ..., M$, so we have really used $2M + 1 = N$ frequencies. In the complex version, we explicitly use $N$ frequencies spaced $\frac{\pi}{A}$ apart. It is traditional that we use the frequencies $\frac{\pi}{A}k$, for $k = 0, 1, ..., N - 1$, although other choices are possible.

Given the (possibly) complex values $f_n = f(\frac{2A}{N}n)$, $n = 0, 1, ..., N - 1$, we model the function $f(x)$ as a finite sum of $N$ complex exponentials:

$$f(x) = \frac{1}{N}\sum_{k=0}^{N-1} F_k \exp(-i\frac{\pi}{A}kx), \tag{8.15}$$

where the coefficients $F_k$ are to be determined from the data $f_n$, $n =$

$0, 1, ..., N-1$. Setting $x = \frac{2A}{N}n$ in Equation (8.15), we have

$$f_n = \frac{1}{N}\sum_{k=0}^{N-1} F_k \exp(-i\frac{2\pi}{N}kn). \tag{8.16}$$

Suppose that $N = 2M + 1$. Using the formula for the sum of a finite geometric progression, we can easily show that

$$\sum_{m=-M}^{M} \exp(imx) = \frac{\sin((M + \frac{1}{2})x)}{\sin(\frac{x}{2})}, \tag{8.17}$$

whenever the denominator is not zero. From Equation (8.17) we can show that

$$\sum_{n=0}^{N-1} \exp(i\frac{2\pi}{N}kn)\exp(-i\frac{2\pi}{N}jn) = 0, \tag{8.18}$$

for $j \neq k$. It follows that the coefficients $F_k$ can be calculated as follows:

$$F_k = \sum_{n=0}^{N-1} f(n)\exp(i\frac{2\pi}{N}kn), \tag{8.19}$$

for $k = 0, 1, ..., N-1$.

Generally, given any (possibly) complex numbers $f_n$, $n = 0, 1, ..., N-1$, the collection of coefficients $F_k$, $k = 0, 1, ..., N-1$, is called its *complex finite Fourier transform.*

### 8.3.1 More Computational Issues

In many applications of signal processing $N$, the number of measurements of the function $f(x)$, can be quite large. We have found a relatively inexpensive way to find the undetermined parameters of the trigonometric model, but even this way poses computational problems when $N$ is large. The computation of a single $a_k$, $b_k$ or $F_k$ requires $N$ multiplications and we have to calculate $N$ of these parameters. Thus, the complexity of the problem is on the order of $N$ squared. Fortunately, there is a fast algorithm, known as the *fast Fourier transform* (FFT), that enables us to perform these calculations in far fewer multiplications.

# Chapter 9

# Transmission and Remote Sensing (Chapter 8)

## 9.1 Chapter Summary

In this chapter we illustrate the roles played by Fourier series and Fourier coefficients in the analysis of signal transmission and remote sensing.

## 9.2 Fourier Series and Fourier Coefficients

We suppose that $f(x)$ is defined for $-L \leq x \leq L$, with Fourier series representation

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(\frac{n\pi}{L}x) + b_n \sin(\frac{n\pi}{L}x). \tag{9.1}$$

The Fourier coefficients are

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos(\frac{n\pi}{L}x)dx, \tag{9.2}$$

and

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin(\frac{n\pi}{L}x)dx. \tag{9.3}$$

In the examples in this chapter, we shall see how Fourier coefficients can arise as data obtained through measurements. However, we shall be able to measure only a finite number of the Fourier coefficients. One issue

that will concern us is the effect on the representation of $f(x)$ if we use some, but not all, of its Fourier coefficients.

Suppose that we have $a_n$ and $b_n$ for $n = 1, 2, ..., N$. It is not unreasonable to try to estimate the function $f(x)$ using the *discrete Fourier transform* (DFT) estimate, which is

$$f_{DFT}(x) = \frac{1}{2}a_0 + \sum_{n=1}^{N} a_n \cos(\frac{n\pi}{L}x) + b_n \sin(\frac{n\pi}{L}x). \tag{9.4}$$

In Figure 9.1 below, the function $f(x)$ is the solid-line figure in both graphs. In the bottom graph, we see the true $f(x)$ and a DFT estimate. The top graph is the result of *band-limited extrapolation*, a technique for predicting missing Fourier coefficients.

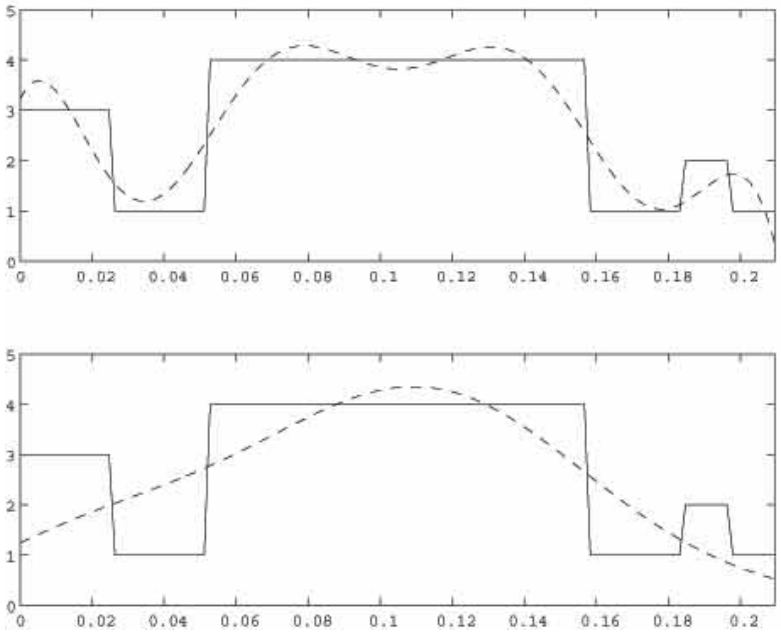



Figure 9.1: The non-iterative band-limited extrapolation method (MDFT) (top) and the DFT (bottom) for $M = 129$, $\Delta = 1$ and $\Omega = \pi/30$.

## 9.3 The Unknown Strength Problem

In this example, we imagine that each point $x$ in the interval $[-L, L]$ is sending a sine function signal at the frequency $\omega$, each with its own strength

$f(x)$; that is, the signal sent by the point $x$ is

$$f(x)\sin(\omega t). \tag{9.5}$$

In our first example, we imagine that the strength function $f(x)$ is unknown and we want to determine it. It could be the case that the signals originate at the points $x$, as with light or radio waves from the sun, or are simply reflected from the points $x$, as is sunlight from the moon or radio waves in radar. Later in this chapter, we shall investigate a related example, in which the points $x$ transmit known signals and we want to determine what is received elsewhere.

### 9.3.1 Measurement in the Far-Field

Now let us consider what is received by a point $P$ on the circumference of a circle centered at the origin and having large radius $D$. The point $P$ corresponds to the angle $\theta$ as shown in Figure 9.2; we use $\theta$ in the interval $[0, \pi]$. It takes a finite time for the signal sent from $x$ at time $t$ to reach $P$, so there is a delay.

We assume that $c$ is the speed at which the signal propagates. Because $D$ is large relative to $L$, we make the *far-field assumption*, which allows us to approximate the distance from $x$ to $P$ by $D - x\cos(\theta)$. Therefore, what $P$ receives at time $t$ is what was sent from $x$ at time $t - \frac{1}{c}(D - x\cos(\theta))$.

At time $t$, the point $P$ receives from $x$ the signal

$$f(x)\Big(\sin(\omega(t-\frac{D}{c}))\cos(\frac{\omega\cos(\theta)}{c}x) + \cos(\omega(t-\frac{D}{c}))\sin(\frac{\omega\cos(\theta)}{c}x)\Big), \tag{9.6}$$

and the point $Q$ corresponding to the angle $\theta + \pi$ receives

$$f(x)\Big(\sin(\omega(t-\frac{D}{c}))\cos(\frac{\omega\cos(\theta)}{c}x) - \cos(\omega(t-\frac{D}{c}))\sin(\frac{\omega\cos(\theta)}{c}x)\Big). \tag{9.7}$$

Adding the quantities in (9.6) and (9.7), we obtain

$$2\Big(f(x)\cos(\frac{\omega\cos(\theta)}{c}x)\Big)\sin(\omega(t-\frac{D}{c})), \tag{9.8}$$

while subtracting the latter from the former, we get

$$2\Big(f(x)\sin(\frac{\omega\cos(\theta)}{c}x)\Big)\cos(\omega(t-\frac{D}{c})). \tag{9.9}$$

Evaluating the signal in Equation (9.8) at the time when

$$\omega(t-\frac{D}{c}) = \frac{\pi}{2},$$

and dividing by 2, we get

$$f(x)\cos(\frac{\omega\cos(\theta)}{c}x),$$

while evaluating the signal in Equation (9.9) at the time when

$$\omega(t-\frac{D}{c})=2\pi$$

and dividing by 2 gives us

$$f(x)\sin(\frac{\omega\cos(\theta)}{c}x).$$

Because $P$ and $Q$ receive signals from all the $x$, not just from one $x$, what $P$ and $Q$ receive at time $t$ involves integrating over all $x$. Therefore, from our measurements at $P$ and $Q$ we obtain the quantities

$$\int_{-L}^{L} f(x)\cos(\frac{\omega\cos(\theta)}{c}x)dx, \tag{9.10}$$

and

$$\int_{-L}^{L} f(x)\sin(\frac{\omega\cos(\theta)}{c}x)dx. \tag{9.11}$$

If we can select an angle $\theta$ for which

$$\frac{\omega\cos(\theta)}{c}=\frac{n\pi}{L}, \tag{9.12}$$

then we have $a_n$ and $b_n$.

### 9.3.2 Limited Data

Note that we will be able to solve Equation (9.12) for $\theta$ only if we have

$$n\leq\frac{L\omega}{\pi c}. \tag{9.13}$$

This tells us that we can measure only finitely many of the Fourier coefficients of $f(x)$. It is common in signal processing to speak of the *wavelength* of a sinusoidal signal; the wavelength associated with a given $\omega$ and $c$ is

$$\lambda=\frac{2\pi c}{\omega}. \tag{9.14}$$

Therefore the number $N$ of Fourier coefficients we can measure is the largest integer not greater than $\frac{2L}{\lambda}$, which is the length of the interval $[-L,L]$, measured in units of wavelength $\lambda$. We get more Fourier coefficients when the product $L\omega$ is larger; this means that when $L$ is small, we want $\omega$ to be large, so that $\lambda$ is small and $N$ is large. As we saw previously, using these finitely many Fourier coefficients to calculate the DFT reconstruction of $f(x)$ can lead to a poor estimate of $f(x)$, particularly when $N$ is small.

### 9.3.3 Can We Get More Data?

As we just saw, we can make measurements at any points $P$ and $Q$ in the far-field; perhaps we do not need to limit ourselves to just those angles that lead to the $a_n$ and $b_n$. It may come as somewhat of a surprise, but from the theory of complex analytic functions we can prove that there is enough data available to us here to reconstruct $f(x)$ perfectly, at least in principle. The drawback, in practice, is that the measurements would have to be free of noise and impossibly accurate. All is not lost, however.

Suppose, for the sake of illustration, that we measure the far-field signals at points $P$ and $Q$ corresponding to angles $\theta$ that satisfy

$$\frac{\omega \cos(\theta)}{c} = \frac{n\pi}{2L}. \tag{9.15}$$

Now we have twice as many data points: we now have

$$A_n = \int_{-2L}^{2L} f(x) \cos(\frac{n\pi}{2L}) dx = \int_{-L}^{L} f(x) \cos(\frac{n\pi}{2L}) dx, \tag{9.16}$$

and

$$B_n = \int_{-2L}^{2L} f(x) \sin(\frac{n\pi}{2L}) dx = \int_{-L}^{L} f(x) \sin(\frac{n\pi}{2L}) dx, \tag{9.17}$$

for $n = 0, 1, ..., 2N$. We say now that our data is *twice over-sampled*.

Notice, however, that we have implicitly assumed that the interval of $x$ values from which signals are coming is now $[-2L, 2L]$, not the true $[-L, L]$; values of $x$ beyond $[-L, L]$ send no signals, so $f(x) = 0$ for those $x$. The data values we now have allow us to get Fourier coefficients $A_n$ and $B_n$ for the function $f(x)$ throughout $[-2L, 2L]$. We have twice the number of Fourier coefficients, but must reconstruct $f(x)$ over an interval that is twice as long. Over half of this interval $f(x) = 0$, so we waste effort if we use the $A_n$ and $B_n$ in the DFT, which will now reconstruct $f(x)$ over the interval $[-2L, 2L]$, on half of which $f(x)$ is known to be zero. But what else can we do?

Considerable research has gone into the use of prior knowledge about $f(x)$ to obtain reconstructions that are better than the DFT. In the example we are now considering, we have prior knowledge that $f(x) = 0$ for $L < |x| \leq 2L$. We can use this prior knowledge to improve our reconstruction. Suppose that we take as our reconstruction the *modified DFT* (MDFT), which is a function defined only for $|x| \leq L$ and having the form

$$f_{MDFT}(x) = \frac{1}{2}c_0 + \sum_{n=1}^{2N} c_n \cos(\frac{n\pi}{2L}x) + d_n \sin(\frac{n\pi}{2L}x), \tag{9.18}$$

where the $c_n$ and $d_n$ are not yet determined. Then we determine the $c_n$ and $d_n$ by requiring that the function $f_{MDFT}(x)$ could be the correct answer; that is, we require that $f_{MDFT}(x)$ be consistent with the measured data. Therefore, we must have

$$\int_{-L}^{L} f_{MDFT}(x)\cos(\frac{n\pi}{2L})dx = A_n, \tag{9.19}$$

and

$$\int_{-L}^{L} f_{MDFT}(x)\sin(\frac{n\pi}{2L})dx = B_n, \tag{9.20}$$

for $n = 0, 1, ..., 2N$. It is important to note now that the $c_n$ and $d_n$ are not the $A_n$ and $B_n$; this is because we no longer have orthogonality. For example, when we calculate the integral

$$\int_{-L}^{L} \cos(\frac{n\pi}{2L})\cos(\frac{m\pi}{2L})dx, \tag{9.21}$$

for $m \neq n$, we do not get zero. To find the $c_n$ and $d_n$ we need to solve a system of linear equations in these unknowns.

The top graph in Figure (9.1) illustrates the improvement over the DFT that can be had using the MDFT. In that figure, we took data that was thirty times over-sampled, not just twice over-sampled, as in our previous discussion. Consequently, we had thirty times the number of Fourier coefficients we would have had otherwise, but for an interval thirty times longer. To get the top graph, we used the MDFT, with the prior knowledge that $f(x)$ was non-zero only within the central thirtieth of the long interval. The bottom graph shows the DFT reconstruction using the larger data set, but only for the central thirtieth of the full period, which is where the original $f(x)$ is non-zero.

### 9.3.4 Other Forms of Prior Knowledge

As we just showed, knowing that we have over-sampled in our measurements can help us improve the resolution in our estimate of $f(x)$. We may have other forms of prior knowledge about $f(x)$ that we can use. If we know something about large-scale features of $f(x)$, but not about finer details, we can use the PDFT estimate, which is a generalization of the MDFT.

For example, we may know that $f(x)$ is non-negative, which we have not assumed explicitly previously in this chapter. Or, we may know that $f(x)$ is approximately zero for most $x$, but contains very sharp peaks at a few places. In more formal language, we may be willing to assume that $f(x)$ contains a few Dirac delta functions in a flat background. There are

non-linear methods, such as the maximum entropy method, the indirect PDFT (IPDFT), and eigenvector methods that can be used to advantage in such cases; these methods are often called *high-resolution methods*.

## 9.4 The Transmission Problem

### 9.4.1 Directionality

Now we turn the table around and suppose that we are designing a broadcasting system, using transmitters at each $x$ in the interval $[-L, L]$. At each $x$ we will transmit $f(x)\sin(\omega t)$, where both $f(x)$ and $\omega$ are chosen by us. We now want to calculate what will be received at each point $P$ in the far-field. We may wish to design the system so that the strengths of the signals received at the various $P$ are not all the same. For example, if we are broadcasting from Los Angeles, we may well want a strong signal in the north and south directions, but weak signals east and west, where there are fewer people to receive the signal. Clearly, our model of a single-frequency signal is too simple, but it does allow us to illustrate several important points about directionality in array processing.

### 9.4.2 The Case of Uniform Strength

For concreteness, we investigate the case in which $f(x) = 1$ for $|x| \leq L$. Since this function is even, we need only the $a_n$. In this case, the measurement of the signal at the point $P$ gives us

$$\frac{2c}{\omega\cos(\theta)}\sin(\frac{\omega\cos(\theta)}{c}), \tag{9.22}$$

whose absolute value is then the strength of the signal at $P$. Is it possible that the strength of the signal at some $P$ is zero?

To have zero signal strength, we need

$$\sin(\frac{L\omega\cos(\theta)}{c}) = 0,$$

without

$$\cos(\theta) = 0.$$

Therefore, we need

$$\frac{L\omega\cos(\theta)}{c} = n\pi, \tag{9.23}$$

for some positive integers $n \geq 1$. Notice that this can happen only if

$$n \leq \frac{L\omega\pi}{c} = \frac{2L}{\lambda}. \tag{9.24}$$

Therefore, if $2L < \lambda$, there can be no $P$ with signal strength zero. The larger $2L$ is, with respect to the wavelength $\lambda$, the more angles at which the signal strength is zero.

We have assumed here that each $x$ in the interval $[-L, L]$ is transmitting, but we can get a similar result using finitely many transmitters in $[-L, L]$. The graphs in Figures 9.3, 9.4, and 9.5 illustrate the sort of transmission patterns that can be designed by varying $\omega$. The figure captions refer to parameters used in a separate discussion, but the pictures are still instructive.

## 9.5 Remote Sensing

A basic problem in remote sensing is to determine the nature of a distant object by measuring signals transmitted by or reflected from that object. If the object of interest is sufficiently remote, that is, is in the *farfield*, the data we obtain by sampling the propagating spatio-temporal field is related, approximately, to what we want by *Fourier transformation*. The problem is then to estimate a function from finitely many (usually noisy) values of its *Fourier transform*. The application we consider here is a common one of remote-sensing of transmitted or reflected waves propagating from distant sources. Examples include optical imaging of planets and asteroids using reflected sunlight, radio-astronomy imaging of distant sources of radio waves, active and passive sonar, and radar imaging.

## 9.6 One-Dimensional Arrays

Now we imagine that the points $P$ are the sources of the signals and we are able to measure the transmissions at points $x$ in $[-L, L]$. The $P$ corresponding to the angle $\theta$ sends $F(\theta)\sin(\omega t)$, where the absolute value of $F(\theta)$ is the strength of the signal coming from $P$. In narrow-band passive sonar, for example, we may have hydrophone sensors placed at various points $x$ and our goal is to determine how much acoustic energy at a specified frequency is coming from different directions. There may be only a few directions contributing significant energy at the frequency of interest.

### 9.6.1 Measuring Fourier Coefficients

To simplify notation, we shall introduce the variable $u = \cos(\theta)$. We then have

$$\frac{du}{d\theta} = -\sin(\theta) = -\sqrt{1-u^2},$$

so that

$$d\theta = -\frac{1}{\sqrt{1-u^2}}du.$$

Now let $G(u)$ be the function

$$G(u) = \frac{F(\arccos(u))}{\sqrt{1-u^2}},$$

defined for $u$ in the interval $[-1, 1]$.

Measuring the signals received at $x$ and $-x$, we can obtain the integrals

$$\int_{-1}^{1} G(u)\cos(\frac{x\omega}{c}u)du, \tag{9.25}$$

and

$$\int_{-1}^{1} G(u)\sin(\frac{x\omega}{c}u)du. \tag{9.26}$$

The Fourier coefficients of $G(u)$ are

$$\frac{1}{2}\int_{-1}^{1} G(u)\cos(n\pi u)du, \tag{9.27}$$

and

$$\frac{1}{2}\int_{-1}^{1} G(u)\sin(n\pi u)du. \tag{9.28}$$

Therefore, in order to have our measurements match Fourier coefficients of $G(u)$ we need

$$\frac{x\omega}{c} = n\pi, \tag{9.29}$$

for some positive integer $n$. Therefore, we need to take measurements at the points $x$ and $-x$, where

$$x = n\frac{\pi c}{\omega} = n\frac{\lambda}{2} = n\Delta, \tag{9.30}$$

where $\Delta = \frac{\lambda}{2}$ is the *Nyquist spacing*. Since $x$ is restricted to $[-L, L]$, there is an upper limit to the $n$ we can use; we must have

$$n \leq \frac{L}{\lambda/2} = \frac{2L}{\lambda}. \tag{9.31}$$

The upper bound $\frac{2L}{\lambda}$, which is the length of our array of sensors, in units of wavelength, is often called the *aperture* of the array.

Once we have some of the Fourier coefficients of the function $G(u)$, we can estimate $G(u)$ for $|u| \leq 1$ and, from that estimate, obtain an estimate of the original $F(\theta)$.

As we just saw, the number of Fourier coefficients of $G(u)$ that we can measure, and therefore the resolution of the resulting reconstruction of $F(\theta)$, is limited by the aperture, that is, the length $2L$ of the array of sensors, divided by the wavelength $\lambda$. One way to improve resolution is to make the array of sensors longer, which is more easily said than done. However, *synthetic-aperture radar* (SAR) effectively does this. The idea of SAR is to mount the array of sensors on a moving airplane. As the plane moves, it effectively creates a longer array of sensors, a *virtual array* if you will. The one drawback is that the sensors in this virtual array are not all present at the same time, as in a normal array. Consequently, the data must be modified to approximate what would have been received at other times.

As in the examples discussed previously, we do have more measurements we can take, if we use values of $x$ other than those described by Equation (9.30). The issue will be what to do with these *over-sampled* measurements.

### 9.6.2 Over-sampling

One situation in which over-sampling arises naturally occurs in sonar array processing. Suppose that an array of sensors has been built to operate at a *design frequency* of $\omega_0$, which means that we have placed sensors at the points $x$ in $[-L, L]$ that satisfy the equation

$$x = n\frac{\pi c}{\omega_0} = n\frac{\lambda_0}{2} = n\Delta_0, \tag{9.32}$$

where $\lambda_0$ is the wavelength corresponding to the frequency $\omega_0$ and $\Delta_0 = \frac{\lambda_0}{2}$ is the Nyquist spacing for frequency $\omega_0$. Now suppose that we want to operate the sensing at another frequency, say $\omega$. The sensors cannot be moved, so we must make due with sensors at the points $x$ determined by the design frequency.

Consider, first, the case in which the second frequency $\omega$ is less than the design frequency $\omega_0$. Then its wavelength $\lambda$ is larger than $\lambda_0$, and the Nyquist spacing $\Delta = \frac{\lambda}{2}$ for $\omega$ is larger than $\Delta_0$. So we have over-sampled.

The measurements taken at the sensors provide us with the integrals

$$\frac{1}{K}\int_{-1}^{1} G(u)\cos(\frac{n\pi}{K}u)du, \tag{9.33}$$

and

$$\frac{1}{K}\int_{-1}^{1} G(u)\sin(\frac{n\pi}{K}u)du, \tag{9.34}$$

where $K = \frac{\omega_0}{\omega} > 1$. These are Fourier coefficients of the function $G(u)$, viewed as defined on the interval $[-K, K]$, which is larger than $[-1, 1]$, and

taking the value zero outside $[-1, 1]$. If we then use the DFT estimate of $G(u)$, it will estimate $G(u)$ for the values of $u$ within $[-1, 1]$, which is what we want, as well as for the values of $u$ outside $[-1, 1]$, where we already know $G(u)$ to be zero. Once again, we can use the modified DFT, the MDFT, to include the prior knowledge that $G(u) = 0$ for $u$ outside $[-1, 1]$ to improve our reconstruction of $G(u)$ and $F(\theta)$. In the over-sampled case the interval $[-1, 1]$ is called *the visible region* (although *audible region* seems more appropriate for sonar), since it contains all the values of $u$ that can correspond to actual angles of arrival of acoustic energy.

### 9.6.3 Under-sampling

Now suppose that the frequency $\omega$ that we want to consider is greater than the design frequency $\omega_0$. This means that the spacing between the sensors is too large; we have *under-sampled*. Once again, however, we cannot move the sensors and must make due with what we have.

Now the measurements at the sensors provide us with the integrals

$$\frac{1}{K}\int_{-1}^{1} G(u)\cos(\frac{n\pi}{K}u)du, \tag{9.35}$$

and

$$\frac{1}{K}\int_{-1}^{1} G(u)\sin(\frac{n\pi}{K}u)du, \tag{9.36}$$

where $K = \frac{\omega_0}{\omega} < 1$. These are Fourier coefficients of the function $G(u)$, viewed as defined on the interval $[-K, K]$, which is smaller than $[-1, 1]$, and taking the value zero outside $[-K, K]$. Since $G(u)$ is not necessarily zero outside $[-K, K]$, treating it as if it were zero there results in a type of error known as *aliasing*, in which energy corresponding to angles whose $u$ lies outside $[-K, K]$ is mistakenly assigned to values of $u$ that lie within $[-K, K]$. Aliasing is a common phenomenon; the strobe-light effect is aliasing, as is the apparent backward motion of the wheels of stage-coaches in cowboy movies. In the case of the strobe light, we are permitted to view the scene at times too far apart for us to sense continuous, smooth motion. In the case of the wagon wheels, the frames of the film capture instants of time too far apart for us to see the true rotation of the wheels.

## 9.7 Higher Dimensional Arrays

Up to now, we have considered sensors placed within a one-dimensional interval $[-L, L]$ and signals propagating within a plane containing $[-L, L]$. In such an arrangement there is a bit of ambiguity; we cannot tell if a

signal is coming from the angle $\theta$ or the angle $\theta + \pi$. When propagating signals can come to the array from any direction in three-dimensional space, there is greater ambiguity. To resolve the ambiguities, we can employ two- and three-dimensional arrays of sensors. To analyze the higher-dimensional cases, it is helpful to use the wave equation.

### 9.7.1 The Wave Equation

In many areas of remote sensing, what we measure are the fluctuations in time of an electromagnetic or acoustic field. Such fields are described mathematically as solutions of certain partial differential equations, such as the *wave equation*. A function $u(x, y, z, t)$ is said to satisfy the *three-dimensional wave equation* if

$$u_{tt} = c^2(u_{xx} + u_{yy} + u_{zz}) = c^2\nabla^2 u, \tag{9.37}$$

where $u_{tt}$ denotes the second partial derivative of $u$ with respect to the time variable $t$ twice and $c > 0$ is the (constant) speed of propagation. More complicated versions of the wave equation permit the speed of propagation $c$ to vary with the spatial variables $x, y, z$, but we shall not consider that here.

We use the method of *separation of variables* at this point, to get some idea about the nature of solutions of the wave equation. Assume, for the moment, that the solution $u(t, x, y, z)$ has the simple form

$$u(t, x, y, z) = f(t)g(x, y, z). \tag{9.38}$$

Inserting this separated form into the wave equation, we get

$$f''(t)g(x, y, z) = c^2 f(t)\nabla^2 g(x, y, z) \tag{9.39}$$

or

$$f''(t)/f(t) = c^2\nabla^2 g(x, y, z)/g(x, y, z). \tag{9.40}$$

The function on the left is independent of the spatial variables, while the one on the right is independent of the time variable; consequently, they must both equal the same constant, which we denote $-\omega^2$. From this we have two separate equations,

$$f''(t) + \omega^2 f(t) = 0, \tag{9.41}$$

and

$$\nabla^2 g(x, y, z) + \frac{\omega^2}{c^2} g(x, y, z) = 0. \tag{9.42}$$

Equation (9.42) is the *Helmholtz equation*.

Equation (9.41) has for its solutions the functions $f(t) = \cos(\omega t)$ and $\sin(\omega t)$. Functions $u(t, x, y, z) = f(t)g(x, y, z)$ with such time dependence are called *time-harmonic* solutions.

### 9.7.2 Planewave Solutions

Suppose that, beginning at time $t = 0$, there is a localized disturbance. As time passes, that disturbance spreads out spherically. When the radius of the sphere is very large, the surface of the sphere appears planar, to an observer on that surface, who is said then to be in the *far field*. This motivates the study of solutions of the wave equation that are constant on planes; the so-called *planewave solutions*.

Let $\mathbf{s} = (x, y, z)$ and $u(\mathbf{s}, t) = u(x, y, z, t) = e^{i\omega t}e^{i\mathbf{k}\cdot\mathbf{s}}$. Then we can show that $u$ satisfies the wave equation $u_{tt} = c^2\nabla^2 u$ for any real vector $\mathbf{k}$, so long as $||\mathbf{k}||^2 = \omega^2/c^2$. This solution is a planewave associated with frequency $\omega$ and *wavevector* $\mathbf{k}$; at any fixed time the function $u(\mathbf{s}, t)$ is constant on any plane in three-dimensional space having $\mathbf{k}$ as a normal vector.

In radar and sonar, the field $u(\mathbf{s}, t)$ being sampled is usually viewed as a discrete or continuous superposition of planewave solutions with various amplitudes, frequencies, and wavevectors. We sample the field at various spatial locations $\mathbf{s}$, for various times $t$. Here we simplify the situation a bit by assuming that all the planewave solutions are associated with the same frequency, $\omega$. If not, we can perform an FFT on the functions of time received at each sensor location $\mathbf{s}$ and keep only the value associated with the desired frequency $\omega$.

### 9.7.3 Superposition and the Fourier Transform

It is notationally convenient now to use the complex exponential functions

$$e^{i\omega t} = \cos(\omega t) + i\sin(\omega t)$$

instead of $\cos(\omega t)$ and $\sin(\omega t)$.

In the continuous superposition model, the field is

$$u(\mathbf{s}, t) = e^{i\omega t}\int F(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{s}}d\mathbf{k}. \tag{9.43}$$

Our measurements at the sensor locations $\mathbf{s}$ give us the values

$$f(\mathbf{s}) = \int F(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{s}}d\mathbf{k}. \tag{9.44}$$

The data are then Fourier transform values of the complex function $F(\mathbf{k})$; $F(\mathbf{k})$ is defined for all three-dimensional real vectors $\mathbf{k}$, but is zero, in theory, at least, for those $\mathbf{k}$ whose squared length $||\mathbf{k}||^2$ is not equal to $\omega^2/c^2$. Our goal is then to estimate $F(\mathbf{k})$ from measured values of its Fourier transform. Since each $\mathbf{k}$ is a normal vector for its planewave field component, determining the value of $F(\mathbf{k})$ will tell us the strength of the planewave component coming from the direction $\mathbf{k}$.

### 9.7.4 The Spherical Model

We can imagine that the sources of the planewave fields are the points $P$ that lie on the surface of a large sphere centered at the origin. For each $P$, the ray from the origin to $P$ is parallel to some wavevector $\mathbf{k}$. The function $F(\mathbf{k})$ can then be viewed as a function $F(P)$ of the points $P$. Our measurements will be taken at points $\mathbf{s}$ inside this sphere. The radius of the sphere is assumed to be orders of magnitude larger than the distance between sensors. The situation is that of astronomical observation of the heavens using ground-based antennas. The sources of the optical or electromagnetic signals reaching the antennas are viewed as lying on a large sphere surrounding the earth. Distance to the sources is not considered now, and all we are interested in are the amplitudes $F(\mathbf{k})$ of the fields associated with each direction $\mathbf{k}$.

### 9.7.5 The Two-Dimensional Array

In some applications the sensor locations are essentially arbitrary, while in others their locations are carefully chosen. Sometimes, the sensors are collinear, as in sonar towed arrays. Figure 9.6 illustrates a line array.

Suppose now that the sensors are in locations $\mathbf{s} = (x, y, 0)$, for various $x$ and $y$; then we have a *planar array* of sensors. Then the dot product $\mathbf{s} \cdot \mathbf{k}$ that occurs in Equation (9.44) is

$$\mathbf{s} \cdot \mathbf{k} = xk_1 + yk_2; \tag{9.45}$$

we cannot *see* the third component, $k_3$. However, since we know the size of the vector $\mathbf{k}$, we can determine $|k_3|$. The only ambiguity that remains is that we cannot distinguish sources on the upper hemisphere from those on the lower one. In most cases, such as astronomy, it is obvious in which hemisphere the sources lie, so the ambiguity is resolved.

The function $F(\mathbf{k})$ can then be viewed as $F(k_1, k_2)$, a function of the two variables $k_1$ and $k_2$. Our measurements give us values of $f(x, y)$, the two-dimensional Fourier transform of $F(k_1, k_2)$. Because of the limitation $||\mathbf{k}|| = \frac{\omega}{c}$, the function $F(k_1, k_2)$ has bounded support. Consequently, its Fourier transform cannot have bounded support. As a result, we can never have all the values of $f(x, y)$, and so cannot hope to reconstruct $F(k_1, k_2)$ exactly, even for noise-free data.

### 9.7.6 The One-Dimensional Array

If the sensors are located at points $\mathbf{s}$ having the form $\mathbf{s} = (x, 0, 0)$, then we have a *line array* of sensors, as we discussed previously. The dot product in Equation (9.44) becomes

$$\mathbf{s} \cdot \mathbf{k} = xk_1. \tag{9.46}$$

Now the ambiguity is greater than in the planar array case. Once we have $k_1$, we know that

$$k_2^2 + k_3^2 = (\frac{\omega}{c})^2 - k_1^2, \tag{9.47}$$

which describes points $P$ lying on a circle on the surface of the distant sphere, with the vector $(k_1, 0, 0)$ pointing at the center of the circle. It is said then that we have a *cone of ambiguity*. One way to resolve the situation is to assume $k_3 = 0$; then $|k_2|$ can be determined and we have remaining only the ambiguity involving the sign of $k_2$. Once again, in many applications, this remaining ambiguity can be resolved by other means.

Once we have resolved any ambiguity, we can view the function $F(\mathbf{k})$ as $F(k_1)$, a function of the single variable $k_1$. Our measurements give us values of $f(x)$, the Fourier transform of $F(k_1)$. As in the two-dimensional case, the restriction on the size of the vectors $\mathbf{k}$ means that the function $F(k_1)$ has bounded support. Consequently, its Fourier transform, $f(x)$, cannot have bounded support. Therefore, we shall never have all of $f(x)$, and so cannot hope to reconstruct $F(k_1)$ exactly, even for noise-free data.

### 9.7.7 Limited Aperture

In both the one- and two-dimensional problems, the sensors will be placed within some bounded region, such as $|x| \leq A$, $|y| \leq B$ for the two-dimensional problem, or $|x| \leq L$ for the one-dimensional case. The size of these bounded regions, in units of wavelength, are the *apertures* of the arrays. The larger these apertures are, the better the resolution of the reconstructions.

In digital array processing there are only finitely many sensors, which then places added limitations on our ability to reconstruction the field amplitude function $F(\mathbf{k})$.

## 9.8 An Example: The Solar-Emission Problem

In [5] Bracewell discusses the *solar-emission* problem. In 1942, it was observed that radio-wave emissions in the one-meter wavelength range were arriving from the sun. Were they coming from the entire disk of the sun or were the sources more localized, in sunspots, for example? The problem then was to view each location on the sun's surface as a potential source of these radio waves and to determine the intensity of emission corresponding to each location.

For electromagnetic waves the propagation speed is the speed of light in a vacuum, which we shall take here to be $c = 3 \times 10^8$ meters per second.

The wavelength $\lambda$ for gamma rays is around one Angstrom, which is $10^{-10}$ meters; for x-rays it is about one millimicron, or $10^{-9}$ meters. The visible spectrum has wavelengths that are a little less than one micron, that is, $10^{-6}$ meters. Shortwave radio has a wavelength around one millimeter; microwaves have wavelengths between one centimeter and one meter. Broadcast radio has a $\lambda$ running from about 10 meters to 1000 meters. The so-called long radio waves can have wavelengths several thousand meters long, prompting clever methods of antenna design for radio astronomy.

The sun has an angular diameter of 30 min. of arc, or one-half of a degree, when viewed from earth, but the needed resolution was more like 3 min. of arc. Such resolution requires a radio telescope 1000 wavelengths across, which means a diameter of 1km at a wavelength of 1 meter; in 1942 the largest military radar antennas were less than 5 meters across. A solution was found, using the method of reconstructing an object from line-integral data, a technique that surfaced again in tomography.

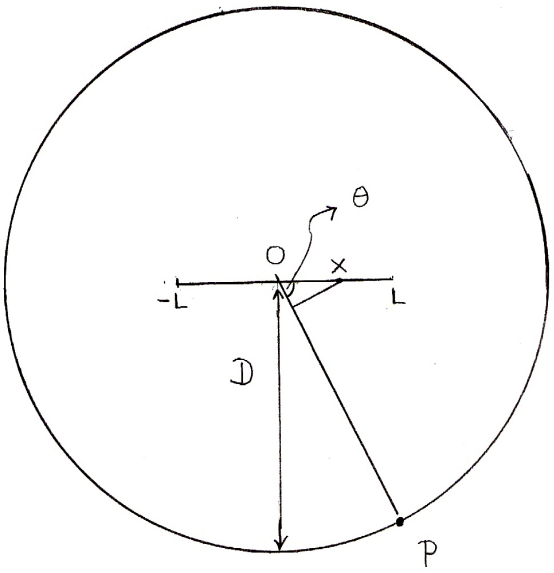



Figure 9.2: Farfield Measurements.

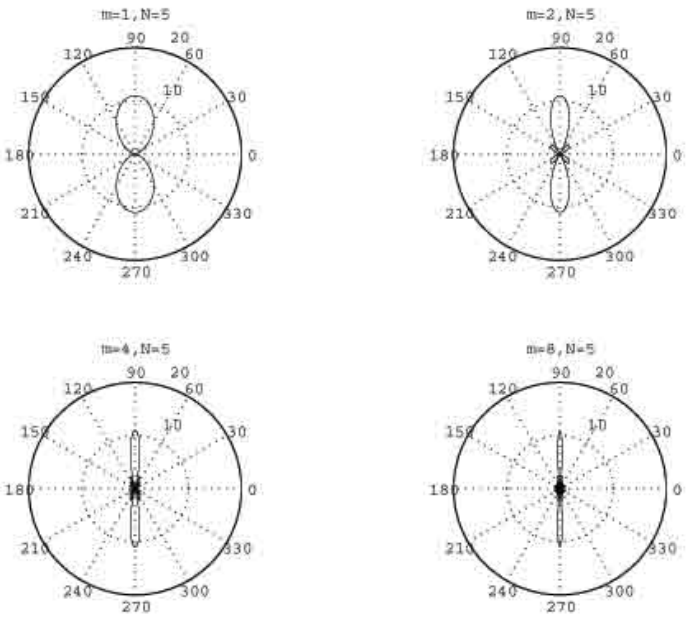



Figure 9.3: Transmission Pattern $A(\theta)$: $m = 1, 2, 4, 8$ and $N = 5$.

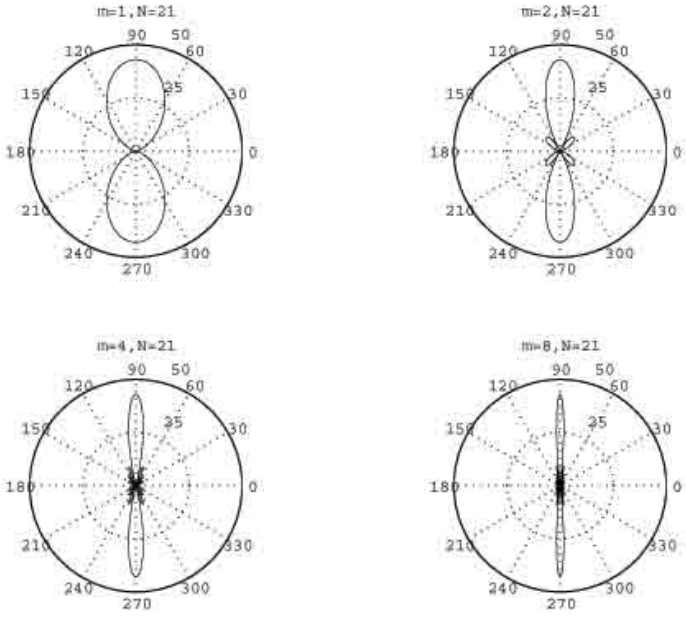



Figure 9.4: Transmission Pattern $A(\theta)$: $m = 1, 2, 4, 8$ and $N = 21$.

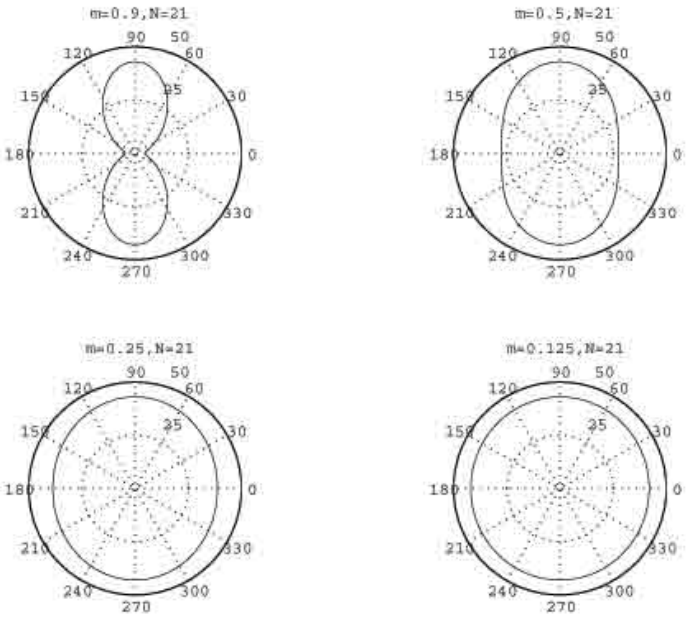



Figure 9.5: Transmission Pattern $A(\theta)$: $m = 0.9, 0.5, 0.25, 0.125$ and $N = 21$.

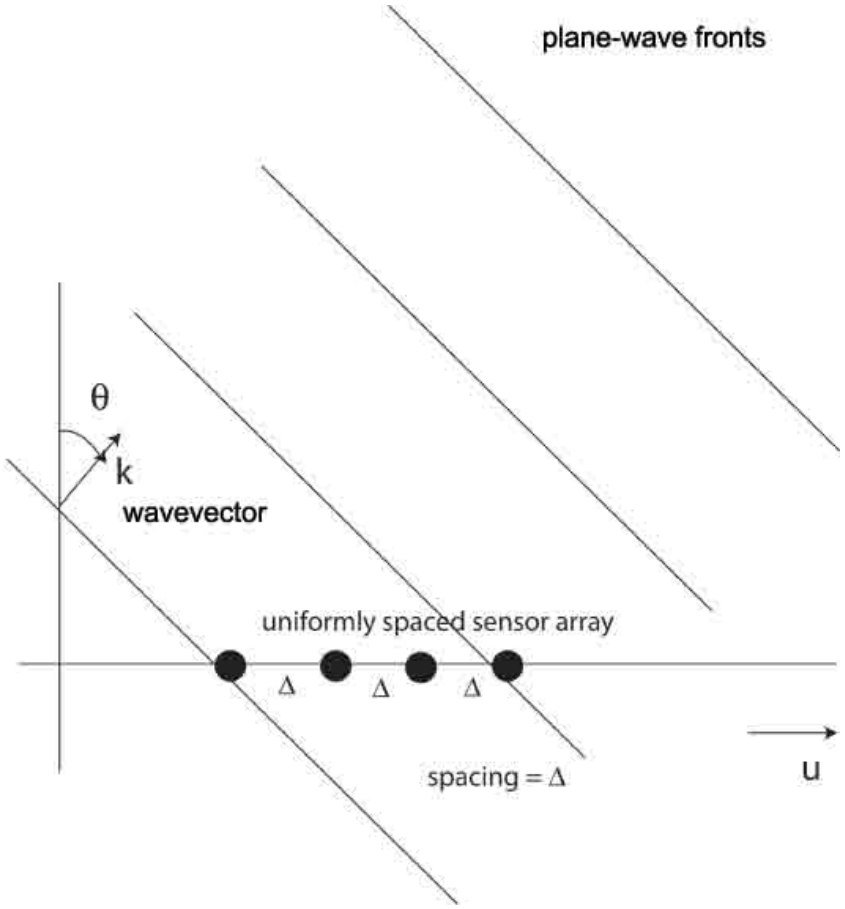



Figure 9.6: A uniform line array sensing a planewave field.

# Chapter 10

# Properties of the Fourier Transform (Chapter 8)

In this chapter we review the basic properties of the Fourier transform.

## 10.1 Fourier-Transform Pairs

Let $f(x)$ be defined for the real variable $x$ in $(-\infty, \infty)$. The *Fourier transform* (FT) of $f(x)$ is the function of the real variable $\omega$ given by

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{i\omega x} dx. \tag{10.1}$$

Having obtained $F(\omega)$ we can recapture the original $f(x)$ from the Fourier-Transform Inversion Formula:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{-i\omega x} d\omega. \tag{10.2}$$

Precisely how we interpret the infinite integrals that arise in the discussion of the Fourier transform will depend on the properties of the function $f(x)$.

### 10.1.1 Decomposing $f(x)$

One way to view Equation (10.2) is that it shows us the function $f(x)$ as a superposition of complex exponential functions $e^{-i\omega x}$, where $\omega$ runs over the entire real line. The use of the minus sign here is simply for notational convenience later. For each fixed value of $\omega$, the complex number $F(\omega) = |F(\omega)|e^{i\theta(\omega)}$ tells us that the amount of $e^{i\omega x}$ in $f(x)$ is $|F(\omega)|$, and that $e^{i\omega x}$ involves a phase shift by $\theta(\omega)$.

### 10.1.2 The Issue of Units

When we write $\cos \pi = -1$, it is with the understanding that $\pi$ is a measure of angle, in radians; the function cos will always have an independent variable in units of radians. By extension, the same is true of the complex exponential functions. Therefore, when we write $e^{ix\omega}$, we understand the product $x\omega$ to be in units of radians. If $x$ is measured in seconds, then $\omega$ is in units of radians per second; if $x$ is in meters, then $\omega$ is in units of radians per meter. When $x$ is in seconds, we sometimes use the variable $\frac{\omega}{2\pi}$; since $2\pi$ is then in units of radians per cycle, the variable $\frac{\omega}{2\pi}$ is in units of cycles per second, or Hertz. When we sample $f(x)$ at values of $x$ spaced $\Delta$ apart, the $\Delta$ is in units of $x$-units per sample, and the reciprocal, $\frac{1}{\Delta}$, which is called the *sampling frequency*, is in units of samples per $x$-units. If $x$ is in seconds, then $\Delta$ is in units of seconds per sample, and $\frac{1}{\Delta}$ is in units of samples per second.

## 10.2 Basic Properties of the Fourier Transform

In this section we present the basic properties of the Fourier transform. Proofs of these assertions are left as exercises.

**Exercise 10.1** *Let $F(\omega)$ be the FT of the function $f(x)$. Use the definitions of the FT and IFT given in Equations (10.1) and (10.2) to establish the following basic properties of the Fourier transform operation:*

- **Symmetry:** The FT of the function $F(x)$ is $2\pi f(-\omega)$. For example, the FT of the function $f(x) = \frac{\sin(\Omega x)}{\pi x}$ is $\chi_\Omega(\omega)$, so the FT of $g(x) = \chi_\Omega(x)$ is $G(\omega) = 2\pi \frac{\sin(\Omega\omega)}{\pi\omega}$.
- **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 $n$th 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 $\omega$:** Let $f, F$, $g, G$ and $h, H$ be FT pairs, with $h(x) = f(x)g(x)$. Then $H(\omega) = \frac{1}{2\pi}(F * G)(\omega)$.

**Definition 10.1** *A function $f : \mathbb{R} \to \mathbb{C}$ is said to be* even *if $f(-x) = f(x)$ for all $x$, and* odd *if $f(-x) = -f(x)$, for all $x$. Note that a typical function is neither even nor odd.*

**Exercise 10.2** *Show that $f$ is an even function if and only if its Fourier transform, $F$, is an even function.*

**Exercise 10.3** *Show that $f$ is real-valued if and only if its Fourier transform $F$ is* conjugate-symmetric*, that is, $F(-\omega) = \overline{F(\omega)}$. Therefore, $f$ is real-valued and even if and only if its Fourier transform $F$ is real-valued and even.*

## 10.3 Some Fourier-Transform Pairs

In this section we present several Fourier-transform pairs.

**Exercise 10.4** *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. Alternatively, perform the integration by completing the square.

Let $u(x)$ be the *Heaviside function* that is $+1$ if $x \geq 0$ and 0 otherwise. Let $\chi_A(x)$ be the *characteristic function* of the interval $[-A, A]$ that is $+1$ for $x$ in $[-A, A]$ 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 10.5** *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 $a$, where $u(x)$ is the Heaviside function.*

**Exercise 10.6** *Show that the FT of* $f(x) = \chi_A(x)$ *is* $F(\omega) = 2\frac{\sin(A\omega)}{\omega}$.

**Exercise 10.7** *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 $u = -x$.

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 10.8** *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 10.9** *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 $f(x)\text{sgn}(x) = f(x)$. Apply the convolution theorem, then compare real and imaginary parts.

## 10.4 Dirac Deltas

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 $\Omega$-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 $\Omega$-band-limited functions $g(x)$.

As $\Omega$ 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 $\omega$. 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

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ix\omega} d\omega. \tag{10.3}$$

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 (10.3) 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 $\omega$ 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 $\omega_0$, 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 10.10** *Use the fact that* $\mathrm{sgn}(x) = 2u(x) - 1$ *and the previous exercise to show that* $f(x) = u(x)$ *has the FT* $F(\omega) = i/\omega + \pi\delta(\omega)$.

**Exercise 10.11** *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

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

## 10.5 More Properties of the Fourier Transform

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

**Exercise 10.12** *Let* $f(x), F(\omega)$ *and* $g(x), G(\omega)$ *be Fourier transform pairs. Use Equation (10.3) 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 10.13** *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* $u$.

## 10.6 Convolution Filters

Let $h(x)$ and $H(\omega)$ be a Fourier-transform pair. We have mentioned several times the basic problem of estimating the function $H(\omega)$ from finitely many values of $h(x)$; for convenience now we use the symbols $h$ and $H$, rather than $f$ and $F$, as we did previously. Sometimes it is $H(\omega)$ that we really want. Other times it is the unmeasured values of $h(x)$ that we want, and we try to estimate them by first estimating $H(\omega)$. Sometimes, neither of these functions is our main interest; it may be the case that what we want is another function, $f(x)$, and $h(x)$ is a distorted version of $f(x)$. For example, suppose that $x$ is time and $f(x)$ represents what a speaker says into a telephone. The phone line distorts the signal somewhat, often diminishing the higher frequencies. What the person at the other end hears is not $f(x)$, but a related signal function, $h(x)$. For another example, suppose that $f(x, y)$ is a two-dimensional picture viewed by someone with poor eyesight. What that person sees is not $f(x, y)$ but a related function, $h(x, y)$, that is a distorted version of the true $f(x, y)$. In both examples, our goal is to recover the original undistorted signal or image. To do this, it helps to model the distortion. Convolution filters are commonly used for this purpose.

### 10.6.1 Blurring and Convolution Filtering

We suppose that what we measure are not values of $f(x)$, but values of $h(x)$, where the Fourier transform of $h(x)$ is

$$H(\omega) = F(\omega)G(\omega).$$

The function $G(\omega)$ describes the effects of the system, the telephone line in our first example, or the weak eyes in the second example, or the refraction of light as it passes through the atmosphere, in optical imaging. If we can use our measurements of $h(x)$ to estimate $H(\omega)$ and if we have some knowledge of the system distortion function, that is, some knowledge of $G(\omega)$ itself, then there is a chance that we can estimate $F(\omega)$, and thereby estimate $f(x)$.

If we apply the Fourier Inversion Formula to $H(\omega) = F(\omega)G(\omega)$, we get

$$h(x) = \frac{1}{2\pi} \int F(\omega)G(\omega)e^{-i\omega x}dx. \tag{10.4}$$

The function $h(x)$ that results is $h(x) = (f * g)(x)$, the *convolution* of the functions $f(x)$ and $g(x)$, with the latter given by

$$g(x) = \frac{1}{2\pi} \int G(\omega)e^{-i\omega x}dx. \tag{10.5}$$

Note that, if $f(x) = \delta(x)$, then $h(x) = g(x)$. In the image processing example, this says that if the true picture $f$ is a single bright spot, the blurred image $h$ is $g$ itself. For that reason, the function $g$ is called the *point-spread function* of the distorting system.

Convolution filtering refers to the process of converting any given function, say $f(x)$, into a different function, say $h(x)$, by convolving $f(x)$ with a fixed function $g(x)$. Since this process can be achieved by multiplying $F(\omega)$ by $G(\omega)$ and then inverse Fourier transforming, such convolution filters are studied in terms of the properties of the function $G(\omega)$, known in this context as the *system transfer function*, or the *optical transfer function* (OTF); when $\omega$ is a frequency, rather than a spatial frequency, $G(\omega)$ is called the *frequency-response function* of the filter. The magnitude of $G(\omega)$, $|G(\omega)|$, is called the *modulation transfer function* (MTF). The study of convolution filters is a major part of signal processing. Such filters provide both reasonable models for the degradation signals undergo, and useful tools for reconstruction.

Let us rewrite Equation (10.4), replacing $F(\omega)$ with its definition, as given by Equation (10.1). Then we have

$$h(x) = \int (\frac{1}{2\pi} \int f(t) e^{i\omega t} dt) G(\omega) e^{-i\omega x} d\omega. \tag{10.6}$$

Interchanging the order of integration, we get

$$h(x) = \int \int f(t) (\frac{1}{2\pi} \int G(\omega) e^{i\omega(t-x)} d\omega) dt. \tag{10.7}$$

The inner integral is $g(x-t)$, so we have

$$h(x) = \int f(t) g(x-t) dt; \tag{10.8}$$

this is the definition of the convolution of the functions $f$ and $g$.

### 10.6.2 Low-Pass Filtering

If we know the nature of the blurring, then we know $G(\omega)$, at least to some degree of precision. We can try to remove the blurring by taking measurements of $h(x)$, then estimating $H(\omega) = F(\omega)G(\omega)$, then dividing these numbers by the value of $G(\omega)$, and then inverse Fourier transforming. The problem is that our measurements are always noisy, and typical functions $G(\omega)$ have many zeros and small values, making division by $G(\omega)$ dangerous, except where the values of $G(\omega)$ are not too small. These values of $\omega$ tend to be the smaller ones, centered around zero, so that we end up with estimates of $F(\omega)$ itself only for the smaller values of $\omega$. The result is a *low-pass filtering* of the object $f(x)$.

To investigate such low-pass filtering, we suppose that $G(\omega) = 1$, for $|\omega| \leq \Omega$, and is zero, otherwise. Then the filter is called the ideal $\Omega$-low-pass filter. In the farfield propagation model, the variable $x$ is spatial, and the variable $\omega$ is spatial frequency, related to how the function $f(x)$ changes spatially, as we move $x$. Rapid changes in $f(x)$ are associated with values of $F(\omega)$ for large $\omega$. For the case in which the variable $x$ is time, the variable $\omega$ becomes frequency, and the effect of the low-pass filter on $f(x)$ is to remove its higher-frequency components.

One effect of low-pass filtering in image processing is to smooth out the more rapidly changing features of an image. This can be useful if these features are simply unwanted oscillations, but if they are important detail, such as edges, the smoothing presents a problem. Restoring such wanted detail is often viewed as removing the unwanted effects of the low-pass filtering; in other words, we try to recapture the missing high-spatial-frequency values that have been zeroed out. Such an approach to image restoration is called *frequency-domain extrapolation* . How can we hope to recover these missing spatial frequencies, when they could have been anything? To have some chance of estimating these missing values we need to have some prior information about the image being reconstructed.

## 10.7 Two-Dimensional Fourier Transforms

More generally, we consider a function $f(x, y)$ of two real variables. Its Fourier transformation is

$$F(\alpha, \beta) = \int \int f(x, y) e^{i(x\alpha + y\beta)} dx dz. \tag{10.9}$$

For example, suppose that $f(x, y) = 1$ for $\sqrt{x^2 + y^2} \leq R$, and zero, otherwise. Then we have

$$F(\alpha, \beta) = \int_{-\pi}^{\pi} \int_{0}^{R} e^{-i(\alpha r \cos\theta + \beta r \sin\theta)} r dr d\theta. \tag{10.10}$$

In polar coordinates, with $\alpha = \rho \cos\phi$ and $\beta = \rho \sin\phi$, we have

$$F(\rho, \phi) = \int_{0}^{R} \int_{-\pi}^{\pi} e^{ir\rho\cos(\theta - \phi)} d\theta r dr. \tag{10.11}$$

The inner integral is well known;

$$\int_{-\pi}^{\pi} e^{ir\rho\cos(\theta - \phi)} d\theta = 2\pi J_0(r\rho), \tag{10.12}$$

where $J_0$ denotes the 0th order Bessel function. Using the identity

$$\int_{0}^{z} t^n J_{n-1}(t) dt = z^n J_n(z), \tag{10.13}$$

we have

$$F(\rho, \phi) = \frac{2\pi R}{\rho} J_1(\rho R). \tag{10.14}$$

Notice that, since $f(x, z)$ is a radial function, that is, dependent only on the distance from $(0, 0)$ to $(x, y)$, its Fourier transform is also radial.

The first positive zero of $J_1(t)$ is around $t = 4$, so when we measure $F$ at various locations and find $F(\rho, \phi) = 0$ for a particular $(\rho, \phi)$, we can estimate $R \approx 4/\rho$. So, even when a distant spherical object, like a star, is too far away to be imaged well, we can sometimes estimate its size by finding where the intensity of the received signal is zero [32].

### 10.7.1 Two-Dimensional Fourier Inversion

Just as in the one-dimensional case, the Fourier transformation that produced $F(\alpha, \beta)$ can be inverted to recover the original $f(x, y)$. The Fourier Inversion Formula in this case is

$$f(x, y) = \frac{1}{4\pi^2} \int \int F(\alpha, \beta) e^{-i(\alpha x + \beta y)} d\alpha d\beta. \tag{10.15}$$

It is important to note that this procedure can be viewed as two one-dimensional Fourier inversions: first, we invert $F(\alpha, \beta)$, as a function of, say, $\beta$ only, to get the function of $\alpha$ and $y$

$$g(\alpha, y) = \frac{1}{2\pi} \int F(\alpha, \beta) e^{-i\beta y} d\beta; \tag{10.16}$$

second, we invert $g(\alpha, y)$, as a function of $\alpha$, to get

$$f(x, y) = \frac{1}{2\pi} \int g(\alpha, y) e^{-i\alpha x} d\alpha. \tag{10.17}$$

If we write the functions $f(x, y)$ and $F(\alpha, \beta)$ in polar coordinates, we obtain alternative ways to implement the two-dimensional Fourier inversion. We shall consider these other ways when we discuss the tomography problem of reconstructing a function $f(x, y)$ from line-integral data.

### 10.7.2 A Discontinuous Function

Consider the function $f(x) = \frac{1}{2A}$, for $|x| \leq A$, and $f(x) = 0$, otherwise. The Fourier transform of this $f(x)$ is

$$F(\omega) = \frac{\sin(A\omega)}{A\omega}, \tag{10.18}$$

for all real $\omega \neq 0$, and $F(0) = 1$. Note that $F(\omega)$ is nonzero throughout the real line, except for isolated zeros, but that it goes to zero as we go to the

infinities. This is typical behavior. Notice also that the smaller the $A$, the slower $F(\omega)$ dies out; the first zeros of $F(\omega)$ are at $|\omega| = \frac{\pi}{A}$, so the main lobe widens as $A$ goes to zero. The function $f(x)$ is not continuous, so its Fourier transform cannot be absolutely integrable. In this case, the Fourier-Transform Inversion Formula must be interpreted as involving convergence in the $L^2$ norm.

# Chapter 11

# Transmission Tomography (Chapter 8)

In this part of the text we focus on transmission tomography. This chapter will provide a detailed description of how the data is gathered, the mathematical model of the scanning process, and the problem to be solved. The emphasis here is on the role of the Fourier transform.

## 11.1 X-ray Transmission Tomography

Although transmission tomography is not limited to scanning living beings, we shall concentrate here on the use of x-ray tomography in medical diagnosis and the issues that concern us in that application. The mathematical formulation will, of course, apply more generally.

In x-ray tomography, x-rays are transmitted through the body along many lines. In some, but not all, cases, the lines will all lie in the same plane. The strength of the x-rays upon entering the body is assumed known, and the strength upon leaving the body is measured. This data can then be used to estimate the amount of attenuation the x-ray encountered along that line, which is taken to be the integral, along that line, of the attenuation function. On the basis of these line integrals, we estimate the attenuation function. This estimate is presented to the physician as one or more two-dimensional images.

## 11.2 The Exponential-Decay Model

As an x-ray beam passes through the body, it encounters various types of matter, such as soft tissue, bone, ligaments, air, each weakening the beam

to a greater or lesser extent. If the intensity of the beam upon entry is $I_{in}$ and $I_{out}$ is its lower intensity after passing through the body, then

$$I_{out} = I_{in} e^{-\int_L f},$$

where $f = f(x, y) \geq 0$ is the *attenuation function* describing the two-dimensional distribution of matter within the slice of the body being scanned and $\int_L f$ is the integral of the function $f$ over the line $L$ along which the x-ray beam has passed. To see why this is the case, imagine the line $L$ parameterized by the variable $s$ and consider the intensity function $I(s)$ as a function of $s$. For small $\Delta s > 0$, the drop in intensity from the start to the end of the interval $[s, s + \Delta s]$ is approximately proportional to the intensity $I(s)$, to the attenuation $f(s)$ and to $\Delta s$, the length of the interval; that is,

$$I(s) - I(s + \Delta s) \approx f(s) I(s) \Delta s.$$

Dividing by $\Delta s$ and letting $\Delta s$ approach zero, we get

$$I'(s) = -f(s) I(s).$$

**Exercise 11.1** *Show that the solution to this differential equation is*

$$I(s) = I(0) \exp(-\int_{u=0}^{u=s} f(u) du).$$

*Hint: Use an integrating factor.*

From knowledge of $I_{in}$ and $I_{out}$, we can determine $\int_L f$. If we know $\int_L f$ for every line in the $x$, $y$-plane we can reconstruct the attenuation function $f$. In the real world we know line integrals only approximately and only for finitely many lines. The goal in x-ray transmission tomography is to estimate the attenuation function $f(x, y)$ in the slice, from finitely many noisy measurements of the line integrals. We usually have prior information about the values that $f(x, y)$ can take on. We also expect to find sharp boundaries separating regions where the function $f(x, y)$ varies only slightly. Therefore, we need algorithms capable of providing such images.

## 11.3 Difficulties to be Overcome

There are several problems associated with this model. X-ray beams are not exactly straight lines; the beams tend to spread out. The x-rays are not monochromatic, and their various frequency components are attenuated at different rates, resulting in *beam hardening*, that is, changes in the spectrum of the beam as it passes through the object (see the appendix on the Laplace transform). The beams consist of photons obeying statistical laws, so our algorithms probably should be based on these laws. How we choose

the line segments is determined by the nature of the problem; in certain cases we are somewhat limited in our choice of these segments. Patients move; they breathe, their hearts beat, and, occasionally, they shift position during the scan. Compensating for these motions is an important, and difficult, aspect of the image reconstruction process. Finally, to be practical in a clinical setting, the processing that leads to the reconstructed image must be completed in a short time, usually around fifteen minutes. This time constraint is what motivates viewing the three-dimensional attenuation function in terms of its two-dimensional slices.

As we shall see, the Fourier transform and the associated theory of convolution filters play important roles in the reconstruction of transmission tomographic images.

The data we actually obtain at the detectors are counts of detected photons. These counts are not the line integrals; they are random quantities whose means, or expected values, are related to the line integrals. The Fourier inversion methods for solving the problem ignore its statistical aspects; in contrast, other methods, such as likelihood maximization, are based on a statistical model that involves Poisson-distributed emissions.

## 11.4 Reconstruction from Line Integrals

We turn now to the underlying problem of reconstructing attenuation functions from line-integral data.

### 11.4.1 The Radon Transform

Our goal is to reconstruct the function $f(x, y) \geq 0$ from line-integral data. Let $\theta$ be a fixed angle in the interval $[0, \pi)$. Form the $t, s$-axis system with the positive $t$-axis making the angle $\theta$ with the positive $x$-axis, as shown in Figure 11.1. Each point $(x, y)$ in the original coordinate system has coordinates $(t, s)$ in the second system, where the $t$ and $s$ are given by

$$t = x \cos\theta + y \sin\theta,$$

and

$$s = -x \sin\theta + y \cos\theta.$$

If we have the new coordinates $(t, s)$ of a point, the old coordinates are $(x, y)$ given by

$$x = t \cos\theta - s \sin\theta,$$

and

$$y = t \sin\theta + s \cos\theta.$$

We can then write the function $f$ as a function of the variables $t$ and $s$. For each fixed value of $t$, we compute the integral

$$\int_L f(x,y)ds = \int f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)ds$$

along the single line $L$ corresponding to the fixed values of $\theta$ and $t$. We repeat this process for every value of $t$ and then change the angle $\theta$ and repeat again. In this way we obtain the integrals of $f$ over every line $L$ in the plane. We denote by $r_f(\theta,t)$ the integral

$$r_f(\theta,t) = \int_L f(x,y)ds.$$

The function $r_f(\theta,t)$ is called the *Radon transform* of $f$.

### 11.4.2 The Central Slice Theorem

For fixed $\theta$ the function $r_f(\theta,t)$ is a function of the single real variable $t$; let $R_f(\theta,\omega)$ be its Fourier transform. Then

$$R_f(\theta,\omega) = \int r_f(\theta,t)e^{i\omega t}dt$$

$$= \int\int f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)e^{i\omega t}dsdt$$

$$= \int\int f(x,y)e^{i\omega(x\cos\theta + y\sin\theta)}dxdy = F(\omega\cos\theta, \omega\sin\theta),$$

where $F(\omega\cos\theta, \omega\sin\theta)$ is the two-dimensional Fourier transform of the function $f(x,y)$, evaluated at the point $(\omega\cos\theta, \omega\sin\theta)$; this relationship is called the *Central Slice Theorem*. For fixed $\theta$, as we change the value of $\omega$, we obtain the values of the function $F$ along the points of the line making the angle $\theta$ with the horizontal axis. As $\theta$ varies in $[0,\pi)$, we get all the values of the function $F$. Once we have $F$, we can obtain $f$ using the formula for the two-dimensional inverse Fourier transform. We conclude that we are able to determine $f$ from its line integrals.

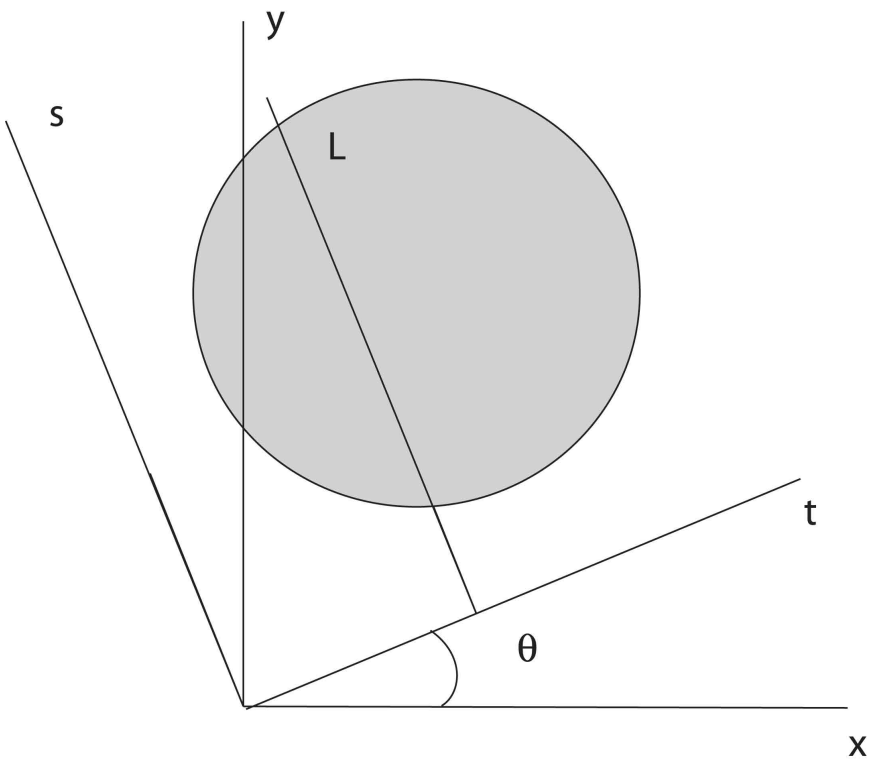



Figure 11.1: The Radon transform of $f$ at $(t, \theta)$ is the line integral of $f$ along line $L$.

# Chapter 12

# The ART and MART (Chapter 15)

## 12.1 Overview

In many applications, such as in image processing, the system of linear equations to be solved is quite large, often several tens of thousands of equations in about the same number of unknowns. In these cases, issues such as the costs of storage and retrieval of matrix entries, the computation involved in apparently trivial operations, such as matrix-vector products, and the speed of convergence of iterative methods demand greater attention. At the same time, the systems to be solved are often under-determined, and solutions satisfying certain additional constraints, such as non-negativity, are required. The ART and the MART are two iterative algorithms that are designed to address these issues.

Both the *algebraic reconstruction technique* (ART) and the *multiplicative algebraic reconstruction technique* (MART) were introduced as two iterative methods for discrete image reconstruction in transmission tomography.

Both methods are what are called *row-action* methods, meaning that each step of the iteration uses only a single equation from the system. The MART is limited to non-negative systems for which non-negative solutions are sought. In the under-determined case, both algorithms find the solution closest to the starting vector, in the two-norm or weighted two-norm sense for ART, and in the cross-entropy sense for MART, so both algorithms can be viewed as solving optimization problems. For both algorithms, the starting vector can be chosen to incorporate prior information about the desired solution. In addition,the ART can be employed in several ways to obtain a least-squares solution, in the over-determined case.

## 12.2 The ART in Tomography

For $i = 1, ..., I$, let $L_i$ be the set of pixel indices $j$ for which the $j$-th pixel intersects the $i$-th line segment, as shown in Figure 12.1, and let $|L_i|$ be the cardinality of the set $L_i$. Let $A_{ij} = 1$ for $j$ in $L_i$, and $A_{ij} = 0$ otherwise. With $i = k(\mathrm{mod}\, I) + 1$, the iterative step of the ART algorithm is

$$x_j^{k+1} = x_j^k + \frac{1}{|L_i|}(b_i - (Ax^k)_i), \tag{12.1}$$

for $j$ in $L_i$, and

$$x_j^{k+1} = x_j^k, \tag{12.2}$$

if $j$ is not in $L_i$. In each step of ART, we take the error, $b_i - (Ax^k)_i$, associated with the current $x^k$ and the $i$-th equation, and distribute it equally over each of the pixels that intersects $L_i$.

A somewhat more sophisticated version of ART allows $A_{ij}$ to include the length of the $i$-th line segment that lies within the $j$-th pixel; $A_{ij}$ is taken to be the ratio of this length to the length of the diagonal of the $j$-pixel.

More generally, ART can be viewed as an iterative method for solving an arbitrary system of linear equations, $Ax = b$.

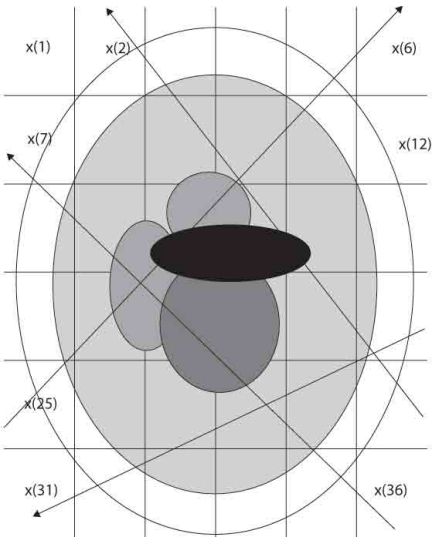



Figure 12.1: Line integrals through a discretized object.

## 12.3 The ART in the General Case

Let $A$ be a complex matrix with $I$ rows and $J$ columns, and let $b$ be a member of $\mathbb{C}^I$. We want to solve the system $Ax = b$.

For each index value $i$, let $H_i$ be the hyperplane of $J$-dimensional vectors given by

$$H_i = \{x|(Ax)_i = b_i\}, \tag{12.3}$$

and $P_i$ the orthogonal projection operator onto $H_i$. Let $x^0$ be arbitrary and, for each nonnegative integer $k$, let $i(k) = k(\mathrm{mod}\, I) + 1$. The iterative step of the ART is

$$x^{k+1} = P_{i(k)}x^k. \tag{12.4}$$

Because the ART uses only a single equation at each step, it has been called a *row-action* method. Figures 12.2 and 12.3 illustrate the behavior of the ART.

### 12.3.1 Calculating the ART

Given any vector $z$ the vector in $H_i$ closest to $z$, in the sense of the Euclidean distance, has the entries

$$x_j = z_j + \overline{A_{ij}}(b_i - (Az)_i)/\sum_{m=1}^{J} |A_{im}|^2. \tag{12.5}$$

To simplify our calculations, we shall assume, throughout this chapter, that the rows of $A$ have been rescaled to have Euclidean length one; that is

$$\sum_{j=1}^{J} |A_{ij}|^2 = 1, \tag{12.6}$$

for each $i = 1, ..., I$, and that the entries of $b$ have been rescaled accordingly, to preserve the equations $Ax = b$. The ART is then the following: begin with an arbitrary vector $x^0$; for each nonnegative integer $k$, having found $x^k$, the next iterate $x^{k+1}$ has entries

$$x_j^{k+1} = x_j^k + \overline{A_{ij}}(b_i - (Ax^k)_i). \tag{12.7}$$

When the system $Ax = b$ has exact solutions the ART converges to the solution closest to $x^0$, in the 2-norm. How fast the algorithm converges will depend on the ordering of the equations and on whether or not we use relaxation. In selecting the equation ordering, the important thing is to avoid particularly bad orderings, in which the hyperplanes $H_i$ and $H_{i+1}$ are nearly parallel.

### 12.3.2 When $Ax = b$ Has Solutions

For the consistent case, in which the system $Ax = b$ has exact solutions, we have the following result.

**Theorem 12.1** *Let $A\hat{x} = b$ and let $x^0$ be arbitrary. Let $\{x^k\}$ be generated by Equation (12.7). Then the sequence $\{||\hat{x} - x^k||_2\}$ is decreasing and $\{x^k\}$ converges to the solution of $Ax = b$ closest to $x^0$.*

### 12.3.3 When $Ax = b$ Has No Solutions

When there are no exact solutions, the ART does not converge to a single vector, but, for each fixed $i$, the subsequence $\{x^{nI+i},\ n = 0, 1, ...\}$ converges to a vector $z^i$ and the collection $\{z^i\,|i = 1, ..., I\}$ is called the *limit cycle*.

The ART limit cycle will vary with the ordering of the equations, and contains more than one vector unless an exact solution exists. There are several open questions about the limit cycle.

**Open Question:** For a fixed ordering, does the limit cycle depend on the initial vector $x^0$? If so, how?

### 12.3.4 The Geometric Least-Squares Solution

When the system $Ax = b$ has no solutions, it is reasonable to seek an approximate solution, such as the *least squares* solution, $x_{LS} = (A^\dagger A)^{-1}A^\dagger b$, which minimizes $||Ax - b||_2$. It is important to note that the system $Ax = b$ has solutions if and only if the related system $WAx = Wb$ has solutions, where $W$ denotes an invertible matrix; when solutions of $Ax = b$ exist, they are identical to those of $WAx = Wb$. But, when $Ax = b$ does not have solutions, the least-squares solutions of $Ax = b$, which need not be unique, but usually are, and the least-squares solutions of $WAx = Wb$ need not be identical. In the typical case in which $A^\dagger A$ is invertible, the unique least-squares solution of $Ax = b$ is

$$(A^\dagger A)^{-1}A^\dagger b, \tag{12.8}$$

while the unique least-squares solution of $WAx = Wb$ is

$$(A^\dagger W^\dagger WA)^{-1}A^\dagger W^\dagger b, \tag{12.9}$$

and these need not be the same.

A simple example is the following. Consider the system

$$x = 1$$

$$x = 2, \tag{12.10}$$

which has the unique least-squares solution $x = 1.5$, and the system

$$2x = 2$$

$$x = 2, \tag{12.11}$$

which has the least-squares solution $x = 1.2$.

**Definition 12.1** *The* geometric least-squares *solution of $Ax = b$ is the least-squares solution of $WAx = Wb$, for $W$ the diagonal matrix whose entries are the reciprocals of the Euclidean lengths of the rows of $A$.*

In our example above, the geometric least-squares solution for the first system is found by using $W_{11} = 1 = W_{22}$, so is again $x = 1.5$, while the geometric least-squares solution of the second system is found by using $W_{11} = 0.5$ and $W_{22} = 1$, so that the geometric least-squares solution is $x = 1.5$, not $x = 1.2$.

**Open Question:** If there is a unique geometric least-squares solution, where is it, in relation to the vectors of the limit cycle? Can it be calculated easily, from the vectors of the limit cycle?

There is a partial answer to the second question. It is known that if the system $Ax = b$ has no exact solution, and if $I = J + 1$, then the vectors of the limit cycle lie on a sphere in $J$-dimensional space having the least-squares solution at its center. This is not true more generally, however.

## 12.4 The MART

The *multiplicative* ART (MART) is an iterative algorithm closely related to the ART. It also was devised to obtain tomographic images, but, like ART, applies more generally; MART applies to systems of linear equations $Ax = b$ for which the $b_i$ are positive, the $A_{ij}$ are nonnegative, and the solution $x$ we seek is to have nonnegative entries. It is not so easy to see the relation between ART and MART if we look at the most general formulation of MART. For that reason, we begin with a simpler case, transmission tomographic imaging, in which the relation is most clearly visible.

### 12.4.1 A Special Case of MART

We begin by considering the application of MART to the transmission tomography problem. For $i = 1, ..., I$, let $L_i$ be the set of pixel indices $j$ for which the $j$-th pixel intersects the $i$-th line segment, and let $|L_i|$ be the

cardinality of the set $L_i$. Let $A_{ij} = 1$ for $j$ in $L_i$, and $A_{ij} = 0$ otherwise. With $i = k(\mathrm{mod}\, I) + 1$, the iterative step of the ART algorithm is

$$x_j^{k+1} = x_j^k + \frac{1}{|L_i|}(b_i - (Ax^k)_i), \tag{12.12}$$

for $j$ in $L_i$, and

$$x_j^{k+1} = x_j^k, \tag{12.13}$$

if $j$ is not in $L_i$. In each step of ART, we take the error, $b_i - (Ax^k)_i$, associated with the current $x^k$ and the $i$-th equation, and distribute it equally over each of the pixels that intersects $L_i$.

Suppose, now, that each $b_i$ is positive, and we know in advance that the desired image we wish to reconstruct must be nonnegative. We can begin with $x^0 > 0$, but as we compute the ART steps, we may lose nonnegativity. One way to avoid this loss is to correct the current $x^k$ multiplicatively, rather than additively, as in ART. This leads to the *multiplicative* ART (MART).

The MART, in this case, has the iterative step

$$x_j^{k+1} = x_j^k \Big( \frac{b_i}{(Ax^k)_i} \Big), \tag{12.14}$$

for those $j$ in $L_i$, and

$$x_j^{k+1} = x_j^k, \tag{12.15}$$

otherwise. Therefore, we can write the iterative step as

$$x_j^{k+1} = x_j^k \Big( \frac{b_i}{(Ax^k)_i} \Big)^{A_{ij}}. \tag{12.16}$$

### 12.4.2 The MART in the General Case

Taking the entries of the matrix $A$ to be either one or zero, depending on whether or not the $j$-th pixel is in the set $L_i$, is too crude. The line $L_i$ may just clip a corner of one pixel, but pass through the center of another. Surely, it makes more sense to let $A_{ij}$ be the length of the intersection of line $L_i$ with the $j$-th pixel, or, perhaps, this length divided by the length of the diagonal of the pixel. It may also be more realistic to consider a strip, instead of a line. Other modifications to $A_{ij}$ may made made, in order to better describe the physics of the situation. Finally, all we can be sure of is that $A_{ij}$ will be nonnegative, for each $i$ and $j$. In such cases, what is the proper form for the MART?

The MART, which can be applied only to nonnegative systems, is a sequential, or row-action, method that uses one equation only at each step of the iteration.

**Algorithm 12.1 (MART)** *Let $x^0$ be any positive vector, and $i = k(\mathrm{mod}\, I)+1$. Having found $x^k$ for positive integer $k$, define $x^{k+1}$ by*

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i}\Big)^{m_i^{-1}A_{ij}}, \tag{12.17}$$

*where $m_i = \max\{A_{ij} \,|j = 1, 2, ..., J\}$.*

Some treatments of MART leave out the $m_i$, but require only that the entries of $A$ have been rescaled so that $A_{ij} \leq 1$ for all $i$ and $j$. The $m_i$ is important, however, in accelerating the convergence of MART.

### 12.4.3 Cross-Entropy

For $a > 0$ and $b > 0$, let the cross-entropy or Kullback-Leibler distance from $a$ to $b$ be

$$KL(a, b) = a \log \frac{a}{b} + b - a, \tag{12.18}$$

with $KL(a, 0) = +\infty$, and $KL(0, b) = b$. Extend to nonnegative vectors coordinate-wise, so that

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{12.19}$$

Unlike the Euclidean distance, the KL distance is not symmetric; $KL(Ax, b)$ and $KL(b, Ax)$ are distinct, and we can obtain different approximate solutions of $Ax = b$ by minimizing these two distances with respect to nonnegative $x$.

### 12.4.4 Convergence of MART

In the consistent case, by which we mean that $Ax = b$ has nonnegative solutions, we have the following convergence theorem for MART.

**Theorem 12.2** *In the consistent case, the MART converges to the unique nonnegative solution of $b = Ax$ for which the distance $\sum_{j=1}^{J} KL(x_j, x_j^0)$ is minimized.*

If the starting vector $x^0$ is the vector whose entries are all one, then the MART converges to the solution that maximizes the Shannon entropy,

$$SE(x) = \sum_{j=1}^{J} x_j \log x_j - x_j. \tag{12.20}$$

As with ART, the speed of convergence is greatly affected by the ordering of the equations, converging most slowly when consecutive equations correspond to nearly parallel hyperplanes.

**Open Question:** When there are no nonnegative solutions, MART does not converge to a single vector, but, like ART, is always observed to produce a limit cycle of vectors. Unlike ART, there is no proof of the existence of a limit cycle for MART.

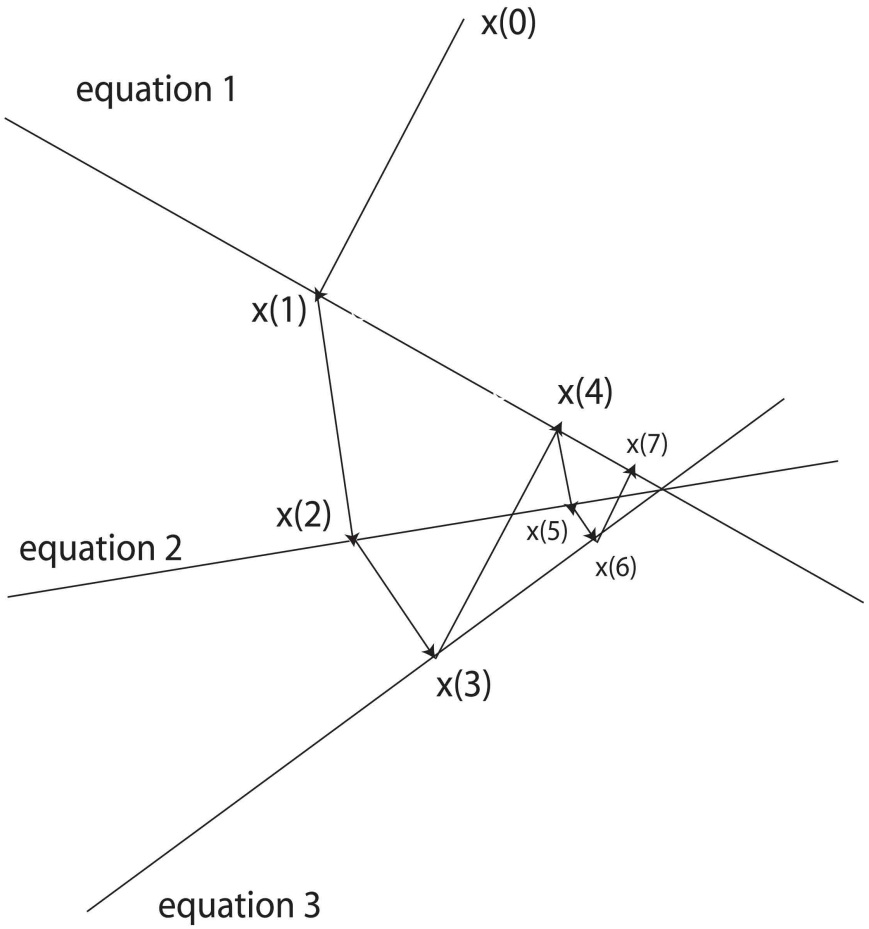



Figure 12.2: The ART algorithm in the consistent case.

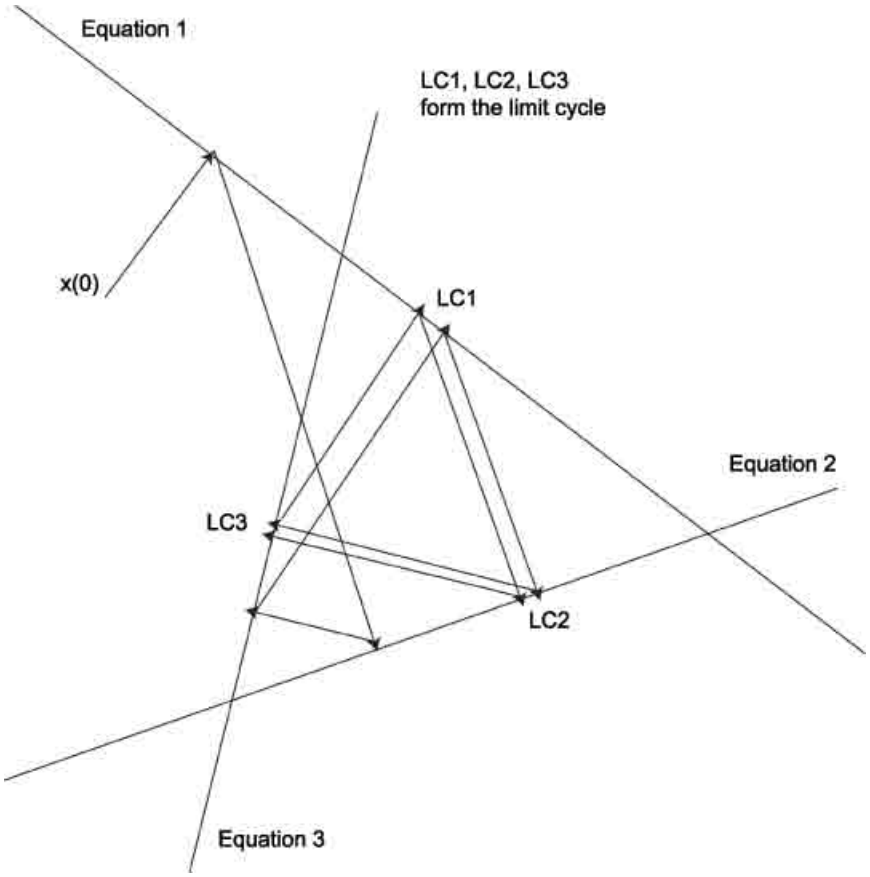



Figure 12.3: The ART algorithm in the inconsistent case.

# Chapter 13

# Some Linear Algebra (Chapter 15)

Linear algebra is the study of linear transformations between vector spaces. Although the subject is not simply matrix theory, there is a close connection, stemming from the role of matrices in representing linear transformations. Throughout this section we shall limit discussion to finite-dimensional vector spaces.

## 13.1 Matrix Algebra

If $A$ and $B$ are real or complex $M$ by $N$ and $N$ by $K$ matrices, respectively, then the product $C = AB$ is defined as the $M$ by $K$ matrix whose entry $C_{mk}$ is given by

$$C_{mk} = \sum_{n=1}^{N} A_{mn} B_{nk}. \tag{13.1}$$

If $x$ is an $N$-dimensional column vector, that is, $x$ is an $N$ by 1 matrix, then the product $b = Ax$ is the $M$-dimensional column vector with entries

$$b_m = \sum_{n=1}^{N} A_{mn} x_n. \tag{13.2}$$

**Exercise 13.1** *Show that, for each $k = 1, ..., K$, $\mathrm{Col}_k(C)$, the $k$th column of the matrix $C = AB$, is*

$$\mathrm{Col}_k(C) = A\mathrm{Col}_k(B).$$

It follows from this exercise that, for given matrices $A$ and $C$, every column of $C$ is a linear combination of the columns of $A$ if and only if there is a third matrix $B$ such that $C = AB$.

The matrix $A^\dagger$ is the *conjugate transpose* of the matrix $A$, that is, the $N$ by $M$ matrix whose entries are

$$(A^\dagger)_{nm} = \overline{A_{mn}} \tag{13.3}$$

When the entries of $A$ are real, $A^\dagger$ is just the *transpose* of $A$, written $A^T$.

**Exercise 13.2** *Let $C = AB$. Show that $B^\dagger A^\dagger = C^\dagger$.*

## 13.2 Linear Independence and Bases

As we shall see shortly, the *dimension* of a *finite-dimensional* vector space will be defined as the number of members of any basis. Obviously, we first need to see what a basis is, and then to convince ourselves that if a vector space $V$ has a basis with $N$ members, then every basis for $V$ has $N$ members.

**Definition 13.1** *The* span *of a collection of vectors $\{u^1, ..., u^N\}$ in $V$ is the set of all vectors $x$ that can be written as linear combinations of the $u^n$; that is, for which there are scalars $c_1, ..., c_N$, such that*

$$x = c_1 u^1 + ... + c_N u^N. \tag{13.4}$$

**Definition 13.2** *A collection of vectors $\{w^1, ..., w^N\}$ in $V$ is called a* spanning set *for a subspace $S$ if the set $S$ is their span.*

**Definition 13.3** *A subset $S$ of a vector space $V$ is called* finite dimensional *if it is contained in the span of a finite set of vectors from $V$.*

This definition tells us what it means to be finite dimensional, but does not tell us what *dimension* means, nor what the actual dimension of a finite dimensional subset is; for that we need the notions of *linear independence* and *basis*.

**Definition 13.4** *A collection of vectors $\{u^1, ..., u^N\}$ in $V$ is* linearly independent *if there is no choice of scalars $\alpha_1, ..., \alpha_N$, not all zero, such that*

$$0 = \alpha_1 u^1 + ... + \alpha_N u^N. \tag{13.5}$$

**Exercise 13.3** *Show that the following are equivalent:*

- *1. the set $\mathcal{U} = \{u^1, ..., u^N\}$ is linearly independent;*
- *2. no $u^n$ is a linear combination of the other members of $\mathcal{U}$;*

- *3. $u^1 \neq 0$ and no $u^n$ is a linear combination of the members of $\mathcal{U}$ that precede it in the list.*

**Definition 13.5** *A collection of vectors $\mathcal{U} = \{u^1, ..., u^N\}$ in $V$ is called a* basis *for a subspace $S$ if the collection is linearly independent and $S$ is their span.*

**Exercise 13.4** *Show that*

- *1. if $\mathcal{U} = \{u^1, ..., u^N\}$ is a spanning set for $S$, then $\mathcal{U}$ is a basis for $S$ if and only if, after the removal of any one member, $\mathcal{U}$ is no longer a spanning set; and*
- *2. if $\mathcal{U} = \{u^1, ..., u^N\}$ is a linearly independent set in $S$, then $\mathcal{U}$ is a basis for $S$ if and only if, after including in $\mathcal{U}$ any new member from $S$, $\mathcal{U}$ is no longer linearly independent.*

## 13.3 Dimension

We turn now to the task of showing that every basis for a finite dimensional vector space has the same number of members. That number will then be used to define the dimension of that subspace.

Suppose that $S$ is a subspace of $V$, that $\{w^1, ..., w^N\}$ is a spanning set for $S$, and $\{u^1, ..., u^M\}$ is a linearly independent subset of $S$. Beginning with $w_1$, we augment the set $\{u^1, ..., u^M\}$ with $w^j$ if $w^j$ is not in the span of the $u^m$ and the $w^k$ previously included. At the end of this process, we have a linearly independent spanning set, and therefore, a basis, for $S$ (Why?). Similarly, beginning with $w^1$, we remove $w^j$ from the set $\{w^1, ..., w^N\}$ if $w^j$ is a linear combination of the $w^k$, $k = 1, ..., j-1$. In this way we obtain a linearly independent set that spans $S$, hence another basis for $S$. The following lemma will allow us to prove that all bases for a subspace $S$ have the same number of elements.

**Lemma 13.1** *Let $G = \{w^1, ..., w^N\}$ be a spanning set for a subspace $S$ in $\mathbb{R}^I$, and $H = \{v^1, ..., v^M\}$ a linearly independent subset of $S$. Then $M \leq N$.*

**Proof:** Suppose that $M > N$. Let $B_0 = G = \{w^1, ..., w^N\}$. To obtain the set $B_1$, form the set $C_1 = \{v^1, w^1, ..., w^N\}$ and remove the first member of $C_1$ that is a linear combination of members of $C_1$ that occur to its left in the listing; since $v^1$ has no members to its left, it is not removed. Since $G$ is a spanning set, $v^1 \neq 0$ is a linear combination of the members of $G$, so that some member of $G$ is a linear combination of $v^1$ and the members of $G$ that precede it in the list; remove the first member of $G$ for which this is true.

We note that the set $B_1$ is a spanning set for $S$ and has $N$ members. Having obtained the spanning set $B_k$, with $N$ members and whose first $k$ members are $v^k, ..., v^1$, we form the set $C_{k+1} = B_k \cup \{v^{k+1}\}$, listing the members so that the first $k+1$ of them are $\{v^{k+1}, v^k, ..., v^1\}$. To get the set $B_{k+1}$ we remove the first member of $C_{k+1}$ that is a linear combination of the members to its left; there must be one, since $B_k$ is a spanning set, and so $v^{k+1}$ is a linear combination of the members of $B_k$. Since the set $H$ is linearly independent, the member removed is from the set $G$. Continuing in this fashion, we obtain a sequence of spanning sets $B_1, ..., B_N$, each with $N$ members. The set $B_N$ is $B_N = \{v^1, ..., v^N\}$ and $v^{N+1}$ must then be a linear combination of the members of $B_N$, which contradicts the linear independence of $H$. ∎

**Corollary 13.1** *Every basis for a subspace $S$ has the same number of elements.*

**Exercise 13.5** *Let $G = \{w^1, ..., w^N\}$ be a spanning set for a subspace $S$ in $\mathbb{R}^I$, and $H = \{v^1, ..., v^M\}$ a linearly independent subset of $S$. Let $A$ be the $I$ by $M$ matrix whose columns are the vectors $v^m$ and $B$ the $I$ by $N$ matrix whose columns are the $w^n$. Prove that there is an $N$ by $M$ matrix $C$ such that $A = BC$. Prove Lemma 13.1 by showing that, if $M > N$, then there is a non-zero vector $x$ with $Cx = 0$.*

**Definition 13.6** *The* dimension *of a subspace $S$ is the number of elements in any basis.*

**Lemma 13.2** *For any matrix $A$, the maximum number of linearly independent rows equals the maximum number of linearly independent columns.*

**Proof:** Suppose that $A$ is an $I$ by $J$ matrix, and that $K \leq J$ is the maximum number of linearly independent columns of $A$. Select $K$ linearly independent columns of $A$ and use them as the $K$ columns of an $I$ by $K$ matrix $U$. Since every column of $A$ must be a linear combination of these $K$ selected ones, there is a $K$ by $J$ matrix $M$ such that $A = UM$. From $A^T = M^T U^T$ we conclude that every column of $A^T$ is a linear combination of the $K$ columns of the matrix $M^T$. Therefore, there can be at most $K$ linearly independent columns of $A^T$. ∎

**Definition 13.7** *The* rank *of $A$ is the maximum number of linearly independent rows or of linearly independent columns of $A$.*

## 13.4 Representing a Linear Transformation

Let $\mathcal{A} = \{a^1, a^2, ..., a^N\}$ be a basis for the finite-dimensional complex vector space $V$. Now that the basis for $V$ is specified, there is a natural association,

an *isomorphism*, between $V$ and the vector space $\mathbb{C}^N$ of $N$-dimensional column vectors with complex entries. Any vector $v$ in $V$ can be written as

$$v = \sum_{n=1}^{N} \gamma_n a^n. \tag{13.6}$$

The column vector $\gamma = (\gamma_1, ..., \gamma_N)^T$ is uniquely determined by $v$ and the basis $\mathcal{A}$ and we denote it by $\gamma = [v]_{\mathcal{A}}$. Notice that the ordering of the list of members of $\mathcal{A}$ matters, so we shall always assume that the ordering has been fixed.

Let $W$ be a second finite-dimensional vector space, and let $T$ be any linear transformation from $V$ to $W$. Let $\mathcal{B} = \{b^1, b^2, ..., b^M\}$ be a basis for $W$. For $n = 1, ..., N$, let

$$Ta^n = A_{1n}b^1 + A_{2n}b^2 + ... + A_{Mn}b^M. \tag{13.7}$$

Then the $M$ by $N$ matrix $A$ having the $A_{mn}$ as entries is said to *represent* $T$, with respect to the bases $\mathcal{A}$ and $\mathcal{B}$.

**Exercise 13.6** *Show that* $[Tv]_{\mathcal{B}} = A[v]_{\mathcal{A}}$.

**Exercise 13.7** *Suppose that $V$, $W$ and $Z$ are vector spaces, with bases $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively. Suppose also that $T$ is a linear transformation from $V$ to $W$ and $U$ is a linear transformation from $W$ to $Z$. Let $A$ represent $T$ with respect to the bases $\mathcal{A}$ and $\mathcal{B}$, and let $B$ represent $U$ with respect to the bases $\mathcal{B}$ and $\mathcal{C}$. Show that the matrix $BA$ represents the linear transformation $UT$ with respect to the bases $\mathcal{A}$ and $\mathcal{C}$.*

## 13.5 Linear Functionals and Duality

When the second vector space $W$ is just the space $C$ of complex numbers, any linear transformation from $V$ to $W$ is called a *linear functional*. The space of all linear functionals on $V$ is denoted $V^*$ and called the *dual space* of $V$. The set $V^*$ is itself a finite-dimensional vector space, so it too has a dual space, $(V^*)^* = V^{**}$.

**Exercise 13.8** *Show that the dimension of $V^*$ is the same as that of $V$. Hint: let $\mathcal{A} = \{a^1, ..., a^N\}$ be a basis for $V$, and for each $m = 1, ..., N$, let $f^m(a^n) = 0$, if $m \neq n$, and $f^m(a^m) = 1$. Show that the collection $\{f^1, ..., f^N\}$ is a basis for $V^*$.*

There is a natural identification of $V^{**}$ with $V$ itself. For each $v$ in $V$, define $J_v(f) = f(v)$ for each $f$ in $V^*$. Then it is easy to establish that $J_v$ is in $V^{**}$ for each $v$ in $V$. The set $J_V$ of all members of $V^{**}$ of the form $J_v$ for some $v$ is a subspace of $V^{**}$.

**Exercise 13.9** *Show that the subspace $J_V$ has the same dimension as $V^{**}$ itself, so that it must be all of $V^{**}$.*

We shall see later that once $V$ has been endowed with an inner product, there is a simple way to describe every linear functional on $V$: for each $f$ in $V^*$ there is a unique vector $v_f$ in $V$ with $f(v) = \langle v, v_f \rangle$, for each $v$ in $V$. As a result, we have an identification of $V^*$ with $V$ itself.

## 13.6 Linear Operators on $V$

When $W = V$, we say that the linear transformation $T$ is a *linear operator* on $V$. In this case, we can also take the basis $\mathcal{B}$ to be $\mathcal{A}$, and say that the matrix $A$ represents the linear operator $T$, with respect to the basis $\mathcal{A}$. We then write $A = [T]_{\mathcal{A}}$.

**Exercise 13.10** *Suppose that $\mathcal{B}$ is a second basis for $V$. Show that there is a unique $N$ by $N$ matrix $Q$ having the property that the matrix $B = QAQ^{-1}$ represents $T$, with respect to the basis $\mathcal{B}$; that is, we can write*

$$[T]_{\mathcal{B}} = Q[T]_{\mathcal{A}}Q^{-1}.$$

*Hint: The matrix $Q$ is the change-of-basis matrix, satisfying*

$$[v]_{\mathcal{B}} = Q[v]_{\mathcal{A}},$$

*for all $v$.*

## 13.7 Diagonalization

Let $T : V \to V$ be a linear operator, $\mathcal{A}$ a basis for $V$, and $A = [T]_{\mathcal{A}}$. As we change the basis, the matrix representing $T$ also changes. We wonder if it is possible to find some basis $\mathcal{B}$ such that $B = [T]_{\mathcal{B}}$ is a diagonal matrix $L$. Let $P = [I]^{\mathcal{A}}_{\mathcal{B}}$ be the change-of basis matrix from $\mathcal{B}$ to $\mathcal{A}$. We would then have $P^{-1}AP = L$, or $A = PLP^{-1}$. When this happens, we say that $A$ has been *diagonalized* by $P$.

Suppose that the basis $\mathcal{B} = \{b^1, ..., b^N\}$ is such that $B = [T]_{\mathcal{B}} = L$, where $L$ is the diagonal matrix $L = \text{diag}\,\{\lambda_1, ..., \lambda_N\}$. Then we have $AP = PL$, which tells us that $p^n$, the $n$-th column of $P$, is an eigenvector of the matrix $A$, with $\lambda_n$ as its eigenvalue. Since $p^n = [b^n]_{\mathcal{A}}$, we have

$$0 = (A - \lambda_n I)p^n = (A - \lambda_n I)[b^n]_{\mathcal{A}} = [(T - \lambda_n I)b^n]_{\mathcal{A}},$$

from which we conclude that

$$(T - \lambda_n I)b^n = 0,$$

or

$$Tb^n = \lambda_n b^n;$$

therefore, $b^n$ is an eigenvector of the linear operator $T$.

## 13.8 Using Matrix Representations

The matrix $A$ has eigenvalues $\lambda_n$, $n = 1, ..., N$ precisely when these $\lambda_n$ are the roots of the *characteristic polynomial*

$$P(\lambda) = \det(A - \lambda I).$$

We would like to be able to define the characteristic polynomial of $T$ itself to be $P(\lambda)$; the problem is that we do not yet know that different matrix representations of $T$ have the same characteristic polynomial.

**Exercise 13.11** *Use the fact that det(GH)=det(G)det(H) for any square matrices G and H to show that*

$$\det([T]_{\mathcal{B}} - \lambda I) = \det([T]_{\mathcal{C}} - \lambda I),$$

*for any bases $\mathcal{B}$ and $\mathcal{C}$ for $V$.*

## 13.9 Matrix Diagonalization and Systems of Linear ODE's

We know that the ordinary linear differential equation

$$x'(t) = ax(t)$$

has the solution

$$x(t) = x(0)e^{at}.$$

In this section we use matrix diagonalization to generalize this solution to systems of linear ordinary differential equations.

Consider the system of linear ordinary differential equations

$$x'(t) = 4x(t) - y(t) \tag{13.8}$$
$$y'(t) = 2x(t) + y(t), \tag{13.9}$$

which we write as $z'(t) = Az(t)$, with

$$A = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix},$$

$$z(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix},$$

and

$$z'(t) = \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix}.$$

We then have

$$\det(A - \lambda I) = (4 - \lambda)(1 - \lambda) + 2 = (\lambda - 2)(\lambda - 3),$$

so the eigenvalues of $A$ are $\lambda = 2$ and $\lambda = 3$.

The vector $u$ given by

$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

solves the system $Au = 2u$ and the vector $v$ given by

$$v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

solves the system $Av = 3v$. Therefore, $u$ and $v$ are linearly independent eigenvectors of $A$. With

$$B = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix},$$

$$B^{-1} = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix},$$

and

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix},$$

we have $A = BDB^{-1}$ and $B^{-1}AB = D$; this is a diagonalization of $A$ using its eigenvalues and eigenvectors.

Note that not every $N$ by $N$ matrix $A$ will have such a diagonalization; we need $N$ linearly independent eigenvectors of $A$, which need not exist. They do exist if the eigenvalues of $A$ are all different, as in the example here, and also if the matrix $A$ is Hermitian or normal. The reader should prove that matrix

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

has no such diagonalization.

Continuing with our example, we let $w(t) = B^{-1}z(t)$ so that $w'(t) = Dw(t)$. Because $D$ is diagonal, this new system is uncoupled;

$$w_1'(t) = 2w_1(t),$$

and

$$w_2'(t) = 3w_2(t).$$

The solutions are then

$$w_1(t) = w_1(0)e^{2t},$$

and

$$w_2(t) = w_2(0)e^{3t}.$$

It follows from $z(t) = Bw(t)$ that

$$x(t) = w_1(0)e^{2t} + w_2(0)e^{3t},$$

and

$$y(t) = 2w_1(0)e^{2t} + w_2(0)e^{3t}.$$

We want to express $x(t)$ and $y(t)$ in terms of $x(0)$ and $y(0)$. To do this we use $z(0) = Bw(0)$, which tells us that

$$x(t) = (-x(0) + y(0))e^{2t} + (2x(0) - y(0))e^{3t},$$

and

$$y(t) = (-2x(0) + 2y(0))e^{2t} + (2x(0) - y(0))e^{3t}.$$

We can rewrite this as

$$z(t) = E(t)z(0),$$

where

$$E(t) = \begin{bmatrix} -e^{2t} + 2e^{3t} & e^{2t} - e^{3t} \\ -2e^{2t} + 2e^{3t} & 2e^{2t} - e^{3t} \end{bmatrix}.$$

What is the matrix $E(t)$?

To mimic the solution $x(t) = x(0)e^{at}$ of the problem $x'(t) = ax(t)$, we try

$$z(t) = e^{tA}z(0),$$

with the matrix exponential defined by

$$e^{tA} = \sum_{n=0}^{\infty} \frac{1}{n!} t^n A^n.$$

Since $A = BDB^{-1}$, it follows that $A^n = BD^nB^{-1}$, so that

$$e^{tA} = Be^{tD}B^{-1}.$$

Since $D$ is diagonal, we have

$$e^{tD} = \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{3t} \end{bmatrix}.$$

A simple calculation shows that

$$e^{tA} = B \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{3t} \end{bmatrix} B^{-1} = \begin{bmatrix} -e^{2t} + 2e^{3t} & e^{2t} - e^{3t} \\ -2e^{2t} + 2e^{3t} & 2e^{2t} - e^{3t} \end{bmatrix} = E(t).$$

Therefore, the solution of the original system is

$$z(t) = e^{tA} z(0).$$

## 13.10 An Inner Product on $V$

For any two column vectors $x = (x_1, ..., x_N)^T$ and $y = (y_1, ..., y_n)^T$ in $\mathbb{C}^N$, their *complex dot product* is defined by

$$x \cdot y = \sum_{n=1}^{N} x_n \overline{y_n} = y^\dagger x,$$

where $y^\dagger$ is the *conjugate transpose* of the vector $y$, that is, $y^\dagger$ is the row vector with entries $\overline{y_n}$.

The association of the elements $v$ in $V$ with the complex column vector $[v]_\mathcal{A}$ can be used to obtain an *inner product* on $V$. For any $v$ and $w$ in $V$, define

$$\langle v, w \rangle = [v]_\mathcal{A} \cdot [w]_\mathcal{A}, \tag{13.10}$$

where the right side is the ordinary complex dot product in $\mathbb{C}^N$. Once we have an inner product on $V$ we can define the *norm* of a vector in $V$ as $\|v\| = \sqrt{\langle v, v \rangle}$.

**Definition 13.8** *A collection of vectors $\{u^1, ..., u^N\}$ in an inner product space $V$ is called* orthonormal *if $\|u^n\|_2 = 1$, for all $n$, and $\langle u^m, u^n \rangle = 0$, for $m \neq n$.*

Note that, with respect to this inner product, the basis $\mathcal{A}$ becomes an orthonormal basis.

We assume, throughout the remainder of this section, that $V$ is an *inner-product space*. For more detail concerning inner products, see the chapter *Appendix: Inner Products and Orthogonality*.

## 13.11 Representing Linear Functionals

Let $f : V \to \mathbb{C}$ be a linear functional on the inner-product space $V$ and let $\mathcal{A} = \{a^1, ..., a^N\}$ be the basis for $V$ used to define the inner product, as in Equation (13.10). The singleton set $\{1\}$ is a basis for the space $W = \mathbb{C}$,

and the matrix $A$ that represents $T = f$ is a 1 by $N$ matrix, or row vector, $A = A_f$ with entries $f(a^n)$. Therefore, for each

$$v = \sum_{n=1}^{N} \alpha_n a^n,$$

in $V$, we have

$$f(v) = A_f[v]_{\mathcal{A}} = \sum_{n=1}^{N} f(a^n)\alpha_n.$$

Consequently, we can write

$$f(v) = \langle v, y_f \rangle,$$

for the vector $y_f$ with $A_f = [y_f]^{\dagger}_{\mathcal{A}}$, or

$$y_f = \sum_{n=1}^{N} \overline{f(a^n)} a^n.$$

So we see that once $V$ has been given an inner product, each linear functional $f$ on $V$ can be thought of as corresponding to a vector $y_f$ in $V$, so that

$$f(v) = \langle v, y_f \rangle.$$

**Exercise 13.12** *Show that the vector $y_f$ associated with the linear functional $f$ is unique by showing that*

$$\langle v, y \rangle = \langle v, w \rangle,$$

*for every $v$ in $V$ implies that $y = w$.*

## 13.12 The Adjoint of a Linear Transformation

Let $T : V \to W$ be a linear transformation from a vector space $V$ to a vector space $W$. The *adjoint* of $T$ is the linear operator $T^* : W^* \to V^*$ defined by

$$(T^* g)(v) = g(Tv), \tag{13.11}$$

for each $g \in W^*$ and $v \in V$.

Once $V$ and $W$ have been given inner products, and $V^*$ and $W^*$ have been identified with $V$ and $W$, respectively, the operator $T^*$ can be defined as a linear operator from $W$ to $V$ as follows. Let $T : V \to W$ be a linear

transformation from an inner-product space $V$ to an inner-product space $W$. For each fixed $w$ in $W$, define a linear functional $f$ on $V$ by

$$f(v) = \langle Tv, w\rangle.$$

By our earlier discussion, $f$ has an associated vector $y_f$ in $V$ such that

$$f(v) = \langle v, y_f\rangle.$$

Therefore,

$$\langle Tv, w\rangle = \langle v, y_f\rangle,$$

for each $v$ in $V$. The *adjoint* of $T$ is the linear transformation $T^*$ from $W$ to $V$ defined by $T^*w = y_f$.

When $W = V$, and $T$ is a linear operator on $V$, then so is $T^*$. In this case, we can ask whether or not $T^*T = TT^*$, that is, whether or not $T$ is *normal*, and whether or not $T = T^*$, that is, whether or not $T$ is *self-adjoint*.

## 13.13 Orthogonality

Two vectors $v$ and $w$ in the inner-product space $V$ are said to be *orthogonal* if $\langle v, w\rangle = 0$. A basis $\mathcal{U} = \{u^1, u^2, ..., u^N\}$ is called an *orthogonal basis* if every two vectors in $\mathcal{U}$ are orthogonal, and *orthonormal* if, in addition, $\|u^n\| = 1$, for each $n$.

**Exercise 13.13** *Let $\mathcal{U}$ and $\mathcal{V}$ be orthonormal bases for the inner-product space $V$, and let $Q$ be the change-of-basis matrix satisfying*

$$[v]_{\mathcal{U}} = Q[v]_{\mathcal{V}}.$$

*Show that $Q^{-1} = Q^\dagger$, so that $Q$ is a unitary matrix.*

**Exercise 13.14** *Let $\mathcal{U}$ be an orthonormal basis for the inner-product space $V$ and $T$ a linear operator on $V$. Show that*

$$[T^*]_{\mathcal{U}} = ([T]_{\mathcal{U}})^\dagger. \tag{13.12}$$

## 13.14 Normal and Self-Adjoint Operators

Let $T$ be a linear operator on an inner-product space $V$. We say that $T$ is *normal* if $T^*T = TT^*$, and *self-adjoint* if $T^* = T$. A square matrix $A$ is said to be *normal* if $A^\dagger A = AA^\dagger$, and *Hermitian* if $A^\dagger = A$.

**Exercise 13.15** *Let $\mathcal{U}$ be an orthonormal basis for the inner-product space $V$. Show that $T$ is normal if and only if $[T]_{\mathcal{U}}$ is a normal matrix, and $T$ is self-adjoint if and only if $[T]_{\mathcal{U}}$ is Hermitian. Hint: use Exercise (13.7).*

**Exercise 13.16** *Compute the eigenvalues for the real square matrix*

$$A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}. \tag{13.13}$$

*Note that the eigenvalues are complex, even though the entries of $A$ are real. The matrix $A$ is not Hermitian.*

**Exercise 13.17** *Show that the eigenvalues of the complex matrix*

$$B = \begin{bmatrix} 1 & 2+i \\ 2-i & 1 \end{bmatrix} \tag{13.14}$$

*are the real numbers $\lambda = 1 + \sqrt{5}$ and $\lambda = 1 - \sqrt{5}$, with corresponding eigenvectors $u = (\sqrt{5}, 2-i)^T$ and $v = (\sqrt{5}, i-2)^T$, respectively.*

**Exercise 13.18** *Show that the eigenvalues of the real matrix*

$$C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{13.15}$$

*are both equal to one, and that the only eigenvectors are non-zero multiples of the vector $(1,0)^T$. Compute $C^TC$ and $CC^T$. Are they equal?*

## 13.15 It is Good to be "Normal"

For a given linear operator, when does there exist an orthonormal basis for $V$ consisting of eigenvectors of $T$? The answer is: When $T$ is normal.

Consider an $N$ by $N$ matrix $A$. We use $A$ to define a linear operator $T$ on the space of column vectors $V = \mathbb{C}^N$ by $Tv = Av$, that is, the operator $T$ works by multiplying each column vector $v$ in $\mathbb{C}^N$ by the matrix $A$. Then $A$ represents $T$ with respect to the usual orthonormal basis $\mathcal{A}$ for $\mathbb{C}^N$. Suppose now that there is an orthonormal basis $\mathcal{U} = \{u^1, ..., u^N\}$ for $\mathbb{C}^N$ such that

$$Au^n = \lambda_n u^n,$$

for each $n$. The matrix representing $T$ in the basis $\mathcal{U}$ is the matrix $B = Q^{-1}AQ$, where $Q$ is the change-of-basis matrix with

$$Q[v]_{\mathcal{U}} = [v]_{\mathcal{A}}.$$

But we also know that $B$ is the diagonal matrix $B = L = \text{diag}(\lambda_1, ..., \lambda_N)$. Therefore, $L = Q^{-1}AQ$, or $A = QLQ^{-1}$.

As we saw in Exercise (13.13), the matrix $Q$ is unitary, that is, $Q^{-1} = Q^\dagger$. Therefore, $A = QLQ^\dagger$. Then we have

$$A^\dagger A = QL^\dagger Q^\dagger QLQ^\dagger = QL^\dagger LQ^\dagger$$

$$= QLL^\dagger Q^\dagger = QLQ^\dagger QL^\dagger Q^\dagger = AA^\dagger,$$

so that

$$A^\dagger A = AA^\dagger,$$

and $A$ is normal.

Two fundamental results in linear algebra are the following.

**Theorem 13.1** *For a linear operator $T$ on a finite-dimensional complex inner-product space $V$ there is an orthonormal basis of eigenvectors if and only if $T$ is normal.*

**Corollary 13.2** *A self-adjoint linear operator $T$ on a finite-dimensional complex inner-product space $V$ has an orthonormal basis of eigenvectors.*

**Exercise 13.19** *Show that the eigenvalues of a self-adjoint linear operator $T$ on a finite-dimensional complex inner-product space are real numbers. Hint: consider $Tu = \lambda_1 u$, and begin with $\lambda\langle u, u\rangle = \langle Tu, u\rangle$.*

Combining the various results obtained so far, we can conclude the following.

**Corollary 13.3** *Let $T$ be a linear operator on a finite-dimensional real inner-product space $V$. Then $V$ has an orthonormal basis consisting of eigenvectors of $T$ if and only if $T$ is self-adjoint.*

We present a proof of the following theorem.

**Theorem 13.2** *For a linear operator $T$ on a finite-dimensional complex inner-product space $V$ there is an orthonormal basis of eigenvectors if and only if $T$ is normal.*

We saw previously that if $V$ has an orthonormal basis of eigenvectors of $T$, then $T$ is a normal operator. We need to prove the converse: if $T$ is normal, then $V$ has an orthonormal basis consisting of eigenvectors of $T$.

A subspace $W$ of $V$ is said to be *$T$-invariant* if $Tw$ is in $W$ whenever $w$ is in $W$. For any $T$-invariant subspace $W$, the restriction of $T$ to $W$, denoted $T_W$, is a linear operator on $W$.

For any subspace $W$, the *orthogonal complement* of $W$ is the space $W^\perp = \{v | \langle w, v\rangle = 0, \text{for all } w \in W\}$.

**Proposition 13.1** *Let $W$ be a $T$-invariant subspace of $V$. Then*

- *(a) if $T$ is self-adjoint, so is $T_W$;*
- *(b) $W^\perp$ is $T^*$-invariant;*
- *(c) if $W$ is both $T$- and $T^*$-invariant, then $(T_W)^* = (T^*)_W$;*

- *(d) if $W$ is both $T$- and $T^*$-invariant, and $T$ is normal, then $T_W$ is normal.*
- *(e) if $T$ is normal and $Tx = \lambda x$, then $T^*x = \overline{\lambda}x$.*

**Exercise 13.20** *Prove Proposition (13.1).*

**Proposition 13.2** *If $T$ is normal, $Tu^1 = \lambda_1 u^1$, $Tu^2 = \lambda_2 u^2$, and $\lambda_1 \neq \lambda_2$, then $\langle u^1, u^2 \rangle = 0$.*

**Exercise 13.21** *Prove Proposition 13.2. Hint: use (e) of Proposition 13.1.*

**Proof of Theorem 13.2** The proof is by induction on the dimension of the inner-product space $V$. To begin with, let $N = 1$, so that $V$ is simply the span of some unit vector $x$. Then any linear operator $T$ on $V$ has $Tx = \lambda x$, for some $\lambda$, and the set $\{x\}$ is an orthonormal basis for $V$.

Now suppose that the theorem is true for every inner-product space of dimension $N - 1$. We know that every linear operator $T$ on $V$ has at least one eigenvector, say $x^1$, since its characteristic polynomial has at least one distinct eigenvalue $\lambda_1$ in $C$. Take $x^1$ to be a unit vector. Let $W$ be the span of the vector $x^1$, and $W^\perp$ the orthogonal complement of $W$. Since $Tx^1 = \lambda_1 x^1$ and $T$ is normal, we know that $T^*x^1 = \overline{\lambda_1}x^1$. Therefore, both $W$ and $W^\perp$ are $T$- and $T^*$-invariant. Therefore, $T_{W^\perp}$ is normal on $W^\perp$. By the induction hypothesis, we know that $W^\perp$ has an orthonormal basis consisting of $N - 1$ eigenvectors of $T_W$, and, therefore, of $T$. Augmenting this set with the original $x^1$, we get an orthonormal basis for all of $V$. ∎

**Corollary 13.4** *A self-adjoint linear operator $T$ on a finite-dimensional complex inner-product space $V$ has an orthonormal basis of eigenvectors.*

**Corollary 13.5** *Let $T$ be a linear operator on a finite-dimensional real inner-product space $V$. Then $V$ has an orthonormal basis consisting of eigenvectors of $T$ if and only if $T$ is self-adjoint.*

Proving the existence of the orthonormal basis uses essentially the same argument as the induction proof given earlier. The eigenvalues of a self-adjoint linear operator $T$ on a finite-dimensional complex inner-product space are real numbers. If $T$ be a linear operator on a finite-dimensional real inner-product space $V$ and $V$ has an orthonormal basis $\mathcal{U} = \{u^1, ..., u^N\}$ consisting of eigenvectors of $T$, then we have

$$Tu^n = \lambda_n u^n = \overline{\lambda_n} u^n = T^* u^n,$$

so, since $T = T^*$ on each member of the basis, these operators are the same everywhere, so $T = T^*$ and $T$ is self-adjoint.

We close with an example of a real 2 by 2 matrix $A$ with $A^T A = AA^T$, but with no eigenvectors in $\mathbb{R}^2$. Take $0 < \theta < \pi$ and $A$ to be the matrix

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{13.16}$$

This matrix represents rotation through an angle of $\theta$ in $\mathbb{R}^2$. Its transpose represents rotation through the angle $-\theta$. These operations obviously can be done in either order, so the matrix $A$ is normal. But there is no non-zero vector in $\mathbb{R}^2$ that is an eigenvector. Clearly, $A$ is not symmetric.

# Part II

# Readings for Applied Mathematics II

# Chapter 14

# Vectors (Chapter 5,6)

## 14.1 Real $N$-dimensional Space

A real $N$-dimensional row vector is a list $\mathbf{x} = (x_1, x_2, ..., x_N)$, where each $x_n$ is a real number. In the context of matrix multiplication, we find it convenient to view $\mathbf{x}$ as a column vector or 1 by $N$ matrix; generally, though we shall view $\mathbf{x}$ as a row vector. We denote by $\mathbb{R}^N$ the set of all such $\mathbf{x}$.

## 14.2 Two Roles for Members of $\mathbb{R}^N$

Members of $\mathbb{R}^N$ play two different roles: they can be points in $N$-dimensional space, or they can be directed line segments in $N$-dimensional space. Consider the case of $\mathbb{R}^2$. The graph of the linear equation

$$3x_1 + 2x_2 = 6 \tag{14.1}$$

is a straight line in the plane. A vector $\mathbf{x} = (x_1, x_2)$ is said to be on this graph if Equation (14.1) holds. For example, $\mathbf{x} = (2, 0)$ is on the graph, as is $\mathbf{y} = (0, 3)$; now both $\mathbf{x}$ and $\mathbf{y}$ are viewed as points in the plane. The vector $\mathbf{a} = (3, 2)$, viewed as a directed line segment, is perpendicular to the graph; it is orthogonal to the directed line segment $\mathbf{b} = \mathbf{x} - \mathbf{y} = (2, -3)$ running from $\mathbf{y}$ to $\mathbf{x}$ that lies along the graph. To see this, note that the dot product $\mathbf{a} \cdot \mathbf{b} = 0$. There is no way to tell from the symbols we use which role a member of $\mathbb{R}^N$ is playing at any given moment; we just have to figure it out from the context.

## 14.3 Vector Algebra and Geometry

There are several forms of multiplication associated with vectors in $\mathbb{R}^N$. The simplest is *multiplication of a vector by a scalar*. By *scalar* we mean a real (or sometimes a complex) number. When we multiply the vector $\mathbf{x} = (2, -3, 6, 1)$ by the scalar 4 we get the vector

$$4\mathbf{x} = (8, -12, 24, 4).$$

The *length* of a vector $\mathbf{x}$ in $\mathbb{R}^N$ is

$$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + ... + x_N^2}. \tag{14.2}$$

The *dot product* $\mathbf{x}\cdot\mathbf{y}$ of two vectors $\mathbf{x} = (x_1, x_2, ..., x_N)$ and $\mathbf{y} = (y_1, y_2, ..., y_N)$ in $\mathbb{R}^N$ is defined by

$$\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + ... + x_Ny_N. \tag{14.3}$$

For the cases of $\mathbb{R}^2$ and $\mathbb{R}^3$ we can give geometric meaning to the dot product; the length of $\mathbf{x}$ is $\sqrt{\mathbf{x} \cdot \mathbf{x}}$ and

$$\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}|\, |\mathbf{y}|\, \cos(\theta),$$

where $\theta$ is the angle between $\mathbf{x}$ and $\mathbf{y}$ when they are viewed as directed line segments positioned to have a common beginning point. We see from this that two vectors are perpendicular (or orthogonal) when their dot product is zero.

For $\mathbb{R}^3$ we also have the *cross product* $\mathbf{x} \times \mathbf{y}$, defined by

$$\mathbf{x} \times \mathbf{y} = (x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_2y_3 - x_3y_2). \tag{14.4}$$

When $\mathbf{x}$ and $\mathbf{y}$ are viewed as directed line segments with a common beginning point, the cross product is viewed as a third directed line segment with the same beginning point, perpendicular to both $\mathbf{x}$ and $\mathbf{y}$, and having for its length the area of the parallelogram formed by $\mathbf{x}$ and $\mathbf{y}$. Therefore, if $\mathbf{x}$ and $\mathbf{y}$ are parallel, there is zero area and the cross product is the zero vector. Note that

$$\mathbf{y} \times \mathbf{x} = -\mathbf{x} \times \mathbf{y}. \tag{14.5}$$

From the relationships

$$\mathbf{x} \cdot (\mathbf{y} \times \mathbf{z}) = \mathbf{y} \cdot (\mathbf{z} \times \mathbf{x}) = \mathbf{z} \cdot (\mathbf{x} \times \mathbf{y}) \tag{14.6}$$

we see that

$$\mathbf{x} \cdot (\mathbf{x} \times \mathbf{y}) = \mathbf{y} \cdot (\mathbf{x} \times \mathbf{x}) = 0. \tag{14.7}$$

The dot product and cross product are relatively new additions to the mathematical tool box. They grew out of the 19th century study of *quaternions*.

## 14.4 Complex Numbers

We may think of complex numbers as members of $\mathbb{R}^2$ with extra algebra imposed, mainly the operator of multiplying two complex numbers to get a third complex number. A complex number $z$ can be written several ways:

$$z = (x, y) = x(1, 0) + y(0, 1) = x + yi,$$

where $i$ is the shorthand for the complex number $i = (0, 1)$. With $w = (u, v) = u + vi$ a second complex number, the product $zw$ is

$$zx = (x + yi)(u + vi) = xu + xvi + yui + uvi^2 = (xu - yv, xv + yu) \quad (14.8)$$

which we obtain by defining $i^2 = (0, 1)(0, 1) = (-1, 0) = -1$. The idea of allowing $-1$ to have a square root was used as a trick in the middle ages to solve certain polynomial equations, and was given a solid mathematical foundation in the early part of the 19th century, with the development of the theory of complex-valued functions of a complex variable (complex analysis).

Complex analysis led to amazing new theorems and mathematical tools, but was limited to two dimensions. As complex analysis was developing, the theory of electromagnetism (EM) was beginning to take shape. The EM theory dealt with the physics of three-dimensional space, while complex analysis dealt only with two-dimensional space. What was needed was a three-dimensional version of complex analysis.

## 14.5 Quaternions

It seemed logical that a three-dimensional version of complex analysis would involve objects of the form

$$a + bi + cj,$$

where $a$, $b$, and $c$ are real numbers, $(1, 0, 0) = 1$, $i = (0, 1, 0)$ and $j = (0, 0, 1)$, and $i^2 = j^2 = -1$ now. Multiplying $a + bi + cj$ by $d + ei + fj$ led to the question What are $ij$ and $ji$? The Irish mathematician Hamilton eventually hit on the answer, but it forced the search to move from three-dimensional space to four-dimensional space.

Hamilton discovered that it was necessary to consider objects of the form $a+bi+cj+dk$, where $1 = (1, 0, 0, 0)$, $i = (0, 1, 0, 0)$, $j = (0, 0, 1, 0)$, and $k = (0, 0, 0, 1)$, and $ij = k = -ji$. With the other rules $i^2 = j^2 = k^2 = -1$, $jk = i = -kj$, and $ki = j = -ik$, we get what are called the *quaternions*. For a while in the latter half of the 19th century it was thought that quaternions would be the main tool for studying EM theory, but that was not what happened.

Let $x = a+bi+cj+dk = (a, \mathbf{A})$, where $\mathbf{A} = (b, c, d)$ is viewed as a vector in $\mathbb{R}^3$, and $y = e + fi + gj + hk = (e, \mathbf{B})$, where $\mathbf{B} = (f, g, h)$ is another member of $\mathbb{R}^3$. When we multiply the quaternion $x$ by the quaternion $y$ to get $xy$, we find that $xy = -yx$ and

$$xy = (ae - \mathbf{A} \cdot \mathbf{B}, a\mathbf{B} + e\mathbf{A} + \mathbf{A} \times \mathbf{B}). \tag{14.9}$$

This tells us that quaternion multiplication employs all four of the notions of multiplication that we have encountered previously: ordinary scalar multiplication, multiplication of a vector by a scalar, the dot product, and the cross product. It didn't take people long to realize that it isn't necessary to use quaternion multiplication all the time; just use the dot product when you need it, and the cross product when you need it. Quaternions were demoted to exercises in abstract algebra texts, while the notions of dot product and cross product became essential tools in vector calculus and EM theory.

# Chapter 15

# A Brief History of Electromagnetism (Chapter 5,6)

## 15.1 Who Knew?

Understanding the connections between magnetism and electricity and exploiting that understanding for technological innovation dominated science in the nineteenth century, and yet no one saw it coming. In the index to Butterfield's classic history of the scientific revolution [9], which he locates roughly from 1300 to 1800, the word "electricity" does not appear. Nobody in 1800 could have imagined that, within a hundred years or so, people would live in cities illuminated by electric light, work with machinery driven by electricity, in factories cooled by electric-powered refrigeration, and go home to listen to a radio and talk to neighbors on a telephone. How we got there is the subject of this essay.

These days, we tend to value science for helping us to predict things like hurricanes, and for providing new technology. The scientific activity we shall encounter in this chapter was not a quest for expanded powers and new devices, but a search for understanding; the expanded powers and new devices came later. The truly fundamental advances do not come from focusing on immediate applications, and, anyway, it is difficult to anticipate what applications will become important in the future. Nobody in 1960 thought that people would want a computer in their living room, just as nobody in 1990 wanted a telephone that took pictures.

Electricity, as we now call it, was not completely unknown, of course. In the late sixteenth century, Gilbert, famous for his studies of magnetism, discovered that certain materials, mainly crystals, could be made attractive

by rubbing them with a cloth. He called these materials *electrics*. Among Gilbert's accomplishments was his overturning of the conventional wisdom about magnets, when he showed, experimentally, that magnets *could* still attract nails after being rubbed with garlic. Sometime after Gilbert, electrostatic repulsion and induction were discovered, making the analogy with magnetism obvious. However, until some way was found to study electricity in the laboratory, the mysteries of electricity would remain hidden and its importance unappreciated.

## 15.2 "What's Past is Prologue"

The history of science is important not simply for its own sake, but as a bridge connecting the arts with the sciences. When we study the history of science, we begin to see science as an integral part of the broader quest by human beings to understand themselves and their world. Progress in science comes not only from finding answers to questions, but from learning to ask better questions. The questions we are able to ask, indeed the observations we are able to make, are conditioned by our society, our history, and our intellectual outlook. Science does not exist in a vacuum. As Shakespeare's line, carved into the wall of the National Archives building in Washington, D.C., suggests, the past sets the stage for what comes next, indeed, for what can come next.

## 15.3 Are We There Yet?

We should be careful when we talk about progress, either within science or more generally. Reasonable people can argue about whether or not the development of atomic weapons ought to be called progress. Einstein and others warned, at the beginning of the atomic age, that the emotional and psychological development of human beings had not kept pace with technological development, that we did not have the capacity to control our technology. It does seem that we have a difficult time concerning ourselves, as a society, with problems that will become more serious in the future, preferring instead the motto "I won't be there. You won't be there."

We can certainly agree, though, that science, overall, has led us to a better, even if not complete, understanding of ourselves and our world and to the technology that is capable of providing decent life and health to far more people than in the past. These successes have given science and scientists a certain amount of political power that is not universally welcomed, however. Recent attempts to challenge the status of science within the community, most notably in the debate over creation "science" and evolution, have really been attempts to lessen the political power of science,

not debates within science itself; the decades long attacks on science by the cigarette industry and efforts to weaken the EPA show clearly that it is not only some religious groups that want the political influence of science diminished.

Many of the issues our society will have to deal with in the near future, including nuclear power, terrorism, genetic engineering, energy, climate change, control of technology, space travel, and so on, involve science and demand a more sophisticated understanding of science on the part of the general public. The recent book *Physics for Future Presidents: the Science Behind the Headlines* [36] discusses many of these topics, supposedly as an attempt by the author to educate presidents-to-be, who will be called on to make decisions, to initiate legislation, and to guide the public debate concerning these issues.

History reminds us that progress need not be permanent. The technological expertise and artistic heights achieved by the Romans, even the mathematical sophistication of Archimedes, were essentially lost, at least in the west, for fifteen hundred years.

History also teaches us how unpredictable the future can be, which is, in fact, the underlying theme of this essay. No one in 1800 could have imagined the electrification that transformed society over the nineteenth century, just as no one in 1900 could have imagined Hiroshima and Nagasaki, only a few decades away, let alone the world of today.

## 15.4 Why Do Things Move?

In his famous "The Origins of Modern Science" [9] Butterfield singles out the problem of motion as the most significant intellectual hurdle the human mind has confronted and overcome in the last fifteen hundred years. The ancients had theories of motion, but for Aristotle, as a scientist perhaps more of a biologist than a physicist, motion as change in location was insignificant compared to motion as qualitative change, as, say, when an acorn grows into a tree. The change experienced by the acorn is clearly oriented toward a goal, to make a tree. By focusing on qualitative change, Aristotle placed too much emphasis on the importance of a goal. His idea that even physical motion was change toward a goal, that objects had a "natural" place to which they "sought" to return, infected science for almost two thousand years.

We must not be too quick to dismiss Aristotle's view, however. General relativity asserts that space-time is curved and that clocks slow down where gravity is stronger. Indeed, a clock on the top of the Empire State Building runs slightly faster than one at street level. As Brian Greene puts it,

*Right now, according to these ideas, you are anchored to the floor because your body is trying to slide down an indentation in space (really,*

*spacetime) caused by the earth. In a sense, all objects "want" to age as slowly as possible* [23].

The one instance of motion as change in location whose importance the ancients appreciated was the motion of the heavens. Aristotle (384-322 B.C.) taught the geocentric theory that the heavens move around the earth. Aristarchus of Samos (310-230 B.C.) had a different view; according to Heath [25], "There is not the slightest doubt that Aristarchus was the first to put forward the heliocentric hypothesis." This probably explains why contemporaries felt that Aristarchus should be indicted for impiety. Ptolemy (100-170 A.D.) based his astronomical system of an earth-centered universe on the theories of Aristotle. Because the objects in the heavens, the moon, the planets and the stars, certainly appear to move rapidly, they must be made of an unearthly material, the *quintessence*.

The recent film "Agora" portrays the Alexandrian mathematician and philosopher Hypatia (350-415 A.D.) as an early version of Copernicus, but this is probably anachronistic. Her death at the hands of a Christian mob seems to have had more to do with rivalries among Christian leaders than with her scientific views and her belief in the heliocentric theory.

So things stood until the middle ages. In the fourteenth century the French theologian Nicole Oresme considered the possibility that the earth rotated daily around its own axis [34]. This hypothesis certainly simplified things considerably, and removed the need for the heavens to spin around the earth daily at enormous speeds. But even Oresme himself was hesitant to push this idea, since it conflicted with scripture.

Gradually, natural philosophers, the term used to describe scientists prior to the nineteenth century, began to take a more serious interest in motion as change in location, due, in part, to their growing interest in military matters and the trajectory of cannon balls. Now, motion on earth and motion of the heavenly bodies came to be studied by some of the same people, such as Galileo, and this set the stage for the unified theory of motion due to gravity that would come later, with Newton.

Copernicus' theory of a sun-centered astronomical system, Tycho Brahe's naked-eye observations of the heavens, Kepler's systematizing of planetary motion, the invention of the telescope and its use by Galileo to observe the pock-marked moon and the mini-planetary system of Jupiter, Galileo's study of balls rolling down inclined planes, and finally Newton's Law of Universal Gravitation marked a century of tremendous progress in the study of motion and put mechanics at the top of the list of scientific paradigms for the next century. Many of the theoretical developments of the eighteenth century involved the expansion of Newton's mechanics to ever more complex systems, so that, by the end of that century, celestial mechanics and potential theory were well developed mathematical subjects.

As we shall see, the early development of the field we now call electromagnetism involved little mathematics. As the subject evolved, the

mathematics of potential theory, borrowed from the study of gravitation and celestial mechanics, was combined with the newly discovered vector calculus and the mathematical treatment of heat propagation to give the theoretical formulation of electromagnetism familiar to us today.

## 15.5 Go Fly a Kite!

The ancients knew about magnets and used them as compasses. Static electricity was easily observed and thought to be similar to magnetism. As had been known for centuries, static electricity exhibited both attraction and repulsion. For that reason, it was argued that there were two distinct types of electricity. Benjamin Franklin opposed this idea, insisting instead on two types of charge, positive and negative. Some progress was made in capturing electricity for study with the invention of the *Leyden jar*, a device for storing relatively large electrostatic charge (and giving rather large shocks). The discharge from the Leyden jar reminded Franklin of lightning and prompted him and others to fly kites in thunderstorms and to discover that lightning would charge a Leyden jar; lightning was electricity. These experiments led to his invention of the lightning rod, a conducting device attached to houses to direct lightning strikes down to the ground.

The obvious analogies with magnetism had been noticed by Gilbert and others in the late sixteenth century, and near the end of the eighteenth century Coulomb found that both magnetic and electrical attraction fell off as the square of the distance, as did gravity, according to Newton. Indeed, the physical connection between magnetism and gravity seemed more plausible than one between magnetism and electricity, and more worth studying. But things were about to change.

## 15.6 Bring in the Frogs!

In 1791 Galvani observed that a twitching of the muscles of a dead frog he was dissecting seemed to be caused by sparks from a nearby discharge of a Leyden jar. He noticed that the sparks need not actually touch the muscles, provided a metal scalpel touched the muscles at the time of discharge. He also saw twitching muscles when the frog was suspended by brass hooks on an iron railing in a thunderstorm. Eventually, he realized that the Leyden jar and thunderstorm played no essential roles; two scalpels of different metals touching the muscles were sufficient to produce the twitching. Galvani concluded that the electricity was in the muscles; it was *animal electricity*.

Believing that the electricity could be within the animals is not as farfetched as it may sound. It was known at the time that there were certain "electric" fish that generated their own electricity and used it to attack

their prey. When these animals were dissected, it was noticed that there were unusual structures within their bodies that other fish did not have. Later, it became clear that these structures were essentially batteries.

## 15.7 Lose the Frogs!

In 1800 Volta discovered that electricity could be produced by two dissimilar metals, copper and zinc, say, in salt water; no animal electricity here, and no further need for the frogs. He had discovered the *battery* and introduced *electrodynamics*. His primitive batteries, eventually called *voltaic piles*, closely resembled the electricity-producing structures found within the bodies of "electric" fish. Only six weeks after Volta's initial report, Nicholson and Carlisle discovered *electrolysis*, the loosening up and separating of distinct atoms in molecules, such as the hydrogen and oxygen atoms in water.

The fact that chemical reactions produced electric currents suggested the reverse, that electrical currents could stimulate chemical reactions; this is *electrochemistry*, which led to the discovery and isolation of many new elements in the decades that followed. In 1807 Humphry Davy isolated some active metals from their liquid compounds and became the first to form sodium, potassium, calcium, strontium, barium, and magnesium.

In 1821 Seebeck found that the electric current would continue as long as the temperatures of the two metals were kept different; this is *thermoelectricity* and provides the basis for the *thermocouple*, which could then be used as a thermometer.

## 15.8 It's a Magnet!

In 1819 Oersted placed a current-carrying wire over a compass, not expecting anything in particular to happen. The needle turned violently perpendicular to the axis of the wire. When Oersted reversed the direction of the current, the needle jerked around 180 degrees. This meant that magnetism and electricity were not just analogous, but intimately related; *electromagnetism* was born. Soon after, Arago demonstrated that a wire carrying an electric current behaved like a magnet. Ampere, in 1820, confirmed that a wire carrying a current *was* a magnet by demonstrating attraction and repulsion between two separate current-carrying wires. He also experimented with wires in various configurations and related the strength of the magnetic force to the strength of the current in the wire. This connection between electric current and magnetism led fairly soon after to the telegraph, and later in the century, to the telephone.

## 15.9 A New World

Electric currents produce magnetism. But can magnets produce electric currents? Can the relationship be reversed? In 1831, Michael Faraday tried to see if a current would be produced in a wire if it was placed in a magnetic field created by another current-carrying wire. The experiment failed, sort of. When the current was turned on in the second wire, generating the magnetic field, the first wire experienced a brief current, but then nothing; when the current was turned off, again a brief current in the first wire. Faraday, an experimental genius who, as a young man, had been an assistant to Davy, and later the inventor of the refrigerator, made the right conjecture that it is not the mere presence of the magnetic field that causes a current, but changes in that magnetic field. He confirmed this conjecture by showing that a current would flow through a coiled wire when a magnetized rod was moved in and out of the coil; he (and, independently, Henry in the United States) had invented *electromagnetic induction* and the *electric generator* and, like Columbus, had discovered a new world.

## 15.10 Do The Math!

Mathematics has yet to appear in our brief history of electromagnetism, but that was about to change. Although Faraday, often described as being innocent of mathematics, developed his concept of *lines of force* in what we would view as an unsophisticated manner, he was a great scientist and his intuition would prove to be remarkably accurate.

In the summer of 1831, the same summer in which the forty-year old Faraday first observed the phenomenon of electromagnetic induction, the creation of an electric current by a changing magnetic field, James Clerk Maxwell was born in Edinburgh, Scotland.

Maxwell's first paper on electromagnetism, "On Faraday's Lines of Force" , appeared in 1855, when he was about 25 years old. The paper involved a mathematical development of the results of Faraday and others and established the mathematical methods Maxwell would use later in his more famous work "On Physical Lines of Force" .

Although Maxwell did not have available all of the compact vector notation we have today, his work was mathematically difficult. The following is an excerpt from a letter Faraday himself sent to Maxwell concerning this point.

*There is one thing I would be glad to ask you. When a mathematician engaged in investigating physical actions and results has arrived at his conclusions, may they not be expressed in common language as fully, clearly and definitely as in mathematical formulae? If so, would it not be a great boon to such as I to express them so? - translating them out of*

*their hieroglyphics, that we may work upon them by experiment.* Hasn't every beginning student of vector calculus and electromagnetism wished that Maxwell and his followers had heeded Faraday's pleas?

As Zajonc relates in [49], reading Faraday, Maxwell was surprised to find a kindred soul, someone who thought mathematically, although he expressed himself in pictures. Maxwell felt that Faraday's use of "lines of force" to coordinate the phenomena of electromagnetism showed him to be "a mathematician of a very high order" .

Maxwell reasoned that, since an electric current sets up a magnetic field, and a changing magnetic field creates an electrical field, there should be what we now call *electromagnetic waves*, as these two types of fields leapfrog across (empty?) space. These waves would obey partial differential equations, called *Maxwell's equations*, although their familiar form came later and is due to Heaviside [20]. Analyzing the mathematical properties of the resulting wave equations, Maxwell discovered that the propagation speed of these waves was the same as that of light, leading to the conclusion that light itself is an electromagnetic phenomenon, distinguished from other electromagnetic radiation only by its frequency. That light also exhibits behavior more particle-like than wave-like is part of the story of the science of the 20th century.

Maxwell predicted that electromagnetic radiation could exist at various frequencies, not only those associated with visible light. Infrared and ultraviolet radiation had been known since early in the century, and perhaps they too were part of a *spectrum* of electromagnetic radiation. After Maxwell's death from cancer at forty-eight, Hertz demonstrated, in 1888, the possibility of electromagnetic radiation at very low frequencies, *radio waves*. In 1895 Röntgen discovered electromagnetic waves at the high-frequency end of the spectrum, the so-called *x-rays*.

## 15.11 Just Dot the i's and Cross the t's?

By the end of the nineteenth century, some scientists felt that all that was left to do in physics was to dot the i's and cross the t's. However, others saw paradoxes and worried that there were problems yet to be solved; how serious these might turn out to be was not always clear.

Maxwell himself had noted, about 1869, that his work on the specific heats of gases revealed conflicts between rigorous theory and experimental findings that he was unable to explain; it seemed that internal vibration of atoms was being "frozen out"at sufficiently low temperatures, something for which classical physics could not account. His was probably the first suggestion that classical physics could be "wrong". There were also the mysteries, observed by Newton, associated with the partial reflection of light by thick glass. Advances in geology and biology had suggested strongly that the

earth and the sun were much older than previously thought, which was not possible, according to the physics of the day; unless a new form of energy was operating, the sun would have burned out a long time ago.

Newton thought that light was a stream of particles. Others at the time, notably Robert Hooke and Christiaan Huygens, felt that light was a wave phenomenon. Both sides were hindered by a lack of a proper scientific vocabulary to express their views. Around 1800 Young demonstrated that a beam of light displayed interference effects similar to water waves. Eventually, his work convinced people that Newton had been wrong on this point and most accepted that light is a wave phenomenon. Faraday, Maxwell, Hertz and others further developed the wave theory of light and related light to other forms of electromagnetic radiation.

In 1887 Hertz discovered the *photo-electric effect*, later offered by Einstein as confirming evidence that light has a particle nature. When light strikes a metal, it can cause the metal to release an electrically charged particle, an electron. If light were simply a wave, there would not be enough energy in the small part of the wave that hits the metal to displace the electron; in 1905 Einstein will argue that light is *quantized*, that is, it consists of individual bundles or particles, later called *photons*, each with enough energy to cause the electron to be released.

It was recognized that there were other problems with the wave theory of light. All known waves required a medium in which to propagate. Sound cannot propagate in a vacuum; it needs air or water or something. The sound waves are actually compressions and rarefactions of the medium, and how fast the waves propagate depends on how fast the material in the medium can perform these movements; sound travels faster in water than in air, for example.

Light travels extremely fast, but does not propagate instantaneously, as Olaus Roemer first demonstrated around 1700. He observed that the eclipses of the moons of Jupiter appeared to happen sooner when Jupiter was moving closer to Earth, and later when it was moving away. He reasoned, correctly, that the light takes a finite amount of time to travel from the moons to Earth, and when Jupiter is moving away the distance is growing longer.

If light travels through a medium, which scientists called the *ether*, then the ether must be a very strange substance indeed. The material that makes up the ether must be able to compress and expand very quickly. Light comes to us from great distances so the ether must extend throughout all of space. The earth moves around the sun, and therefore through this ether, at a great speed, and yet there are no friction effects, while very much slower winds produce a great deal of weathering. Light can also be polarized, so the medium must be capable of supporting transverse waves, not just longitudinal waves, as in acoustics. To top it all off, the Michelson-Morley experiment, performed in Cleveland in 1887, failed to detect the presence

of the ether. The notion that there is a physical medium that supports the propagation of light would not go away, however. Late in his long life Lord Kelvin (William Thomson) wrote "One word characterizes the most strenuous efforts ... that I have made perseveringly during fifty-five years: that word is FAILURE." Thomson refused to give up his efforts to combine the mathematics of electromagnetism with the mechanical picture of the world.

## 15.12 Seeing is Believing

If radio waves can travel through an invisible ether, and if hypnotists can *mesmerize* their subjects, why can't human beings communicate telepathically with each other and with the dead? Why should atoms exist when we cannot see them, while ghosts must not, even when, as some claimed, they have shown up in photographs? When is seeing believing?

In the late 1800's the experimental physicist William Crooke claimed to have discovered *radiant matter* [14]. When he passed an electric current through a glass tube filled with a low-pressure gas, a small object within the tube could be made to move from one end to the other, driven, so Crooke claimed, by radiant particles of matter, later called *cathode rays*, streaming from one end of the tube to the other. Crooke then went on, without much success, to find material explanation for some of the alleged effects of spiritualism. He felt that it ought to be possible for humans to receive transmissions in much the same way as a radio receives signals. It was a time of considerable uncertainty, and it was not clear that Crooke's radiant matter, atoms, x-rays, radio waves, radioactivity, and the ether were any more real than ghosts, table tapping, and communicating with the dead; they all called into question established physics.

Crooke felt that scientists had a calling to investigate all these mysteries, and should avoid preconceptions about what was true or false. Others accused him of betraying his scientific calling and of being duped by spiritualists. Perhaps remembering that even the word "scientist" was unknown prior to the 1830's, they knew, nevertheless, that, if the history of the nineteenth century taught them anything, it was that there were also serious problems on the horizon of which they were completely unaware.

## 15.13 If You Can Spray Them, They Exist

Up through the seventeenth century, philosophy, especially the works of Aristotle, had colored the way scientists looked at the physical world. By the end of the nineteenth century, most scientists would have agreed that philosophy had been banished from science, that statements that could not be empirically verified, that is, metaphysics, had no place in science. But

philosophy began to sneak back in, as questions about causality and the existence of objects we cannot see, such as atoms, started to be asked [1]. Most scientists are probably *realists*, believing that the objects they study have an existence independent of the instruments used to probe them. On the other side of the debate, *positivists*, or, at least, the more extreme positivists, hold that we have no way of observing an observer-independent reality, and therefore cannot verify that there is such a reality. Positivists hold that scientific theories are simply instruments used to hold together observed facts and make predictions. They do accept that the theories describe an *empirical* reality that is the same for all observers, but not a reality independent of observation. At first, scientists felt that it was safe for them to carry on without worrying too much about these philosophical points, but quantum theory would change things [26].

The idea that matter is composed of very small indivisible *atoms* goes back to the ancient Greek thinkers Democritus and Epicurus. The philosophy of Epicurus was popularized during Roman times by Lucretius, in his lengthy poem *De Rerum Natura* ("On the Nature of Things" ), but this work was lost to history for almost a thousand years. The discovery, in 1417, of a medieval copy of the poem changed the course of history, according to the author Stephen Greenblatt [22]. Copies of the poem became widely distributed throughout Europe and eventually influenced the thinking of Galileo, Freud, Darwin, Einstein, Thomas Jefferson, and many others. But it wasn't until after Einstein's 1905 paper on Brownian motion and subsequent experimental confirmations of his predictions that the actual existence of atoms was more or less universally accepted.

I recall reading somewhere about a conversation between a philosopher of science and an experimental physicist, in which the physicist was explaining how he sprayed an object with positrons. The philosopher then asked him if he really believed that positrons exist. The physicist answered, "If you can spray them, they exist."

## 15.14 What's Going On Here?

Experiments with cathode rays revealed that they were deflected by magnets, unlike any form of radiation similar to light, and unresponsive to gravity. Maybe they were very small electrically charged particles. In 1897 J.J. Thomson established that the cathode rays were, indeed, electrically charged particles, which he called *electrons*. For this discovery he was awarded the Nobel Prize in Physics in 1906. Perhaps there were two fundamental objects in nature, the atoms of materials and the electrons. However, Volta's experiments suggested the electrons were within the materials and involved in chemical reactions. In 1899 Thomson investigated the photo-electric effect and found that cathode rays could be produced

by shining light on certain metals; the photo-electric effect revealed that electrons were inside the materials. Were they between the atoms, or inside the atoms? If they were within the atoms, perhaps their number and configuration could help explain Mendeleev's periodic table and the variety of elements found in nature.

In 1912, Max von Laue demonstrated that Röntgen's x-ray beams can be diffracted; this provided a powerful tool for determining the structure of crystals and molecules and later played an important role in the discovery of the double-helix structure of DNA. In 1923, the French physicist Louis de Broglie suggested that moving particles, such as electrons, should exhibit wave-like properties characterized by a wave-length. In particular, he suggested that beams of electrons sent through a narrow aperture could be diffracted. In 1937 G.P. Thomson, the son of J.J. Thomson, shared the Nobel Prize in Physics with Clinton Davisson for their work demonstrating that beams of electrons can be diffracted. As someone once put it, "The father won the prize for showing that electrons are particles, and the son won it for showing that they aren't." Some suggested that, since beams of electrons exhibited wave-like properties, they should give rise to the sort of interference effects Young had shown were exhibited by beams of light. The first laboratory experiment showing double-slit interference effects of beams of electrons was performed in 1989.

J.J. Thomson also discovered that the kinetic energy of the emitted electrons depended not at all on the intensity of the light, but only on its frequency. This puzzling aspect of the photo-electric effect prompted Einstein to consider the possibility that light is quantized, that is, it comes in small "packages", or *light quanta*, later called *photons*. Einstein proposed quantization of light energy in his 1905 work on the photo-electric effect. It was this work, not his theories of special and general relativity, that eventually won for Einstein the 1921 Nobel Prize in Physics.

Einstein's 1905 paper that deals with the photo-electric effect is really a paper about the particle nature of light. But this idea met with great resistance, and it was made clear to Einstein that his prize was not for the whole paper, but for that part dealing with the photo-electric effect. He was even asked not to mention the particle nature of light in his Nobel speech.

Around 1900 Max Planck had introduced quantization in his derivation of the energy distribution as a function of frequency in black-body radiation. Scholars have suggested that he did this simply for computational convenience, and did not intend, at that moment, to abandon classical physics. Somewhat later Plank and others proposed that the energy might need to be quantized, in order to explain the absence of what Ehrenfest called the *ultraviolet catastrophe* in black-body radiation.

Were the electrons the only sub-atomic particles? No, as Rutherford's discovery of the atomic nucleus in 1911 would reveal. And what is radioac-

tivity, anyway? The new century was dawning, and all these questions were in the air. It was about 1900, Planck had just discovered the quantum theory, Einstein was in the patent office, where he would remain until 1909, Bohr and Schrödinger schoolboys, Heisenberg not yet born. A new scientific revolution was about to occur, and, as in 1800, nobody could have guessed what was coming next [35].

## 15.15 The Year of the Golden Eggs

As Rigden relates in [39], toward the end of his life Einstein looked back to 1905, when he was twenty-six, and told Leo Szilard, "They were the happiest years of my life. Nobody expected me to lay golden eggs." It is appropriate to end our story in 1905 because it was both an end and a beginning. In five great papers published in that year, Einstein solved several of the major outstanding problems that had worried physicists for years, but the way he answered them was revolutionary and began a whole new era of physics. After 1905 the development of electromagnetism merges with that of quantum mechanics, and becomes too big a story to relate here.

The problems that attracted Einstein involved apparent contradictions, and his answers were surprising. Is matter continuous or discrete? It is discrete; atoms do exist. Is light wave-like or particle-like? It is both. Are the laws of thermodynamics absolute or statistical? They are statistical. Are the laws of physics the same for observers moving with uniform velocity relative to one another? Yes; in particular, each will measure the speed of light to be the same. And, by the way, our notion of three-dimensional space and a separate dimension of time is wrong (special relativity), and gravity and acceleration are really the same thing (general relativity). Is inertial mass the same as gravitational mass? Yes. And what is mass, anyway? It is really energy, as $E = mc^2$ tells us.

## 15.16 Do Individuals Matter?

Our brief history of electromagnetism has focused on a handful of extraordinary people. But how important are individuals in the development of science, or in the course of history generally? An ongoing debate among those who study history is over the role of the Great Man [13]. On one side of the debate is the British writer and hero-worshipper Carlyle: "Universal history, the history of what man has accomplished in this world, is at bottom the History of the Great Men who have worked here." On the other side is the German political leader Bismarck: "The statesman's task is to hear God's footsteps marching through history, and to try to catch on to His coattails as He marches past."

If Mozart had never lived, nobody else would have composed his music. If Picasso had never lived, nobody else would have painted his pictures. If Winston Churchill had never lived, or had he died of his injuries when, in 1930, he was hit by a car on Fifth Avenue in New York City, western Europe would probably be different today. If Hitler had died in 1930, when the car he was riding in was hit by a truck, recent history would certainly be different, in ways hard for us to imagine. But, I think the jury is still out on this debate, at least as it applies to science.

I recently came across the following, which I think makes this point well. Suppose that you were forced to decide which one of these four things to"consign to oblivion" , that is, make it never to have happened: Mozart's opera *Don Giovanni*, Chaucer's *Canterbury Tales*, Newton's *Principia*, or Eiffel's tower. Which one would you choose? The answer has to be Newton's *Principia*; it is the only one of the four that is not irreplaceable.

If Newton had never lived, we would still have Leibniz's calculus. Newton's Law of Universal Gravitation would have been discovered by someone else. If Faraday had never lived, we would still have Henry's discovery of electromagnetic induction. If Darwin had never lived, someone else would have published roughly the same ideas, at about the same time; in fact, Alfred Russel Wallace did just that. If Einstein had not lived, somebody else, maybe Poincaré, would have hit on roughly the same ideas, perhaps a bit later. Relativity would have been discovered by someone else. The fact that light behaves both like a wave and like a particle would have become apparent to someone else. The fact that atoms do really exist would have been demonstrated by someone else, although perhaps in a different way.

Nevertheless, just as Mozart's work is unique, even though it was obviously influenced by the times in which he composed and is clearly in the style of the late 18th century, Darwin's view of what he was doing differed somewhat from the view taken by Wallace, and Einstein's work reflected his own fascination with apparent contradiction and a remarkable ability, "to think outside the box" , as the currently popular expression has it. Each of the people we have encountered in this brief history made a unique contribution, even though, had they not lived, others would probably have made their discoveries, one way or another.

People matter in another way, as well. Science is the work of individual people just as art, music and politics are. The book of nature, as some call it, is not easily read. Science is a human activity. Scientists are often mistaken and blind to what their training and culture prevent them from seeing. The history of the development of science is, like all history, our own story.

## 15.17 What's Next?

The twentieth century has taught us that all natural phenomena are based on two physical principles, quantum mechanics and relativity. The combination of special relativity and quantum mechanics led to a unification of three of the four fundamental forces of nature, electromagnetic force and the weak and strong nuclear forces, originally thought to be unrelated. The remaining quest is to combine quantum mechanics with general relativity, which describes gravity. Such a unification seems necessary if one is to solve the mysteries posed by *dark matter* and *dark energy* [6], which make up most of the *stuff* of the universe, but of which nothing is known and whose existence can only be inferred from their gravitational effects. Perhaps what will be needed is a *paradigm shift*, to use Kuhn's popular phrase; perhaps the notion of a *fundamental particle*, or even of an *observer*, will need to be abandoned.

The June 2010 issue of *Scientific American* contains an article called "Twelve events that will change everything". The article identifies twelve events, both natural and man-made, that could happen at any time and would transform society. It also rates the events in terms of how likely they are to occur: fusion energy (very unlikely); extraterrestrial intelligence, nuclear exchange, and asteroid collision (unlikely); deadly pandemic, room-temperature superconductors, and extra dimensions (50-50); cloning of a human, machine self-awareness, and polar meltdown (likely); and creation of life, and Pacific earthquake (almost certain). Our brief study of the history of electromagnetism should convince us that the event that will *really* change everything is not on this list nor on anyone else's list. As Brian Greene suggests [23], people in the year 2100 may look back on today as the time when the first primitive notions of parallel universes began to take shape.

## 15.18 Unreasonable Effectiveness

As Butterfield points out in [9], science became modern in the period 1300 to 1800 not when experiment and observation replaced adherence to the authority of ancient philosophers, but when the experimentation was performed under the control of mathematics. New mathematical tools, logarithms, algebra, analytic geometry, and calculus, certainly played an important role, but so did mathematical thinking, measuring quantities, rather than speculating about qualities, idealizing and abstracting from a physical situation, and the like. Astronomy and mechanics were the first to benefit from this new approach. Paradoxically, our understanding of electromagnetism rests largely on a century or more of intuition, conjecture, experimentation and invention that was almost completely free of math-

ematics. To a degree, this was because the objects of interest, magnets and electricity, were close at hand and, increasingly, available for study. In contrast, Newton's synthesis of terrestrial and celestial gravitation was necessarily largely a mathematical achievement; observational data was available, but experimentation was not possible.

With Maxwell and the mathematicians, electromagnetism became a modern science. Now electromagnetism could be studied with a pencil and paper, as well as with generators. Consequences of the equations could be tested in the laboratory and used to advance technology. The incompleteness of the theory, with regard to the ether, the arrow of time, the finite speed of light, also served to motivate further theoretical and experimental investigation.

As electromagnetism, in particular, and physics, generally, became more mathematical, studies of the very small (nuclear physics), the very large (the universe), and the very long ago (cosmology) became possible. The search for unifying theories of everything became mathematical studies, the consequences of the theories largely beyond observation [43].

One of the great mysteries of science is what the physicist Eugene Wigner called "the unreasonable effectiveness of mathematics". Maxwell's mathematics suggested to him that visible light was an electromagnetic phenomenon, occupying only a small part of an electromagnetic spectrum, and to Hertz that there might be radio waves. Dirac's mathematics suggested to him the existence of anti-matter, positrons with the mass of an electron, but with a positive charge, and with the bizarre property that, when a positron hits an electron, their masses disappear, leaving only energy. What was fantastic science fiction in 1930 is commonplace today, as anyone who has had a positron-emission-tomography (PET) scan is well aware. Mathematics pointed to the existence of the Higgs boson, recently discovered at CERN.

In 2000 the mathematical physicist Ed Witten wrote a paper describing the physics of the century just ending [46]. Even the title is revealing; the quest is for *mathematical* understanding. He points out that, as physics became more mathematical in the first half of the twentieth century, with relativity and non-relativistic quantum mechanics, it had a broad influence on mathematics itself. The equations involved were familiar to the mathematicians of the day, even if the applications were not, and their use in physics prompted further mathematical development, and the emergence of new fields, such as functional analysis. In contrast, the physics of the second half of the century involves mathematics, principally quantum concepts applied to fields, not just particles, the foundations of which are not well understood by mathematicians. This is mathematics with which even the mathematicians are not familiar. Providing a mathematical foundation for the standard model for particle physics should keep the mathematicians of the next century busy for a while. The most interesting sentence in [46]

is *The quest to understand string theory may well prove to be a central theme in physics of the twenty-first century.* Are physicists now just trying to understand their own mathematics, instead of the physical world?

## 15.19 Coming Full Circle

As we have seen, prior to Maxwell, electromagnetism was an experimental science. With the coming of quantum mechanics, it became a mathematical study. Advances came from equations like Dirac's, more than from laboratories.

Within the last couple of decades, however, the circle has begun to close. As scientists began to use computers to study their equations, strange phenomena began to emerge: sensitive dependence on initial conditions in the equations used to study the weather; chaotic behavior of sequences of numbers generated by apparently simple formulas; fractal images appearing when these simple formulas were displayed graphically. At first, it was thought that the strange behavior was coming from numerical errors, but soon similar behavior was observed in natural systems. Chaos theory, complexity and the study of emergent phenomena are the products of computer-driven experimental mathematics.

# Chapter 16

# Changing Variables in Multiple Integrals (Chapter 5,6)

## 16.1 Mean-Value Theorems

In this section we review mean-value theorems for several types of functions.

### 16.1.1 The Single-Variable Case

The mean-value theorem that we learn in Calculus I can be expressed as follows:

**Theorem 16.1** *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable real-valued function of a single real variable. Then for any real numbers $a$ and $b$ there is a third real number $c$ between $a$ and $b$ such that*

$$\Delta f(a) = f(b) - f(a) = f'(c)(b - a) = f'(c)\Delta a. \tag{16.1}$$

When we take $b = a + da$, where $da$ is an infinitesmal, we have

$$df(a) = f'(a)da. \tag{16.2}$$

### 16.1.2 The Multi-variate Case

Now consider a differentiable real-valued function of $J$ real variables, $F : \mathbb{R}^J \to \mathbb{R}$. There is a mean-value theorem for this case, as well.

**Theorem 16.2** *For any* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $\mathbb{R}^J$ *there is* $\mathbf{c}$ *on the line segment between* $\mathbf{a}$ *and* $\mathbf{b}$ *such that*

$$\Delta F(\mathbf{a}) = F(\mathbf{b}) - F(\mathbf{a}) = \nabla F(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a}) = \nabla F(\mathbf{c}) \cdot \Delta \mathbf{a}. \tag{16.3}$$

**Proof:** We prove this mean-value theorem using the previous one. Any point $\mathbf{x}$ on the line segment joining $\mathbf{a}$ with $\mathbf{b}$ has the form

$$\mathbf{x} = \mathbf{a} + t(\mathbf{b} - \mathbf{a}) = (1 - t)\mathbf{a} + t\mathbf{b},$$

for some $t$ in the interval $[0, 1]$. We then define

$$f(t) = F(\mathbf{a} + t(\mathbf{b} - \mathbf{a})). \tag{16.4}$$

The chain rule tells us that

$$f'(t) = \nabla F(\mathbf{a} + t(\mathbf{b} - \mathbf{a})) \cdot (\mathbf{b} - \mathbf{a}). \tag{16.5}$$

Now we apply Equation (16.3) to get

$$\Delta F(\mathbf{a}) = f(1) - f(0) = f'(\tau)(1 - 0)$$

$$= \nabla F(\mathbf{a} + \tau(\mathbf{b} - \mathbf{a})) \cdot (\mathbf{b} - \mathbf{a}) = \nabla F(\mathbf{c}) \cdot \Delta \mathbf{a}, \tag{16.6}$$

where $\mathbf{c} = \mathbf{a} + \tau(\mathbf{b} - \mathbf{a})$. ∎

When $\mathbf{b} - \mathbf{a} = \mathbf{da}$ we can write

$$dF(\mathbf{a}) = F(\mathbf{b}) - F(\mathbf{a}) = \nabla F(\mathbf{a}) \cdot \mathbf{da}. \tag{16.7}$$

### 16.1.3 The Vector-Valued Multi-variate Case

Our objective in this chapter is to examine the rules for change of coordinates when we integrate functions defined on $\mathbb{R}^J$. This leads us to consider functions $\mathbf{r} : \mathbb{R}^J \to \mathbb{R}^J$. We write

$$\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), ..., r_J(\mathbf{x})). \tag{16.8}$$

Each of the functions $r_j$ is a real-valued function of $J$ real variables, so we can apply the mean-value theorem of the previous section, using $F = r_j$. Then we get

$$dr_j(\mathbf{a}) = \nabla r_j(\mathbf{a}) \cdot \mathbf{da}. \tag{16.9}$$

We extend this to $\mathbf{r}$, writing

$$\mathbf{dr}(\mathbf{a}) = (dr_1(\mathbf{a}), ..., dr_J(\mathbf{a})), \tag{16.10}$$

so that the *vector differential* of $\mathbf{r}$ at $\mathbf{a}$ is

$$\mathbf{dr}(\mathbf{a}) = (\nabla r_1(\mathbf{a}), ..., \nabla r_J(\mathbf{a})) \cdot \mathbf{da}. \tag{16.11}$$

Writing $\mathbf{a} = (a_1, ..., a_J)$, $\mathbf{da} = (da_1, ..., da_J)$, and

$$\frac{\partial \mathbf{r}}{\partial a_j}(\mathbf{a}) = (\frac{\partial r_1}{\partial a_j}(\mathbf{a}), ..., \frac{\partial r_J}{\partial a_j}(\mathbf{a})), \tag{16.12}$$

we have

$$\mathbf{dr}(\mathbf{a}) = \sum_{j=1}^{J} \frac{\partial \mathbf{r}}{\partial a_j}(\mathbf{a})\, da_j. \tag{16.13}$$

## 16.2 The Vector Differential for Three Dimensions

Let $\mathbf{r} = (x, y, z)$ be the vector from the origin in three-dimensional space to the point $(x, y, z)$ in rectangular coordinates. Suppose that there is another coordinate system, $(u, v, w)$, such that $x = f(u, v, w)$, $y = g(u, v, w)$ and $z = h(u, v, w)$. Then, with $\mathbf{a} = (u, v, w)$, we write

$$\frac{\partial \mathbf{r}}{\partial u}(\mathbf{a}) = (\frac{\partial x}{\partial u}(\mathbf{a}), \frac{\partial y}{\partial u}(\mathbf{a}), \frac{\partial z}{\partial u}(\mathbf{a})), \tag{16.14}$$

$$\frac{\partial \mathbf{r}}{\partial v}(\mathbf{a}) = (\frac{\partial x}{\partial v}(\mathbf{a}), \frac{\partial y}{\partial v}(\mathbf{a}), \frac{\partial z}{\partial v}(\mathbf{a})), \tag{16.15}$$

and

$$\frac{\partial \mathbf{r}}{\partial w}(\mathbf{a}) = (\frac{\partial x}{\partial w}(\mathbf{a}), \frac{\partial y}{\partial w}(\mathbf{a}), \frac{\partial z}{\partial w}(\mathbf{a})). \tag{16.16}$$

The vector differential $d\mathbf{r}$ is then

$$d\mathbf{r}(\mathbf{a}) = \frac{\partial \mathbf{r}}{\partial u}(\mathbf{a})du + \frac{\partial \mathbf{r}}{\partial v}(\mathbf{a})dv + \frac{\partial \mathbf{r}}{\partial w}(\mathbf{a})dw, \tag{16.17}$$

which we obtain by applying the mean value theorem of the previous section, viewing each of the functions $x(u, v, w)$, $y(u, v, w)$, and $z(u, v, w, )$ as one of the $r_j$. We view $d\mathbf{r}$ as the diagonal of an infinitesimal parallelepiped with one corner at the point $(x, y, z)$. We want to compute the volume of this parallelepiped.

The vectors $\mathbf{A} = \frac{\partial \mathbf{r}}{\partial u}du$, $\mathbf{B} = \frac{\partial \mathbf{r}}{\partial v}dv$ and $\mathbf{C} = \frac{\partial \mathbf{r}}{\partial w}dw$ are then three vectors forming the sides of the parallelepiped. The volume of the parallelepiped is then the absolute value of the vector triple product $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$.

The triple product $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$ is the determinant of the three by three Jacobian matrix

$$J(x, y, z) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{bmatrix}, \tag{16.18}$$

multiplied by $du\,dv\,dw$. Therefore the infinitesimal volume element $dV$ is

$$dV = |\det(J)| du dv dw. \tag{16.19}$$

For example, let us consider spherical coordinates, $(\rho, \phi, \theta)$.

Now we have

$$x = f(\rho, \phi, \theta) = \rho \sin\phi \cdot \cos\theta, \tag{16.20}$$

$$y = g(\rho, \phi, \theta) = \rho \sin\phi \cdot \sin\theta, \tag{16.21}$$

and

$$z = h(\rho, \phi, \theta) = \rho \cos\phi. \tag{16.22}$$

Then the Jacobian matrix is

$$J(x, y, z) = \begin{bmatrix} \sin\phi \cdot \cos\theta & \rho\cos\phi \cdot \cos\theta & -\rho\sin\phi \cdot \sin\theta \\ \sin\phi \cdot \sin\theta & \rho\cos\phi \cdot \sin\theta & \rho\sin\phi \cdot \cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{bmatrix}, \tag{16.23}$$

and

$$\det(J) = \rho^2 \sin\phi. \tag{16.24}$$

Therefore, the infinitesimal volume element in spherical coordinates is

$$\rho^2 \sin\phi \, d\rho d\phi d\theta. \tag{16.25}$$

Similar formulas hold in two dimensions, as the example of polar coordinates shows.

In the polar-coordinates system $(\rho, \theta)$ in two dimensions we have $x = \rho\cos\theta$, and $y = \rho\sin\theta$. Then the Jacobian matrix is

$$J(x, y) = \begin{bmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{bmatrix}, \tag{16.26}$$

and

$$\det(J) = \rho. \tag{16.27}$$

Therefore, the infinitesimal area element in polar coordinates is

$$\rho d\rho d\theta. \tag{16.28}$$

# Chapter 17

# Div, Grad, Curl (Chapter 5,6)

When we begin to study vector calculus, we encounter a number of new concepts, *divergence*, *gradient*, *curl*, and so on, all related to the *del* operator, $\nabla$. Shortly thereafter, we are hit with a blizzard of formulas relating these concepts. It is all rather abstract and students easily lose their way. It occurred to Prof. Schey of MIT to present these ideas to his students side-by-side with the basics of electrostatics, which, after all, was one of the main applications that drove the development of the vector calculus in the first place. Eventually, he wrote a small book [40], which is now a classic. These notes are based, in part, on that book.

## 17.1 The Electric Field

The basic principles of the electrostatics are the following:

- 1. there are positive and negative electrical charges, and like charges repel, unlike charges attract;
- 2. the force is a *central* force, that is, the force that one charge exerts on another is directed along the ray between them and, by Coulomb's Law, its strength falls off as the square of the distance between them;
- 3. *super-position* holds, which means that the force that results from multiple charges is the vector sum of the forces exerted by each one separately.

Apart from the first principle, this is a good description of gravity and magnetism as well. According to Newton, every massive body exerts a

gravitational force of attraction on every other massive body. A space craft heading to the moon feels the attractive force of both the earth and the moon. For most of the journey, the craft is trying to escape the earth, and the effect of the moon pulling the craft toward itself is small. But, at some point in the journey, the attraction of the moon becomes stronger than that of the earth, and the craft is mainly being pulled toward the moon. Even before the space craft was launched, something existed up there in space, waiting for a massive object to arrive and experience attractive force. This something is the *gravitational field* due to the totality of massive bodies doing the attracting. Einstein and others showed that gravity is a bit more complicated than that, but this is a story for another time and a different teller.

Faraday, working in England in the first half of the nineteenth century, was the first to apply this idea of a *field* to electrostatics. He reasoned that a distribution of electrical charges sets up something analogous to a gravitational field, called an *electric field*, such that, once another charge is placed within that field, it has a force exerted on it. The important idea here is that something exists *out there* even when there is no charge present to experience this force, just as with the gravitational field. There are also magnetic fields, and the study of the interaction of electric and magnetic fields is the focus of *electromagnetism*.

## 17.2 The Electric Field Due To A Single Charge

Suppose there is charge $q$ at the origin in three-dimensional space. The electric field resulting from this charge is

$$\mathbf{E}(x,y,z) = \frac{q}{x^2+y^2+z^2}\mathbf{u}(x,y,z), \tag{17.1}$$

where

$$\mathbf{u}(x,y,z) = \left(\frac{x}{\sqrt{x^2+y^2+z^2}}, \frac{y}{\sqrt{x^2+y^2+z^2}}, \frac{z}{\sqrt{x^2+y^2+z^2}}\right)$$

is the unit vector pointing from $(0,0,0)$ to $(x,y,z)$. The electric field can be written in terms of its component functions, that is,

$$\mathbf{E}(x,y,z) = (E_1(x,y,z), E_2(x,y,z), E_3(x,y,z)),$$

where

$$E_1(x,y,z) = \frac{qx}{(x^2+y^2+z^2)^{3/2}},$$

$$E_2(x,y,z) = \frac{qy}{(x^2+y^2+z^2)^{3/2}},$$

and

$$E_3(x,y,z) = \frac{qz}{(x^2+y^2+z^2)^{3/2}}.$$

It is helpful to note that these component functions are the three first partial derivatives of the function

$$\phi(x,y,z) = \frac{-q}{\sqrt{x^2+y^2+z^2}}. \tag{17.2}$$

## 17.3 Gradients and Potentials

Because of the super-position principle, even when the electric field is the result of multiple charges it will still be true that the component functions of the field are the three partial derivatives of some scalar-valued function $\phi(x,y,z)$. This function is called the *potential function* for the field.

For any scalar-valued function $f(x,y,z)$, the *gradient* of $f$ at the point $(x,y,z)$ is the vector of its first partial derivatives at $(x,y,z)$, that is,

$$\nabla f(x,y,z) = (\frac{\partial f}{\partial x}(x,y,z), \frac{\partial f}{\partial y}(x,y,z), \frac{\partial f}{\partial z}(x,y,z));$$

the vector-valued function $\nabla f$ is called the *gradient field* of $f$. Therefore, the electric field $\mathbf{E}$ is the gradient field of its potential function.

## 17.4 Gauss's Law

Let's begin by looking at Gauss's Law, and then we'll try to figure out what it means.

**Gauss's Law:**

$$\int\int_S \mathbf{E}\cdot\mathbf{n}\,dS = 4\pi \int\int\int_V \rho\,dV. \tag{17.3}$$

The integral on the left side is the integral over the surface $S$, while the integral on the right side is the triple integral over the volume $V$ enclosed by the surface $S$. We must remember to think of integrals as summing, so on the left we are summing something over the surface, while on the right we are summing something else over the enclosed volume.

### 17.4.1 The Charge Density Function

The function $\rho = \rho(x,y,z)$ assigns to each point in space a number, the charge density at that point. The vector $\mathbf{n} = \mathbf{n}(x,y,z)$ is the outward unit normal vector to the surface at the point $(x,y,z)$ on the surface, that is, it is a unit vector pointing directly out of the surface at the point $(x,y,z)$.

### 17.4.2 The Flux

The dot product

$$\mathbf{E} \cdot \mathbf{n} = \mathbf{E}(x, y, z) \cdot \mathbf{n}(x, y, z)$$

is the amplitude, that is, plus or minus the magnitude, of the component of the electric field vector $\mathbf{E}(x, y, z)$ that points directly out of the surface. The surface integral on the left side of Equation (17.3) is a measure of the outward *flux* of the electric field through the surface. If there were no charges inside the surface $S$ there would be no outward flux. Gauss's Law tells us that the total outward flux that does exist is due to how much charge there is inside the surface, that is, to the totality of charge density inside the surface.

Our goal is to find a convenient way to determine the electric field everywhere, assuming we know the charge density function everywhere. Gauss's Law is only a partial answer, since it seems to require lots of surface and volume integrals.

## 17.5 A Local Gauss's Law and Divergence

Gauss's Law involves arbitrary surfaces and the volumes they enclose. It would be more helpful if the law could be expressed *locally*, at each point in space separately. To achieve this, we consider a fixed point $(x, y, z)$ in space, and imagine this point to be the center of a sphere. We apply Gauss's Law to this sphere and get the flux through its surface. Now we imagine shrinking the sphere down to its center point. As we shall show later, in the limit, the ratio of the flux to the volume of the sphere, as the radius of the sphere goes to zero, is the *divergence* of the field $\mathbf{E}$, whose value at the point $(x, y, z)$ is the number

$$\operatorname{div} \mathbf{E}(x, y, z) = \frac{\partial E_1}{\partial x}(x, y, z) + \frac{\partial E_2}{\partial y}(x, y, z) + \frac{\partial E_3}{\partial z}(x, y, z). \qquad (17.4)$$

For notational convenience, we also write the divergence function as

$$\operatorname{div} \mathbf{E} = \nabla \cdot \mathbf{E},$$

where the symbol

$$\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$$

is the *del* operator.

When we apply the same limiting process to the integral on the right side of Gauss's Law, we just get $4\pi\rho(x, y, z)$. Therefore, the *local* or *differential* form of Gauss's Law becomes

$$\operatorname{div} \mathbf{E}(x, y, z) = 4\pi\rho(x, y, z). \qquad (17.5)$$

This is also the first of the four *Maxwell's Equations*. When we substitute $\operatorname{div}\mathbf{E}(x,y,z)$ for $4\pi\rho(x,y,z)$ in Equation (17.3) we get

$$\int\int_S \mathbf{E}\cdot\mathbf{n}\,dS = \int\int\int_V \operatorname{div}\mathbf{E}(x,y,z)\,dV, \tag{17.6}$$

which is the *Divergence Theorem*.

Our goal is to determine the electric field from knowledge of the charge density function $\rho$. The partial differential equation in (17.5) is not enough, by itself, since it involves three different unknown functions, $E_1$, $E_2$, and $E_3$, and only one known function $\rho$. The next step in solving the problem involves the potential function for the electric field.

### 17.5.1 The Laplacian

For a scalar-valued function $f(x,y,z)$ the *Laplacian* is

$$\nabla^2 f(x,y,z) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \nabla\cdot(\nabla f).$$

For a vector-valued function

$$\mathbf{F}(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z)),$$

the symbol $\nabla^2\mathbf{F}$ is the vector-valued function whose components are the Laplacians of the individual $F_1$, $F_2$, and $F_3$, that is,

$$\nabla^2\mathbf{F} = (\nabla^2 F_1, \nabla^2 F_2, \nabla^2 F_3).$$

## 17.6 Poisson's Equation and Harmonic Functions

As we discussed earlier, the component functions of the electric field are the three first partial derivatives of a single function, $\phi(x,y,z)$, the *electrostatic potential function*. Our goal then is to find the potential function. When we calculate the divergence of the electric field using $\phi$ we find that

$$\operatorname{div}\mathbf{E}(x,y,z) = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = \nabla\cdot(\nabla\phi) = \nabla^2\phi.$$

Therefore, the differential form of Gauss's Law can be written as

$$\nabla^2\phi(x,y,z) = 4\pi\rho(x,y,z); \tag{17.7}$$

this is called *Poisson's Equation*. In any region of space where there are no charges, that is, where $\rho(x,y,z) = 0$, we have

$$\nabla^2\phi(x,y,z) = 0. \tag{17.8}$$

Functions that satisfy Equation (17.8) are called *harmonic functions*. The reader may know that both the real and imaginary parts of a complex-valued analytic function are harmonic functions of two variables. This connection between electrostatics and complex analysis motivated the (ultimately fruitless) search for a three-dimensional extension of complex analysis.

## 17.7 The Curl

The divergence of a vector field is a local measure of the flux, which we may think of as outward flow of something. The curl is a measure of the rotation of the something.

For any vector field $\mathbf{F}(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$, the *curl* of $\mathbf{F}$ is the vector field

$$\operatorname{curl}\mathbf{F}(x,y,z) = \nabla \times \mathbf{F} = (\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}, \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}). \quad (17.9)$$

A useful identity involving the curl is the following:

$$\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2 \mathbf{F}. \quad (17.10)$$

### 17.7.1 An Example

The curve $\mathbf{r}(t)$ in three-dimensional space given by

$$\mathbf{r}(t) = (x(t), y(t), z(t)) = r(\cos\theta(t), \sin\theta(t), 0)$$

can be viewed as describing the motion of a point moving in time, revolving counter-clockwise around the $z$-axis. The velocity vector at each point is

$$\mathbf{v}(t) = \mathbf{r}'(t) = (x'(t), y'(t), z'(t)) = \frac{d\theta}{dt}(-r\sin\theta(t), r\cos\theta(t), 0).$$

Suppressing the dependence on $t$, we can write the velocity vector field as

$$\mathbf{v}(x,y,z) = \frac{d\theta}{dt}(-y, x, 0).$$

Then

$$\operatorname{curl}\mathbf{v}(x,y,z) = (0,0,2\omega),$$

where $\omega = \frac{d\theta}{dt}$ is the angular velocity. The divergence of the velocity field is

$$\operatorname{div}\mathbf{v}(x,y,z) = 0.$$

The motion here is rotational; there is no outward flow of anything. Here the curl describes how fast the rotation is, and indicates the axis of rotation; the fact that there is no outward flow is indicated by the divergence being zero.

### 17.7.2 Solenoidal Fields

When the divergence of a vector field is zero, the field is said to be *solenoidal*; the velocity field in the previous example is solenoidal. The second of Maxwell's four equations is that the magnetic field is solenoidal.

### 17.7.3 The Curl of the Electrostatic Field

We can safely assume that the mixed second partial derivatives of the potential function $\phi$ satisfy

$$\frac{\partial^2 \phi}{\partial x \partial y} = \frac{\partial^2 \phi}{\partial y \partial x},$$

$$\frac{\partial^2 \phi}{\partial x \partial z} = \frac{\partial^2 \phi}{\partial z \partial x},$$

and

$$\frac{\partial^2 \phi}{\partial z \partial y} = \frac{\partial^2 \phi}{\partial y \partial z}.$$

It follows, therefore, that, because the electrostatic field has a potential, its curl is zero. The third of Maxwell's Equations (for electrostatics) is

$$\operatorname{curl} \mathbf{E}(x, y, z) = 0. \tag{17.11}$$

## 17.8 The Magnetic Field

We denote by $\mathbf{B}(x, y, z)$ a magnetic field. In the static case, in which neither the magnetic field nor the electric field is changing with respect to time, there is no connection between them. The equations that describe this situation are

**Maxwell's Equations for the Static Case:**

- 1. $\operatorname{div} \mathbf{E} = 4\pi\rho$;
- 2. $\operatorname{curl} \mathbf{E} = 0$;
- 3. $\operatorname{div} \mathbf{B} = 0$;
- 4. $\operatorname{curl} \mathbf{B} = 0$.

It is what happens in the dynamic case, when the electric and magnetic fields change with time, that is interesting.

Ampere discovered that a wire carrying a current acts like a magnet. When the electric field changes with time, there is a current density vector

field **J** proportional to the rate of change of the electric field, and Item 4 above is replaced by Ampere's Law:

$$\text{curl}\,\mathbf{B} = a\frac{\partial \mathbf{E}}{\partial t},$$

where $a$ is some constant. Therefore, the curl of the magnetic field is proportional to the rate of change of the electric field with respect to time.

Faraday (and also Henry) discovered that moving a magnet inside a wire coil creates a current in the wire. When the magnetic field is changing with respect to time, the electric field has a non-zero curl proportional to the rate at which the magnetic field is changing. Then Item 2 above is replaced by

$$\text{curl}\,\mathbf{E} = b\frac{\partial \mathbf{B}}{\partial t},$$

where $b$ is some constant. Therefore, the curl of the electric field is proportional to the rate of change of the magnetic field. It is this mutual dependence that causes electromagnetic waves: as the electric field changes, it creates a changing magnetic field, which, in turn, creates a changing electric field, and so on.

## 17.9 Electro-magnetic Waves

We consider now the behavior of electric and magnetic fields that are changing with time, in a region of space where there are no charges or currents. Maxwell's Equations are then

- 1. $\text{div}\,\mathbf{E} = 0$;
- 2. $\text{curl}\,\mathbf{E} = -b\frac{\partial \mathbf{B}}{\partial t}$;
- 3. $\text{div}\,\mathbf{B} = 0$;
- 4. $\text{curl}\,\mathbf{B} = a\frac{\partial \mathbf{E}}{\partial t}$.

We then have

$$\nabla \times (\nabla \times \mathbf{E}) = -b(\nabla \times \frac{\partial B}{\partial t}) = -b\frac{\partial}{\partial t}(\nabla \times B) = -ab\frac{\partial}{\partial t}(\frac{\partial \mathbf{E}}{\partial t}) = -ab\frac{\partial^2 \mathbf{E}}{\partial t^2}.$$

Using Equation (17.10), we can also write

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} = \nabla\text{div}\,\mathbf{E} - \nabla^2\mathbf{E} = -\nabla^2\mathbf{E}.$$

Therefore, we have

$$\nabla^2\mathbf{E} = ab\frac{\partial^2 \mathbf{E}}{\partial t^2},$$

which means that, for each $i = 1, 2, 3$, the component function $E_i$ satisfies the three-dimensional wave equation

$$\frac{\partial^2 E_i}{\partial t^2} = c^2 \nabla^2 E_i.$$

The same is true for the component functions of the magnetic field. Here the constant $c$ is the speed of propagation of the wave, which turns out to be the speed of light. It was this discovery that suggested to Maxwell that light is an electromagnetic phenomenon.

# Chapter 18

# Kepler's Laws of Planetary Motion (Chapter 5,6)

## 18.1 Introduction

Kepler worked from 1601 to 1612 in Prague as the Imperial Mathematician. Taking over from Tycho Brahe, and using the tremendous amount of data gathered by Brahe from naked-eye astronomical observation, he formulated three laws governing planetary motion. Fortunately, among his tasks was the study of the planet Mars, whose orbit is quite unlike a circle, at least relatively speaking. This forced Kepler to consider other possibilities and ultimately led to his discovery of elliptic orbits. These laws, which were the first "natural laws" in the modern sense, served to divorce astronomy from theology and philosophy and marry it to physics. At last, the planets were viewed as material bodies, not unlike earth, floating freely in space and moved by physical forces acting on them. Although the theology and philosophy of the time dictated uniform planetary motion and circular orbits, nature was now free to ignore these demands; motion of the planets could be non-uniform and the orbits other than circular.

Although the second law preceded the first, Kepler's Laws are usually enumerated as follows:

- 1. the planets travel around the sun not in circles but in elliptical orbits, with the sun at one focal point;
- 2. a planet's speed is not uniform, but is such that the line segment from the sun to the planet sweeps out equal areas in equal time intervals; and, finally,

- 3. for all the planets, the time required for the planet to complete one orbit around the sun, divided by the 3/2 power of its average distance from the sun, is the same constant.

These laws, particularly the third one, provided strong evidence for Newton's law of universal gravitation. How Kepler discovered these laws without the aid of analytic geometry and differential calculus, with no notion of momentum, and only a vague conception of gravity, is a fascinating story, perhaps best told by Koestler in [31].

Around 1684, Newton was asked by Edmund Halley, of Halley's comet fame, what the path would be for a planet moving around the sun, if the force of gravity fell off as the square of the distance from the sun. Newton responded that it would be an ellipse. Kepler had already declared that planets moved along elliptical orbits with the sun at one focal point, but his findings were based on observation and imagination, not deduction from physical principles. Halley asked Newton to provide a proof. To supply such a proof, Newton needed to write a whole book, the *Principia*, published in 1687, in which he had to deal with such mathematically difficult questions as what the gravitational force is on a point when the attracting body is not just another point, but a sphere, like the sun.

With the help of vector calculus, a later invention, Kepler's laws can be derived as consequences of Newton's inverse square law for gravitational attraction.

## 18.2 Preliminaries

We consider a body with constant mass $m$ moving through three-dimensional space along a curve

$$\mathbf{r}(t) = (x(t), y(t), z(t)),$$

where $t$ is time and the sun is the origin. The velocity vector at time $t$ is then

$$\mathbf{v}(t) = \mathbf{r}'(t) = (x'(t), y'(t), z'(t)),$$

and the acceleration vector at time $t$ is

$$\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{r}''(t) = (x''(t), y''(t), z''(t)).$$

The *linear momentum* vector is

$$\mathbf{p}(t) = m\mathbf{v}(t).$$

One of the most basic laws of motion is that the vector $\mathbf{p}'(t) = m\mathbf{v}'(t) = m\mathbf{a}(t)$ is equal to the external force exerted on the body. When a body, or more precisely, the center of mass of the body, does not change location, all it can do is rotate. In order for a body to rotate about an axis a *torque*

is required. Just as work equals force times distance moved, work done in rotating a body equals torque times angle through which it is rotated. Just as force is the time derivative of $\mathbf{p}(t)$, the linear momentum vector, we find that torque is the time derivative of something else, called the *angular momentum vector*.

## 18.3 Torque and Angular Momentum

Consider a body rotating around the origin in two-dimensional space, whose position at time $t$ is

$$\mathbf{r}(t) = (r\cos\theta(t), r\sin\theta(t)).$$

Then at time $t + \Delta t$ it is at

$$\mathbf{r}(t + \Delta t) = (r\cos(\theta(t) + \Delta\theta), r\sin(\theta(t) + \Delta\theta)).$$

Therefore, using trig identities, we find that the change in the $x$-coordinate is approximately

$$\Delta x = -r\Delta\theta\sin\theta(t) = -y(t)\Delta\theta,$$

and the change in the $y$-coordinate is approximately

$$\Delta y = r\Delta\theta\cos\theta(t) = x(t)\Delta\theta.$$

The infinitesimal work done by a force $\mathbf{F} = (F_x, F_y)$ in rotating the body through the angle $\Delta\theta$ is then approximately

$$\Delta W = F_x\Delta x + F_y\Delta y = (F_y x(t) - F_x y(t))\Delta\theta.$$

Since work is torque times angle, we define the torque to be

$$\tau = F_y x(t) - F_x y(t).$$

The entire motion is taking place in two dimensional space. Nevertheless, it is convenient to make use of the concept of cross product of three-dimensional vectors to represent the torque. When we rewrite

$$\mathbf{r}(t) = (x(t), y(t), 0),$$

and

$$\mathbf{F} = (F_x, F_y, 0),$$

we find that

$$\mathbf{r}(t) \times \mathbf{F} = (0, 0, F_y x(t) - F_x y(t)) = (0, 0, \tau) = \tau.$$

Now we use the fact that the force is the time derivative of the vector $\mathbf{p}(t)$ to write

$$\tau = (0, 0, \tau) = \mathbf{r}(t) \times \mathbf{p}'(t).$$

**Exercise 18.1** *Show that*

$$\mathbf{r}(t) \times \mathbf{p}'(t) = \frac{d}{dt}(\mathbf{r}(t) \times \mathbf{p}(t)). \tag{18.1}$$

By analogy with force as the time derivative of linear momentum, we define torque as the time derivative of angular momentum, which, from the calculations just performed, leads to the definition of the *angular momentum vector* as

$$\mathbf{L}(t) = \mathbf{r}(t) \times \mathbf{p}(t).$$

We need to say a word about the word "vector". In our example of rotation in two dimensions we introduced the third dimension as merely a notational convenience. It is convenient to be able to represent the torque as $\mathbf{L}'(t) = (0, 0, \tau)$, but when we casually call $L(t)$ the angular momentum vector, physicists would tell us that we haven't yet shown that angular momentum is a "vector" in the physicists' sense. Our example was too simple, they would point out. We had rotation about a single fixed axis that was conveniently chosen to be one of the coordinate axes in three-dimensional space. But what happens when the coordinate system changes?

Clearly, they would say, physical objects rotate and have angular momentum. The earth rotates around an axis, but this axis is not always the same axis; the axis wobbles. A well thrown football rotates around its longest axis, but this axis changes as the ball flies through the air. Can we still say that the angular momentum can be represented as

$$\mathbf{L}(t) = \mathbf{r}(t) \times \mathbf{p}(t)?$$

In other words, we need to know that the torque is still the time derivative of $\mathbf{L}(t)$, even as the coordinate system changes. In order for something to be a "vector"in the physicists' sense, it needs to behave properly as we switch coordinate systems, that is, it needs to *transform as a vector* [15]. In fact, all is well. This definition of $\mathbf{L}(t)$ holds for bodies moving along more general curves in three-dimensional space, and we can go on calling $\mathbf{L}(t)$ the angular momentum vector. Now we begin to exploit the special nature of the gravitational force.

## 18.4 Gravity is a Central Force

We are not interested here in arbitrary forces, but in the gravitational force that the sun exerts on the body, which has special properties that we shall exploit. In particular, this gravitational force is a *central force*.

**Definition 18.1** *We say that the force is a* central force *if*

$$\mathbf{F}(t) = h(t)\mathbf{r}(t),$$

*for each $t$, where $h(t)$ denotes a scalar function of $t$; that is, the force is central if it is proportional to $\mathbf{r}(t)$ at each $t$.*

**Proposition 18.1** *If $\mathbf{F}(t)$ is a central force, then $\mathbf{L}'(t) = \mathbf{0}$, for all $t$, so that $\mathbf{L} = \mathbf{L}(t)$ is a constant vector and $L = ||\mathbf{L}(t)|| = ||\mathbf{L}||$ is a constant scalar, for all $t$.*

**Proof:** From Equation (18.1) we have

$$\mathbf{L}'(t) = \mathbf{r}(t) \times \mathbf{p}'(t) = \mathbf{r}(t) \times \mathbf{F}(t) = h(t)\mathbf{r}(t) \times \mathbf{r}(t) = \mathbf{0}.$$

∎

We see then that the angular momentum vector $\mathbf{L}(t)$ is *conserved* when the force is central.

**Proposition 18.2** *If $\mathbf{L}'(t) = \mathbf{0}$, then the curve $\mathbf{r}(t)$ lies in a plane.*

**Proof:** We have

$$\mathbf{r}(t) \cdot \mathbf{L} = \mathbf{r}(t) \cdot \mathbf{L}(t) = \mathbf{r}(t) \cdot \Big(\mathbf{r}(t) \times \mathbf{p}(t)\Big),$$

which is the volume of the parallelepiped formed by the three vectors $\mathbf{r}(t)$, $\mathbf{r}(t)$ and $\mathbf{p}(t)$, which is obviously zero. Therefore, for every $t$, the vector $\mathbf{r}(t)$ is orthogonal to the constant vector $\mathbf{L}$. So, the curve lies in a plane with normal vector $\mathbf{L}$. ∎

## 18.5 The Second Law

We know now that, since the force is central, the curve described by $\mathbf{r}(t)$ lies in a plane. This allows us to use polar coordinate notation [42]. We write

$$\mathbf{r}(t) = \rho(t)(\cos\theta(t), \sin\theta(t)) = \rho(t)\mathbf{u}_r(t),$$

where $\rho(t)$ is the length of the vector $\mathbf{r}(t)$ and

$$\mathbf{u}_r(t) = \frac{\mathbf{r}(t)}{||\mathbf{r}(t)||} = (\cos\theta(t), \sin\theta(t))$$

is the unit vector in the direction of $\mathbf{r}(t)$. We also define

$$\mathbf{u}_\theta(t) = (-\sin\theta(t), \cos\theta(t)),$$

so that

$$\mathbf{u}_\theta(t) = \frac{d}{d\theta}\mathbf{u}_r(t),$$

and

$$\mathbf{u}_r(t) = -\frac{d}{d\theta}\mathbf{u}_\theta(t).$$

**Exercise 18.2** *Show that*

$$\mathbf{p}(t) = m\rho'(t)\mathbf{u}_r(t) + m\rho(t)\frac{d\theta}{dt}\mathbf{u}_\theta(t). \tag{18.2}$$

**Exercise 18.3** *View the vectors $\mathbf{r}(t)$, $\mathbf{p}(t)$, $\mathbf{u}_r(t)$ and $\mathbf{u}_\theta(t)$ as vectors in three-dimensional space, all with third component equal to zero. Show that*

$$\mathbf{u}_r(t) \times \mathbf{u}_\theta(t) = \mathbf{k} = (0, 0, 1),$$

*for all $t$. Use this and Equation (18.2) to show that*

$$\mathbf{L} = \mathbf{L}(t) = \Big(m\rho(t)^2\frac{d\theta}{dt}\Big)\mathbf{k},$$

*so that $L = m\rho(t)^2\frac{d\theta}{dt}$, the moment of inertia times the angular velocity, is constant.*

Let $t_0$ be some arbitrary time, and for any time $t \geq t_0$ let $A(t)$ be the area swept out by the planet in the time interval $[t_0, t]$. Then $A(t_2) - A(t_1)$ is the area swept out in the time interval $[t_1, t_2]$.

In the very short time interval $[t, t + \Delta t]$ the vector $\mathbf{r}(t)$ sweeps out a very small angle $\Delta\theta$, and the very small amount of area formed is then approximately

$$\Delta A = \frac{1}{2}\rho(t)^2\Delta\theta.$$

Dividing by $\Delta t$ and taking limits, as $\Delta t \to 0$, we get

$$\frac{dA}{dt} = \frac{1}{2}\rho(t)^2\frac{d\theta}{dt} = \frac{L}{2m}.$$

Therefore, the area swept out between times $t_1$ and $t_2$ is

$$A(t_2) - A(t_1) = \int_{t_1}^{t_2}\frac{dA}{dt}dt = \int_{t_1}^{t_2}\frac{L}{2m}dt = \frac{L(t_2 - t_1)}{2m}.$$

This is Kepler's Second Law.

## 18.6 The First Law

We saw previously that the angular momentum vector is conserved when the force is central. When Newton's inverse-square law holds, there is another conservation law; the *Runge-Lenz vector* is also conserved. We shall use this fact to derive the First Law.

Let $M$ denote the mass of the sun, and $G$ Newton's gravitational constant.

**Definition 18.2** *The force obeys Newton's inverse square law if*

$$\mathbf{F}(t) = h(t)\mathbf{r}(t) = -\frac{mMG}{\rho(t)^3}\mathbf{r}(t).$$

Then we can write

$$\mathbf{F}(t) = -\frac{mMG}{\rho(t)^2}\frac{\mathbf{r}(t)}{||\mathbf{r}(t)||} = -\frac{mMG}{\rho(t)^2}\mathbf{u}_r(t).$$

**Definition 18.3** *The* Runge-Lenz vector *is*

$$\mathbf{K}(t) = \mathbf{p}(t) \times \mathbf{L}(t) - k\mathbf{u}_r(t),$$

*where* $k = m^2MG$.

**Exercise 18.4** *Show that the velocity vectors* $\mathbf{r}'(t)$ *lie in the same plane as the curve* $\mathbf{r}(t)$.

**Exercise 18.5** *Use the rule*

$$\mathbf{A} \times (\mathbf{A} \times \mathbf{B}) = (\mathbf{A} \cdot \mathbf{B})\mathbf{A} - (\mathbf{A} \cdot \mathbf{A})\mathbf{B}$$

*to show that* $\mathbf{K}'(t) = \mathbf{0}$, *so that* $\mathbf{K} = \mathbf{K}(t)$ *is a constant vector and* $K = ||\mathbf{K}||$ *is a constant scalar.*

So the Runge-Lenz vector is conserved when the force obeys Newton's inverse square law.

**Exercise 18.6** *Use the rule in the previous exercise to show that the constant vector* $\mathbf{K}$ *also lies in the plane of the curve* $\mathbf{r}(t)$.

**Exercise 18.7** *Show that*

$$\mathbf{K} \cdot \mathbf{r}(t) = L^2 - k\rho(t).$$

It follows from this exercise that

$$L^2 - k\rho(t) = \mathbf{K} \cdot \mathbf{r}(t) = K\rho(t)\cos\alpha(t),$$

where $\alpha(t)$ is the angle between the vectors $\mathbf{K}$ and $\mathbf{r}(t)$. From this we get

$$\rho(t) = L^2/(k + K\cos\alpha(t)).$$

For $k > K$, this is the equation of an ellipse having eccentricity $e = K/k$. This is Kepler's First Law.

Kepler initially thought that the orbits were "egg-shaped" , but later came to realize that they were ellipses. Although Kepler did not have the analytical geometry tools to help him, he was familiar with the mathematical development of ellipses in the *Conics*, the ancient book by Apollonius, written in Greek in Alexandria about 200 BC. Conics, or conic sections, are the terms used to describe the two-dimensional curves, such as ellipses, parabolas and hyperbolas, formed when a plane intersects an infinite double cone (think "hour-glass").

Apollonius was interested in astronomy and Ptolemy was certainly aware of the work of Apollonius, but it took Kepler to overcome the bias toward circular motion and introduce conic sections into astronomy. As related by Bochner [3], there is a bit of mystery concerning Kepler's use of the *Conics*. He shows that he is familiar with a part of the *Conics* that existed only in Arabic until translated into Latin in 1661, well after his time. How he gained that familiarity is the mystery.

## 18.7 The Third Law

As the planet moves around its orbit, the closest distance to the sun is

$$\rho_{\min} = L^2/(k + K),$$

and the farthest distance is

$$\rho_{\max} = L^2/(k - K).$$

The average of these two is

$$a = \frac{1}{2}\Big(\rho_{\min} + \rho_{\max}\Big) = 2kL^2/(k^2 - K^2);$$

this is the semi-major axis of the ellipse. The semi-minor axis has length $b$, where

$$b^2 = a^2(1 - e^2).$$

Therefore,

$$b = \frac{L\sqrt{a}}{\sqrt{k}}.$$

The area of this ellipse is $\pi ab$. But we know from the first law that the area of the ellipse is $\frac{L}{2m}$ times the time $T$ required to complete a full orbit. Equating the two expressions for the area, we get

$$T^2 = \frac{4\pi^2}{MG}a^3.$$

This is the third law.

The first two laws deal with the behavior of one planet; the third law is different. The third law describes behavior that is common to all the planets in the solar system, thereby suggesting a universality to the force of gravity.

## 18.8 Dark Matter and Dark Energy

Ordinary matter makes up only a small fraction of the "stuff" in the universe. About 25 percent of the stuff is *dark matter* and over two thirds is *dark energy*. Because neither of these interacts with electromagnetic radiation, evidence for their existence is indirect.

Suppose, for the moment, that a planet moves in a circular orbit of radius $a$, centered at the sun. The orbital time is $T$, so, by Kepler's third law, the speed with which the planet orbits the sun is $\sqrt{\frac{MG}{a}}$, so the farther away the planet the slower it moves. Spiral galaxies are like large planetary systems, with some stars nearer to the center of the galaxy than others. We would expect those stars farther from the center of mass of the galaxy to be moving more slowly, but this is not the case. One explanation for this is that there is more mass present, *dark mass* we cannot detect, spread throughout the galaxy and not concentrated just near the center.

According to Einstein, massive objects can bend light. This *gravitational lensing*, distorting the light from distant stars, has been observed by astronomers, but cannot be simply the result of the observable mass present; there must be more mass out there. Again, this provides indirect evidence for dark mass.

The universe is expanding. Until fairly recently, it was believed that, although it was expanding, the rate of expansion was decreasing; the mass in the universe was exerting gravitational pull that was slowing down the rate of expansion. The question was whether or not the expansion would eventually stop and contraction begin. When the rate of expansion was measured, it was discovered that the rate was increasing, not decreasing. The only possible explanation for this seemed to be that *dark energy* was operating and with sufficient strength to overcome not just the pull of ordinary matter, but of the dark matter as well. Understanding dark matter and dark energy is one of the big challenges for physicists of the twenty-first century.

## 18.9 From Kepler to Newton

Our goal, up to now, has been to show how Kepler's three laws can be derived from Newton's inverse-square law, which, of course, is not how Kepler obtained the laws. Kepler arrived at his laws empirically, by studying the astronomical data. Newton was aware of Kepler's laws and they influenced his work on universal gravitation. When asked what would explain Kepler's elliptical orbits, Newton replied that he had calculated that an inverse-square law would do it. Newton found that the force required to cause the moon to deviate from a tangent line was approximately that given by an inverse-square fall-off in gravity.

It is interesting to ask if the inverse-square law can be derived from Kepler's three laws; the answer is yes, as we shall see in this section. What follows is taken from [21].

We found previously that

$$\frac{dA}{dt} = \frac{1}{2}\rho(t)^2\frac{d\theta}{dt} = \frac{L}{2m} = c. \tag{18.3}$$

Differentiating with respect to $t$, we get

$$\rho(t)\rho'(t)\frac{d\theta}{dt} + \frac{1}{2}\rho(t)^2\frac{d^2\theta}{dt^2} = 0, \tag{18.4}$$

so that

$$2\rho'(t)\frac{d\theta}{dt} + \rho(t)\frac{d^2\theta}{dt^2} = 0. \tag{18.5}$$

From this, we shall prove that the force is central, directed towards the sun.

As we did earlier, we write the position vector $\mathbf{r}(t)$ as

$$\mathbf{r}(t) = \rho(t)\mathbf{u}_r(t),$$

so, suppressing the dependence on the time $t$, and using the identities

$$\frac{d\mathbf{u}_r}{dt} = \mathbf{u}_\theta\frac{d\theta}{dt},$$

and

$$\frac{d\mathbf{u}_\theta}{dt} = -\mathbf{u}_r\frac{d\rho}{dt},$$

we write the velocity vector as

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\mathbf{u}_r}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\mathbf{u}_r}{d\theta}\frac{d\theta}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\theta}{dt}\mathbf{u}_\theta,$$

and the acceleration vector as

$$\mathbf{a} = \frac{d^2\rho}{dt^2}\mathbf{u}_r + \frac{d\rho}{dt}\frac{d\mathbf{u}_r}{dt} + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \rho\frac{d^2\theta}{dt^2}\mathbf{u}_\theta + \rho\frac{d\theta}{dt}\frac{d\mathbf{u}_\theta}{dt}$$

$$= \frac{d^2\rho}{dt^2}\mathbf{u}_r + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \rho\frac{d^2\theta}{dt^2}\mathbf{u}_\theta - \rho\frac{d\theta}{dt}\frac{d\theta}{dt}\mathbf{u}_r.$$

Therefore, we have

$$\mathbf{a} = \Big(\frac{d^2\rho}{dt^2} - \rho(\frac{d\theta}{dt})^2\Big)\mathbf{u}_r + \Big(2\frac{d\rho}{dt}\frac{d\theta}{dt} + \rho\frac{d^2\theta}{dt^2}\Big)\mathbf{u}_\theta.$$

Using Equation (18.4), this reduces to

$$\mathbf{a} = \Big(\frac{d^2\rho}{dt^2} - \rho(\frac{d\theta}{dt})^2\Big)\mathbf{u}_r, \tag{18.6}$$

which tells us that the acceleration, and therefore the force, is directed along the line joining the planet to the sun; it is a central force.

**Exercise 18.8** *Prove the following two identities:*

$$\frac{d\rho}{dt} = \frac{d\rho}{d\theta}\frac{d\theta}{dt} = \frac{2c}{\rho^2}\frac{d\theta}{dt} \tag{18.7}$$

*and*

$$\frac{d^2\rho}{dt^2} = \frac{4c^2}{\rho^4}\frac{d^2\rho}{d\theta^2} - \frac{8c^2}{\rho^5}\Big(\frac{d\rho}{d\theta}\Big)^2. \tag{18.8}$$

Therefore, we can write the acceleration vector as

$$\mathbf{a} = \left(\frac{4c^2}{\rho^4}\frac{d^2\rho}{d\theta^2} - \frac{8c^2}{\rho^5}\Big(\frac{d\rho}{d\theta}\Big)^2 - \frac{4c^2}{\rho^3}\right)\mathbf{u}_r.$$

To simplify, we substitute $u = \rho^{-1}$.

**Exercise 18.9** *Prove that the acceleration vector can be written as*

$$\mathbf{a} = \left(4c^2u^2\left(-\frac{1}{u^2}\frac{d^2u}{d\theta^2} + \frac{2}{u^3}\Big(\frac{du}{d\theta}\Big)^2\right) - 8c^2u^5\Big(-\frac{1}{u^2}\frac{du}{d\theta}\Big)^2 - 4c^2u^3\right)\mathbf{u}_r,$$

*so that*

$$\mathbf{a} = -4c^2u^2\Big(\frac{d^2u}{d\theta^2} + u\Big)\mathbf{u}_r. \tag{18.9}$$

Kepler's First Law tells us that

$$\rho(t) = \frac{L^2}{k + K\cos\alpha(t)} = \frac{a(1-e^2)}{1 + e\cos\alpha(t)},$$

where $e = K/k$ and $a$ is the semi-major axis. Therefore,

$$u = \frac{1 + e\cos\alpha(t)}{a(1-e^2)}.$$

Using Equation (18.9), we can write the acceleration as

$$\mathbf{a} = -\frac{4c^2}{a(1-e^2)}u^2\mathbf{u}_r = -\frac{4c^2}{a(1-e^2)}r^{-2}\mathbf{u}_r,$$

which tells us that the force obeys an inverse-square law. We still must show that this same law applies to each of the planets, that is, that the constant $\frac{c^2}{a(1-e^2)}$ does not depend on the particular planet.

**Exercise 18.10** *Show that*

$$\frac{c^2}{a(1-e^2)} = \frac{\pi^2 a^3}{T^2},$$

*which is independent of the particular planet, according to Kepler's Third Law.*

## 18.10 Newton's Own Proof of the Second Law

Although Newton invented calculus, he relied on geometry for many of his mathematical arguments. A good example is his proof of Kepler's Second Law.

He begins by imagining the planet at the point 0 in Figure 18.1. If there were no force coming from the sun, then, by the principle of inertia, the planet would continue in a straight line, with constant speed. The distance $\Delta$ from the point 0 to the point 1 is the same as the distance from 1 to 2 and the same as the distance from 2 to 3. The areas of the three triangles formed by the sun and the points 0 and 1, the sun and the points 1 and 2, and the sun and the points 2 and 3 are all equal, since they all equal half of the base $\Delta$ times the height $H$. Therefore, in the absence of a force from the sun, the planet sweeps out equal areas in equal times. Now what happens when there is a force from the sun?

Newton now assumes that $\Delta$ is very small, and that during the short time it would have taken for the planet to move from 1 to 3 there is a force on the planet, directed toward the sun. Because of the small size of $\Delta$, he safely assumes that the direction of this force is unchanged and is directed

along the line from 2, the midpoint of 1 and 3, to the sun. The effect of such a force is to pull the planet away from 3, along the line from 3 to 4. The areas of the two triangles formed by the sun and the points 2 and 3 and the sun and the points 2 and 4 are both equal to half of the distance from the sun to 2, times the distance from 2 to $B$. So we still have equal areas in equal times.

We can corroborate Newton's approximations using vector calculus. Consider the planet at 2 at time $t = 0$. Suppose that the acceleration is $\mathbf{a}(t) = (b, c)$, where $(b, c)$ is a vector parallel to the line segment from the sun to 2. Then the velocity vector is $\mathbf{v}(t) = t(b, c) + (0, \Delta)$, where, for simplicity, we assume that, in the absence of the force from the sun, the planet travels at a speed of $\Delta$ units per second. The position vector is then

$$\mathbf{r}(t) = \frac{1}{2}t^2(b, c) + t(0, \Delta) + \mathbf{r}(0).$$

At time $t = 1$, instead of the planet being at 3, it is now at

$$\mathbf{r}(1) = \frac{1}{2}(b, c) + (0, \Delta) + \mathbf{r}(0).$$

Since the point 3 corresponds to the position $(0, \Delta) + \mathbf{r}(0)$, we see that the point 4 lies along the line from 3 parallel to the vector $(b, c)$.

## 18.11 Armchair Physics

Mathematicians tend to ignore things like units, when they do calculus problems. Physicists know that you can often learn a lot just by paying attention to the units involved, or by asking questions like what happens to velocity when length is converted from feet to inches and time from minutes to seconds. This is sometimes called "armchair physics" . To illustrate, we apply this approach to Kepler's Third Law.

### 18.11.1 Rescaling

Suppose that the spatial variables $(x, y, z)$ are replaced by $(\alpha x, \alpha y, \alpha z)$ and time changed from $t$ to $\beta t$. Then velocity, since it is distance divided by time, is changed from $v$ to $\alpha\beta^{-1}v$. Velocity squared, and therefore kinetic and potential energies, are changed by a factor of $\alpha^2\beta^{-2}$.

### 18.11.2 Gravitational Potential

The gravitational potential function $\phi(x, y, z)$ associated with the gravitational field due to the sun is given by

$$\phi(x, y, z) = \frac{-C}{\sqrt{x^2 + y^2 + z^2}}, \tag{18.10}$$

where $C > 0$ is some constant and we assume that the sun is at the origin. The gradient of $\phi(x, y, z)$ is

$$\nabla\phi(x, y, z) = \Big(\frac{-C}{x^2 + y^2 + z^2}\Big)\Big(\frac{x}{\sqrt{x^2 + y^2 + z^2}}, \frac{y}{\sqrt{x^2 + y^2 + z^2}}, \frac{z}{\sqrt{x^2 + y^2 + z^2}}\Big).$$

The gravitational force on a massive object at point $(x, y, z)$ is therefore a vector of magnitude $\frac{C}{x^2+y^2+z^2}$, directed from $(x, y, z)$ toward $(0, 0, 0)$, which says that the force is central and falls off as the reciprocal of the distance squared.

The potential function $\phi(x, y, z)$ is $(-1)$-homogeneous, meaning that when we replace $x$ with $\alpha x$, $y$ with $\alpha y$, and $z$ with $\alpha z$, the new potential is the old one times $\alpha^{-1}$.

We also know, though, that when we rescale the space variables by $\alpha$ and time by $\beta$ the potential energy is multiplied by a factor of $\alpha^2\beta^{-2}$. It follows that

$$\alpha^{-1} = \alpha^2\beta^{-2},$$

so that

$$\beta^2 = \alpha^3. \tag{18.11}$$

Suppose that we have two planets, $P_1$ and $P_2$, orbiting the sun in circular orbits, with the length of the the orbit of $P_2$ equal to $\alpha$ times that of $P_1$. We can view the orbital data from $P_2$ as that from $P_1$, after a rescaling of the spatial variables by $\alpha$. According to Equation (18.11), the orbital time of $P_2$ is then that of $P_1$ multiplied by $\beta = \alpha^{3/2}$. This is Kepler's Third Law.

Kepler took several decades to arrive at his third law, which he obtained not from basic physical principles, but from analysis of observational data. Could he have saved himself much time and effort if he had stayed in his armchair and considered rescaling, as we have just done? No. The importance of Kepler's Third Law lies in its universality, the fact that it applies not just to one planet but to all. We have implicitly assumed universality by postulating a potential function that governs the gravitational field from the sun.

### 18.11.3 Gravity on Earth

We turn now to the gravitational pull of the earth on an object near its surface. We have just seen that the potential function is proportional to the reciprocal of the distance from the center of the earth to the object. Let the radius of the earth be $R$ and let the object be at a height $h$ above the surface of the earth. Then the potential is

$$\phi(R + h) = \frac{-B}{R + h},$$

for some constant $B$. The potential at the surface of the earth is

$$\phi(R) = \frac{-B}{R}.$$

The potential difference between the object at height $h$ and the surface of the earth is then

$$PD(h) = \frac{B}{R} - \frac{B}{R+h} = B\Big(\frac{1}{R} - \frac{1}{R+h}\Big) = B\Big(\frac{R+h-R}{R(R+h)}\Big).$$

If $h$ is very small relative to $R$, then we can say that

$$PD(h) = \frac{B}{R^2}h,$$

so is linear in $h$. The potential difference is therefore 1-homogeneous; if we rescale the spatial variables by $\alpha$ the potential difference is also rescaled by $\alpha$. But, as we saw previously, the potential difference is also rescaled by $\alpha^2\beta^{-2}$. Therefore,

$$\alpha = \alpha^2\beta^{-2},$$

or

$$\beta = \alpha^{1/2}.$$

This makes sense. Consider a ball dropped from a tall building. In order to double the time of fall (multiply $t$ by $\beta = 2$) we must quadruple the height from which it is dropped (multiply $h$ by $\alpha = \beta^2 = 4$).

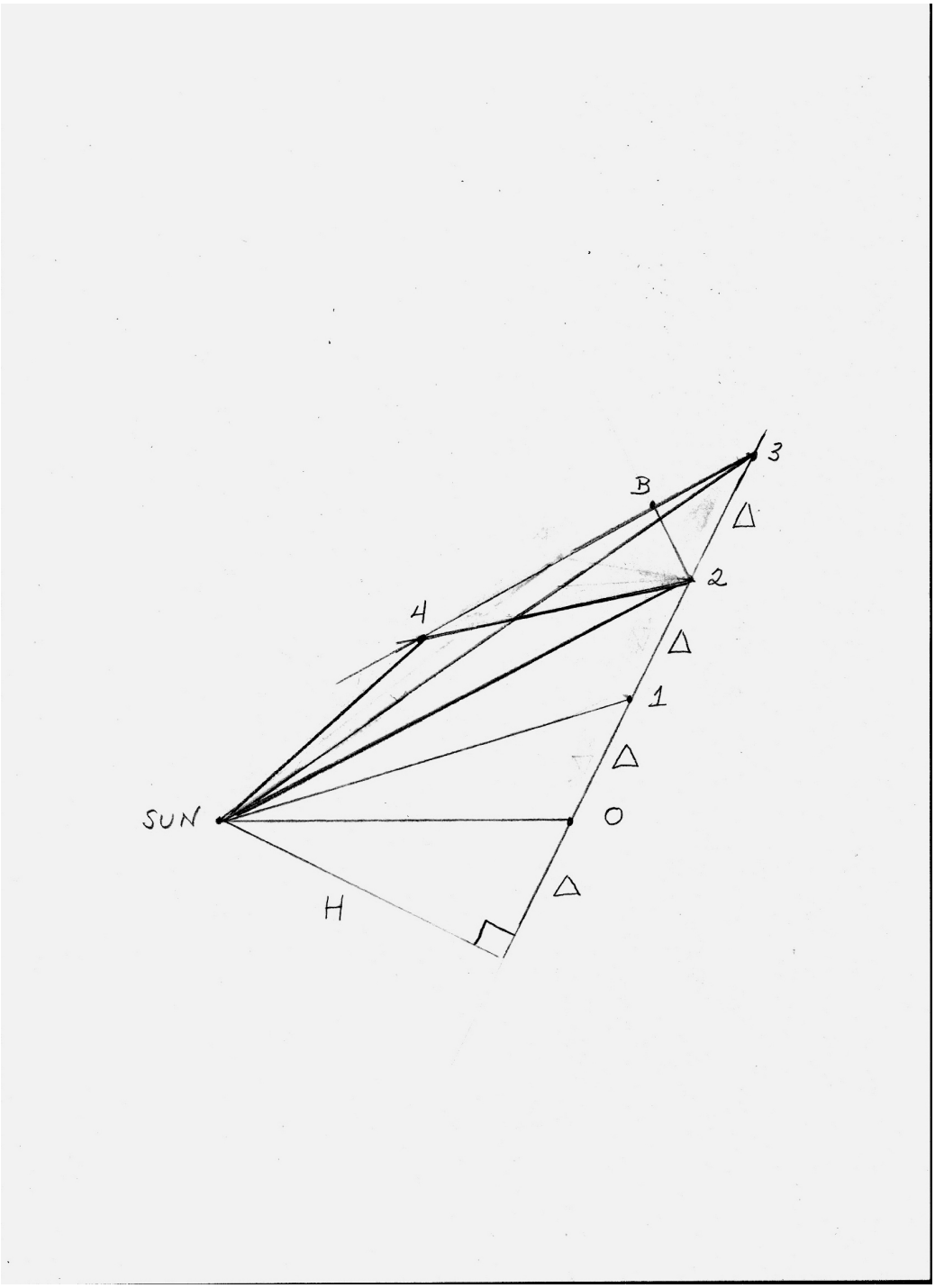



Figure 18.1: Newton's Own Diagram.

# Chapter 19

# Green's Theorem and Related Topics (Chapter 5,6,13)

## 19.1 Introduction

Green's Theorem in two dimensions can be interpreted in two different ways, both leading to important generalizations, namely Stokes's Theorem and the Divergence Theorem. In addition, Green's Theorem has a number of corollaries that involve normal derivatives, Laplacians, and harmonic functions, and that anticipate results in analytic function theory, such as the Cauchy Integral Theorems. A good reference is the book by Flanigan [16].

### 19.1.1 Some Terminology

A subset $D$ of $\mathbb{R}^2$ is said to be *open* if, for every point $x$ in $D$, there is $\epsilon > 0$, such that the ball centered at $x$, with radius $\epsilon$ is completely contained within $D$. The set $D$ is *connected* if it is not the union of two disjoint non-empty open sets. The set $D$ is said to be a *domain* if $D$ is non-empty, open and connected. A subset $B$ of $\mathbb{R}^2$ is *bounded* if it is a subset of a ball of finite radius. The *boundary* of a set $D$, denoted $\partial D$, is the set of all points $x$, in $D$ or not, such that every ball centered at $x$ contains points in $D$ and points not in $D$.

Because we shall be interested in theorems that relate the behavior of functions inside a domain to their behavior on the boundary of that domain, we need to limit our discussion to those domains that have nice boundaries. A *Jordan curve* is a piece-wise smooth closed curve that does not cross

itself. A *Jordan domain* is a bounded domain, whose boundary consists of finitely many, say $k$, disjoint Jordan curves, parameterized in such a way that as a point moves around the curve with increasing parameter, the domain always lies to the left; this is *positive orientation*. Then the domain is called a *$k$-connected Jordan domain*. For example, a ball in $\mathbb{R}^2$ is 1-connected, while an annulus is 2-connected; Jordan domains can have holes in them.

### 19.1.2 Arc-Length Parametrization

Let $C$ be a curve in space with parameterized form

$$\mathbf{r}(t) = (x(t), y(t), z(t)).$$

For each $t$, let $s(t)$ be the distance along the curve from the point $\mathbf{r}(0)$ to the point $\mathbf{r}(t)$. The function $s(t)$ is invertible, so that we can also express $t$ as a function of $s$, $t = t(s)$. Then $s(t)$ is called the *arc-length*. We can then rewrite the parametrization, using as the parameter the variable $s$ instead of $t$; that is, the curve $C$ can be described as

$$\mathbf{r}(s) = \mathbf{r}(t(s)) = (x(t(s)), y(t(s)), z(t(s))). \tag{19.1}$$

Then

$$\mathbf{r}'(t) = \frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt} = (\frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds})\frac{ds}{dt}. \tag{19.2}$$

The vector

$$\mathbf{T}(s) = \frac{d\mathbf{r}}{ds} = (\frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds}) \tag{19.3}$$

has length one, since

$$ds^2 = dx^2 + dy^2 + dz^2, \tag{19.4}$$

and $v = \frac{ds}{dt}$, the *speed* along the curve, satisfies

$$(\frac{ds}{dt})^2 = (\frac{dx}{dt})^2 + (\frac{dy}{dt})^2 + (\frac{dz}{dt})^2. \tag{19.5}$$

## 19.2 Green's Theorem in Two Dimensions

Green's Theorem for two dimensions relates double integrals over domains $D$ to line integrals around their boundaries $\partial D$. Theorems such as this can be thought of as two-dimensional extensions of integration by parts. Green published this theorem in 1828, but it was known earlier to Lagrange and Gauss.

**Theorem 19.1 (Green-2D)** *Let $P(x,y)$ and $Q(x,y)$ have continuous first partial derivatives for $(x,y)$ in a domain $\Omega$ containing both Jordan domain $D$ and $\partial D$. Then*

$$\oint_{\partial D} Pdx + Qdy = \int\int_D \Big(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\Big)dxdy. \tag{19.6}$$

Let the boundary $\partial D$ be the positively oriented parameterized curve

$$\mathbf{r}(t) = (x(t), y(t)).$$

Then, for each $t$, the vector

$$\mathbf{r}'(t) = (x'(t), y'(t))$$

is tangent to the curve at the point $\mathbf{r}(t)$. The vector

$$\mathbf{N}(t) = (y'(t), -x'(t))$$

is perpendicular to $\mathbf{r}'(t)$ and is outwardly normal to the curve at the point $\mathbf{r}(t)$. The integrand on the left side of Equation (19.6) can be written in two ways:

$$Pdx + Qdy = (P,Q) \cdot \mathbf{r}'(t)dt, \tag{19.7}$$

or as

$$Pdx + Qdy = (Q,-P) \cdot \mathbf{N}(t)dt. \tag{19.8}$$

In Equation (19.7) we use the dot product of the vector field $\mathbf{F} = (P,Q)$ with a tangent vector; this point of view will be extended to Stokes's Theorem. In Equation (19.8) we use the dot product of the vector field $\mathbf{G} = (Q,-P)$ with a normal vector; this formulation of Green's Theorem, also called Gauss's Theorem in the plane, will be extended to the Divergence Theorem, also called Gauss's Theorem in three dimensions. Either of these extensions therefore can legitimately be called Green's Theorem in three dimensions.

## 19.3 Proof of Green-2D

First, we compute the line integral $\oint Pdx + Qdy$ around a small rectangle in $D$ and then sum the result over all such small rectangles in $D$. For convenience, we assume the parameter $s$ is arc-length.

Consider the rectangle with vertices $(x_0, y_0)$, $(x_0+\Delta x, y_0)$, $(x_0+\Delta x, y_0+\Delta y)$, and $(x_0, y_0+\Delta y)$, where $\Delta x$ and $\Delta y$ are very small positive quantities. The boundary curve is counter-clockwise. The line integrals along the four sides are as follows:

- The right side:

$$\int_{y_0}^{y_0+\Delta y} Q(x_0+\Delta x, y)dy; \tag{19.9}$$

- The top:

$$\int_{x_0+\Delta x}^{x_0} P(x, y_0+\Delta y)dx; \tag{19.10}$$

- The left side:

$$\int_{y_0+\Delta y}^{y_0} Q(x_0, y)dy; \tag{19.11}$$

- The bottom:

$$\int_{x_0}^{x_0+\Delta x} P(x, y_0)dx. \tag{19.12}$$

Now consider the double integral

$$\int\int_{\Delta} (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y})dx\,dy, \tag{19.13}$$

where $\Delta$ denotes the infinitesmal rectangular region. We write the first half of this integral as

$$\int_{y_0}^{y_0+\Delta y} \Big( \int_{x_0}^{x_0+\Delta x} Q_x(x,y)dx \Big) dy,$$

$$= \int_{y_0}^{y_0+\Delta y} \Big( Q(x_0+\Delta x, y) - Q(x_0, y) \Big) dy,$$

$$= \int_{y_0}^{y_0+\Delta y} Q(x_0+\Delta x, y)dy - \int_{y_0}^{y_0+\Delta y} Q(x_0, y)dy,$$

which is the sum of the two integrals in lines 19.9 and 19.11. In the same way, we can show that the second half of the double integral is equal to the line integrals along the top and bottom of $\Delta$.

Now consider the contributions to the double integral

$$\int\int_{D} (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y})dx\,dy, \tag{19.14}$$

which is the sum of each of the double integrals over all the small rectangles $\Delta$ in $D$. When we add up the contributions of all these infinitesimal rectangles, we need to note that rectangles adjacent to one another contribute

nothing to the line integral from their shared edge, since the unit outward normals are opposite in direction. Consequently, the sum of all the line integrals around the small rectangles reduces to the line integral around the boundary of $D$, since this is the only curve without any shared edges. The double integral in Equation (19.14) is then the line integral around the boundary only, which is the assertion of Green-2D.

Note that we have used the assumption that $Q_x$ and $P_y$ are continuous when we replaced the double integral with iterated single integrals and when we reversed the order of integration.

## 19.4 Extension to Three Dimensions

### 19.4.1 Stokes's Theorem

The first extension of Green-2D to three dimensions that we shall discuss is Stokes's Theorem. The statement of Stokes's Theorem involves a curve $C$ in space and a surface $S$ that is a *capping surface* for $C$. A good illustration of a capping surface is the soap bubble formed when we blow air through a soapy ring; the ring is $C$ and the bubble formed is $S$.

**Theorem 19.2** *Let $C$ be a Jordan curve in space with unit tangent $\mathbf{T}(s) = \frac{d\mathbf{r}}{ds}$ and $S$ a capping surface for $C$, with outward unit normal vectors $\mathbf{n}(s)$. Let $\mathbf{F}(x,y,z) = (P(x,y,z), Q(x,y,z), R(x,y,z))$ be a vector field. The curl of $\mathbf{F}$ is the vector field*

$$\text{curl}(\mathbf{F}) = (R_y - Q_z, P_z - R_x, Q_x - P_y), \tag{19.15}$$

*where $R_y = \frac{\partial R}{\partial y}$. Then*

$$\oint_C \mathbf{F} \cdot \mathbf{T} ds = \int\int_S \mathbf{n} \cdot \text{curl}(\mathbf{F}) dS. \tag{19.16}$$

**Proof:** For convenience, we shall assume that there is a region $D$ in the $x, y$ plane and a real-valued function $f(x,y)$, defined for $(x,y)$ in $D$, such that the surface $S$ is the graph of $f(x,y)$, that is, each point $(x,y,z)$ on $S$ has the form $(x,y,z) = (x,y,f(x,y))$. The boundary curve of $D$, denoted $\partial D$, is the curve in the $x, y$ plane directly below the curve $C$.

Since we can write

$$\mathbf{F} = (P,Q,R) = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$$

and

$$\nabla \times \mathbf{F} = \nabla \times (P\mathbf{i}) + \nabla \times (Q\mathbf{j}) + \nabla \times (R\mathbf{k}),$$

we focus on proving the theorem for the simpler case of $Q = R = 0$. Note that we have

$$\nabla \times (P\mathbf{i}) = \frac{\partial P}{\partial z}\mathbf{j} - \frac{\partial P}{\partial y}\mathbf{k},$$

so that

$$\nabla \times (P\mathbf{i}) \cdot \mathbf{n} = \frac{\partial P}{\partial z}\mathbf{n} \cdot \mathbf{j} - \frac{\partial P}{\partial y}\mathbf{n} \cdot \mathbf{k}. \tag{19.17}$$

The vector $\mathbf{r}(x, y, z) = (x, y, f(x, y))$ from the origin to the point $(x, y, z)$ on the surface $S$ then has

$$\frac{\partial \mathbf{r}}{\partial y} = \mathbf{j} + \frac{\partial f}{\partial y}\mathbf{k}.$$

The vector $\frac{\partial \mathbf{r}}{\partial y}$ is tangent to the surface at $(x, y, z)$, and so it is perpendicular to the unit outward normal. This means that

$$\mathbf{n} \cdot \mathbf{j} + \frac{\partial f}{\partial y}\mathbf{n} \cdot \mathbf{k} = 0,$$

so that

$$\mathbf{n} \cdot \mathbf{j} = -\frac{\partial f}{\partial y}\mathbf{n} \cdot \mathbf{k}. \tag{19.18}$$

Therefore, using Equations (19.17) and (19.18), we have

$$\nabla \times (P\mathbf{i}) \cdot \mathbf{n}\, dS = -\Big(\frac{\partial P}{\partial z}\frac{\partial f}{\partial y} + \frac{\partial P}{\partial y}\Big)\mathbf{n} \cdot \mathbf{k} dS. \tag{19.19}$$

Note, however, that

$$\frac{\partial P}{\partial z}\frac{\partial f}{\partial y} + \frac{\partial P}{\partial y} = \frac{\partial F}{\partial y},$$

where $F(x, y) = P(x, y, f(x, y))$. Therefore, recalling that

$$\mathbf{n} \cdot \mathbf{k}\, dS = dxdy,$$

we get

$$\nabla \times (P\mathbf{i}) \cdot \mathbf{n}\, dS = -\frac{\partial F}{\partial y} dxdy. \tag{19.20}$$

By Green 2-D, we have

$$\int\int_S \nabla \times (P\mathbf{i}) \cdot \mathbf{n}\, dS = \int\int_D -\frac{\partial F}{\partial y} dxdy = \oint_{\partial D} F dx.$$

But we also have

$$\oint_{\partial D} F dx = \oint_C P dx,$$

since $F(x, y) = P(x, y, f(x, y))$. Similar calculations for the other two coordinate-direction components establish the assertion of the theorem. ∎

Suppose that $\mathbf{F} = (P, Q, 0)$ and the surface $S$ is a Jordan domain $D$ in $\mathbb{R}^2$, with $C = \partial D$. Then

$$\text{curl}(\mathbf{F}) = (0, 0, Q_x - P_y),$$

and $\mathbf{n} = (0, 0, 1)$. Therefore,

$$\mathbf{n} \cdot \text{curl}(\mathbf{F}) = Q_x - P_y.$$

Also,

$$\mathbf{F} \cdot \mathbf{T} = P dx + Q dy.$$

We see then that Stokes's Theorem has Green-2D as a special case.

Because the curl of a vector field is defined only for three-dimensional vector fields, it is not obvious that the curl and Stokes's Theorem extend to higher dimensions. They do, but the extensions involve more complicated calculus on manifolds and the integration of $(n-1)$-forms over a suitably oriented boundary of an oriented $n$-manifold; see Fleming [17] for the details.

### 19.4.2 The Divergence Theorem

Equation (19.8) suggests that we consider surface integrals of functions having the form $\mathbf{F} \cdot \mathbf{n}$, where $\mathbf{n}$ is the outward unit normal to the surface at each point. The Divergence Theorem, also called Gauss's Theorem in three dimensions, is one result in this direction.

**Theorem 19.3** *Let $S$ be a closed surface enclosing the volume $V$. Let $\mathbf{F} = (P, Q, R)$ be a vector field with divergence*

$$\text{div}(\mathbf{F}) = P_x + Q_y + R_z.$$

*Then*

$$\int\int_S \mathbf{F} \cdot \mathbf{n}\, dS = \int\int\int_V \text{div}(\mathbf{F})\, dV. \tag{19.21}$$

**Proof:** We first prove the theorem for a small cube with vertices $(x, y, z)$, $(x, y + \Delta y, z), (x, y, z + \Delta z)$ and $(x, y + \Delta y, z + \Delta z)$ forming the left side wall, and the vertices $(x + \Delta x, y, z), (x + \Delta x, y + \Delta y, z), (x + \Delta x, y, z + \Delta z)$ and $(x + \Delta x, y + \Delta y, z + \Delta z)$ forming the right side wall. The unit outward normal for the side wall containing the first four of the eight vertices is $n = (-1, 0, 0)$; for the other side wall, it is $n = (1, 0, 0)$. For the first side wall the flux is the normal component of the field times the area of the wall, or

$$-P(x, y, z)\Delta y\, \Delta z,$$

while for the second side wall, it is

$$P(x + \Delta x, y, z)\Delta y\, \Delta z.$$

The total outward flux through these two walls is then

$$\Big(P(x + \Delta x, y, z) - P(x, y, z)\Big)\Delta y\, \Delta z,$$

or

$$\Big(\frac{P(x + \Delta x, y, z) - P(x, y, z)}{\Delta x}\Big)\Delta x \Delta y \Delta z.$$

Taking limits, we get

$$\frac{\partial P}{\partial x}(x, y, z)\, dV.$$

We then perform the same calculations for the other four walls. Finally, having proved the theorem for small cubes, we view the entire volume as a sum of small cubes and add up the total flux for all the cubes. Because outward flux from one cube's wall is inward flux for its neighbor, they cancel out, except when a wall has no neighbor; this means that the only outward flux that remains is through the surface. This is what the theorem says. ∎

If we let $R = 0$ and imagine the volume shrinking down to a two-dimensional planar domain $D$, with $S$ compressing down to its boundary, $\partial D$, the unit normal vector becomes

$$\mathbf{n} = (\frac{dy}{ds}, -\frac{dx}{ds}),$$

and Equation (19.21) reduces to Equation (19.6).

## 19.5 When is a Vector Field a Gradient Field?

The following theorem is classical and extends the familiar "test for exactness" .

**Theorem 19.4** *Let $\mathbf{F} : D \subseteq \mathbb{R}^N \to \mathbb{R}^N$ be continuously differentiable on an open convex set $D_0 \subseteq D$, with*

$$\mathbf{F}(x) = (F_1(x), F_2(x), ..., F_N(x)).$$

*Then there is a differentiable function $f : D_0 \to \mathbb{R}^N$ such that $\mathbf{F}(x) = \nabla f(x)$ for all $x$ in $D_0$ if and only if*

$$\frac{\partial F_m}{\partial x_n} = \frac{\partial F_n}{\partial x_m},$$

*for all $m$ and $n$; in other words, the Jacobian matrix of $\mathbf{F}$ is symmetric.*

**Proof:** If $\mathbf{F}(x) = \nabla f(x)$ for all $x$ in $D_0$ and is continuously differentiable, then the second partial derivatives of $f(x)$ are continuous, so that the mixed second partial derivatives of $f(x)$ are independent of the order of differentiation.

For notational convenience, we present the proof of the converse only for the case of $N = 3$; the proof is the same in general.

Without loss of generality, we assume that the origin is a member of the set $D_0$. Define $f(x, y, z)$ by

$$f(x, y, z) = \int_0^x F_1(u, 0, 0)du + \int_0^y F_2(x, u, 0)du + \int_0^z F_3(x, y, u)du.$$

We prove that $\frac{\partial f}{\partial x}(x, y, z) = F_1(x, y, z)$.

The partial derivative of the first integral, with respect to $x$, is $F_1(x, 0, 0)$. The partial derivative of the second integral, with respect to $x$, obtained by differentiating under the integral sign, is

$$\int_0^y \frac{\partial F_2}{\partial x}(x, u, 0)du,$$

which, by the symmetry of the Jacobian matrix, is

$$\int_0^y \frac{\partial F_1}{\partial y}(x, u, 0)du = F_1(x, y, 0) - F_1(x, 0, 0).$$

The partial derivative of the third integral, with respect to $x$, obtained by differentiating under the integral sign, is

$$\int_0^z \frac{\partial F_3}{\partial x}(x, y, u)du,$$

which, by the symmetry of the Jacobian matrix, is

$$\int_0^z \frac{\partial F_1}{\partial z}(x, y, u)du = F_1(x, y, z) - F_1(x, y, 0).$$

We complete the proof by adding these three integral values. Similar calculations show that $\nabla f(x) = \mathbf{F}(x)$. ∎

Theorem 19.4 tells us that, for a three-dimensional field

$$\mathbf{F}(x, y, z) = (F_1(x, y, z), F_2(x, y, z), F_3(x, y, z)),$$

there is a real-valued function $f(x, y, z)$ with $\mathbf{F}(x, y, z) = \nabla f(x, y, z)$ for all $(x, y, z)$ if and only if the curl of $\mathbf{F}(x, y, z)$ is zero. It follows from Stokes's Theorem that the integral $\oint_C \mathbf{F} \cdot \mathbf{T}ds$ is zero for every closed curve $C$. Consequently, for any path $C$ connecting points $A$ and $B$, the integral

$\int_C \mathbf{F} \cdot \mathbf{T} ds$ is independent of the path and depends only on the points $A$ and $B$; then we can write

$$\int_C \mathbf{F} \cdot \mathbf{T} ds = \int_A^B \mathbf{F} \cdot \mathbf{T} ds. \tag{19.22}$$

In addition, the potential function $f(x, y, z)$ can be chosen to be

$$f(x, y, z) = \int_{(x_0,y_0,z_0)}^{(x,y,z)} \mathbf{F} \cdot \mathbf{T} ds, \tag{19.23}$$

where $(x_0, y_0, z_0)$ is an arbitrarily selected point in space.

When $\mathbf{F}(x, y, z)$ denotes a force field, the integral $\int_A^B \mathbf{F} \cdot \mathbf{T} ds$ is the work done against the force in moving from $A$ to $B$. When $\mathbf{F} = \nabla f$, this work is simply the change in the potential function $f$. Such force fields are called *conservative*.

## 19.6 Corollaries of Green-2D

### 19.6.1 Green's First Identity

Let $u(x, y)$ be a differentiable real-valued function of two variables, with gradient

$$\nabla u(x, y) = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}).$$

Let $D$ be a Jordan domain with boundary $C = \partial D$. The directional derivative of $u$, in the direction of the unit outward normal $\mathbf{n}$, is

$$\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}.$$

When the curve $C$ is parameterized by arc-length, the unit outward normal is

$$\mathbf{n} = (\frac{dy}{ds}, -\frac{dx}{ds}),$$

and

$$\oint_C \frac{\partial u}{\partial \mathbf{n}} \, ds = \oint_C -u_y dx + u_x dy. \tag{19.24}$$

**Theorem 19.5 (Green I)** *Let $\nabla^2 q$ denote the Laplacian of the function $q(x, y)$, that is,*

$$\nabla^2 q = q_{xx} + q_{yy}.$$

*Then*

$$\int\int_D (\nabla p) \cdot (\nabla q) \, dxdy = \oint_C p \frac{\partial q}{\partial \mathbf{n}} \, ds - \int\int_D p \nabla^2 q \, dxdy. \tag{19.25}$$

**Proof:** Evaluate the line integral using Green-2D, with $P = -pq_y$ and $Q = pq_x$. ∎

### 19.6.2 Green's Second Identity

An immediate corollary is Green's Second Identity (Green II).

**Theorem 19.6 (Green II)**

$$\oint_C p\frac{\partial q}{\partial \mathbf{n}} - q\frac{\partial p}{\partial \mathbf{n}}\, ds = \int\int_D p\nabla^2 q - q\nabla^2 p\, dxdy. \tag{19.26}$$

### 19.6.3 Inside-Outside Theorem

The Inside-Outside Theorem, which is a special case of Gauss's Theorem in the plane, follows immediately from Green II.

**Theorem 19.7 (Inside-Outside Theorem)**

$$\oint_C \frac{\partial q}{\partial \mathbf{n}}\, ds = \int\int_D \nabla^2 q\, dxdy. \tag{19.27}$$

### 19.6.4 Green's Third Identity

Green's Third Identity (Green III) is more complicated than the previous ones. Let $w$ be any point inside the Jordan domain $D$ in $\mathbb{R}^2$ and hold $w$ fixed. For variable $z$ in the plane, let $r = |z - w|$. A function is said to be *harmonic* if its Laplacian is identically zero. We show now that the function $p(z) = \log r$ is harmonic for any $z$ in any domain that does not contain $w$. With $z = (x, y)$ and $w = (a, b)$, we have

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

so that

$$p(z) = p(x,y) = \frac{1}{2}\log\Big((x-a)^2 + (y-b)^2\Big).$$

Then

$$p_x(x,y) = \frac{x-a}{(x-a)^2 + (y-b)^2},$$

$$p_{xx}(x,y) = \frac{(y-b)^2 - (x-a)^2}{((x-a)^2 + (y-b)^2)^2},$$

$$p_y(x,y) = \frac{y-b}{(x-a)^2 + (y-b)^2},$$

and

$$p_{yy}(x,y) = \frac{(x-a)^2 - (y-b)^2}{((x-a)^2 + (y-b)^2)^2}.$$

Clearly, we have

$$p_{xx} + p_{yy} = 0,$$

and so $p$ is harmonic in any region not including $w$.

The theorem is the following:

**Theorem 19.8 (Green III)**

$$q(w) = \frac{1}{2\pi} \int \int_D \log r \, \nabla^2 q \, dxdy$$

$$-\frac{1}{2\pi} \oint_C \log r \, \frac{\partial q}{\partial \mathbf{n}} \, ds + \frac{1}{2\pi} \oint_C q \frac{\partial \log r}{\partial \mathbf{n}} \, ds. \tag{19.28}$$

The two line integrals in Equation (19.28) are known as the *logarithmic single-layer potential* and *logarithmic double-layer potential*, respectively, of the function $q$.

Notice that we cannot apply Green II directly to the domain $D$, since $\log r$ is not defined at $z = w$. The idea is to draw a small circle $C'$ centered at $w$, with interior $D'$ and consider the new domain that is the original $D$, without the ball $D'$ around $w$ and its boundary; the new domain has a hole in it, but that is acceptable. Then apply Green II, and finally, let the radius of the ball go to zero. There are two key steps in the calculation.

First, we use the fact that, for the small circle, $\mathbf{n} = \mathbf{r}/\|\mathbf{r}\|$ to show that

$$\frac{\partial p}{\partial \mathbf{n}} = \nabla p \cdot \mathbf{n} = \frac{1}{\rho},$$

where $\rho$ is the radius of the small circle $C'$ centered at $w$ and $\mathbf{r} = (z - w)$ is the vector from $w$ to $z$. Then

$$p(z) = \frac{1}{2} \log(\|\mathbf{r}\|^2),$$

so that

$$\nabla p(z) = \frac{1}{2} \frac{1}{\|\mathbf{r}\|^2} \nabla \|\mathbf{r}\|^2 = \frac{1}{\|\mathbf{r}\|^2} \mathbf{r}.$$

Therefore, for $z$ on $C'$, we have

$$\frac{\partial p}{\partial \mathbf{n}} = \nabla p(z) \cdot \mathbf{n} = \frac{1}{\rho}.$$

Then

$$\oint_{C'} q \frac{\partial p}{\partial \mathbf{n}} ds = \frac{1}{\rho} \oint_{C'} q ds,$$

which, as the radius of $C'$ goes to zero, is just $2\pi q(w)$.

Second, we note that the function $\frac{\partial q}{\partial \mathbf{n}}$ is continuous, and therefore bounded by some constant $K > 0$ on the circle $C'$; the constant $K$ can be chosen to be independent of $\rho$, for $\rho$ sufficiently close to zero. Consequently, we have

$$\left| \oint_{C'} \log r \, \frac{\partial q}{\partial \mathbf{n}} \, ds \right| \leq K |\log \rho \oint_{C'} ds| = 2\pi K |\rho \log \rho|.$$

Since $\rho \log \rho$ goes to zero, as $\rho$ goes to zero, this integral vanishes, in the limit.

Equation (19.28) tells us that if $q$ is a harmonic function in $D$, then its value at any point $w$ inside $D$ is completely determined by what the functions $q$ and $\frac{\partial q}{\partial \mathbf{n}}$ do on the boundary $C$. Note, however, that the normal derivative of $q$ depends on values of $q$ near the boundary, not just on the boundary. In fact, $q(w)$ is completely determined by $q$ alone on the boundary, via

$$q(w) = -\frac{1}{2\pi} \oint_C q(z) \frac{\partial}{\partial \mathbf{n}} G(z, w)\, ds,$$

where $G(z, w)$ is the *Green's function* for the domain $D$.

According to the heat equation, the temperature $u(x, y, t)$ in a two-dimensional region at time $t$ is governed by the partial differential equation

$$\frac{\partial u}{\partial t} = c\nabla^2 u,$$

for some constant $c > 0$. When a steady-state temperature has been reached, the function $u(x, y, t)$ no longer depends on $t$ and the resulting function $f(x, y)$ of $(x, y)$ only satisfies $\nabla^2 f = 0$; that is, $f(x, y)$ is harmonic. Imagine the region being heated by maintaining a temperature distribution around the boundary of the region. It is not surprising that such a steady-state temperature distribution throughout the region should be completely determined by the temperature distribution around the boundary of the region.

## 19.7 Application to Complex Function Theory

In Complex Analysis the focus is on functions $f(z)$ where both $z$ and $f(z)$ are complex variables, for each $z$. Because $z = x + iy$, for $x$ and $y$ real variables, we can always write

$$f(z) = u(x, y) + iv(x, y),$$

where $u(x, y)$ and $v(x, y)$ are both real-valued functions of two real variables. So it looks like there is nothing new here; complex function theory looks like the theory of any two real-valued functions glued together to form a complex-valued function. There is an important difference, however.

The most important functions in complex analysis are the functions that are *analytic* in a domain $D$ in the complex plane. Such functions will have the property that, for any closed curve $C$ in $D$,

$$\oint_C f(z) dz = 0.$$

Writing $dz = dx + idy$, we have

$$\oint_C f(z)dz = \oint_C (udx - vdy) + i \oint_C (vdx + udy).$$

From Green 2-D it follows that we want $u_x = v_y$ and $u_y = -v_x$; these are called the Cauchy-Riemann (CR) equations. The CR equations are a consequence of the differentiability of $f(z)$. The point here is that complex function theory is not just the theory of unrelated real-valued functions glued together; the functions $u(x, y)$ and $v(x, y)$ must be related in the sense of the CR equations in order for the function $f(z)$ to fit the theory.

If $f(z)$ is an analytic function of the complex variable $z$, then, because of the CR equations, the real and imaginary parts of $f(z)$, the functions $u(x, y)$ and $v(x, y)$, are real-valued harmonic functions of two variables. Using Green III, we can obtain *Cauchy's Integral Formula*, which shows that the value of $f$ at any point $w$ within the domain $D$ is determined by the value of the function at the points $z$ of the boundary:

$$f(w) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - w} dz. \tag{19.29}$$

This formula occurs in complex function theory under slightly weaker assumptions than we use here. We shall assume that $f(z)$ is continuously differentiable, so that the real and imaginary parts of $f$ are continuously differentiable; we need this for Green 2-D. In complex function theory, it is shown that the continuity of $f'(z)$ is a consequence of analyticity. According to complex function theory, we may, without loss of generality, consider only the case in which $C$ is the circle of radius $\rho$ centered at $w = (a, b)$, and $D$ is the region enclosed by $C$, which is what we shall do.

We know from Green III that

$$u(w) = \frac{1}{2\pi} \int \int_D \log r \, \nabla^2 u \, dxdy$$

$$-\frac{1}{2\pi} \oint_C \log r \, \frac{\partial u}{\partial \mathbf{n}} \, ds + \frac{1}{2\pi} \oint_C u \frac{\partial \log r}{\partial \mathbf{n}} \, ds, \tag{19.30}$$

with a similar expression involving $v$. Because $u$ is harmonic, Equation (19.30) reduces to

$$u(w) = -\frac{1}{2\pi} \oint_C \log r \, \frac{\partial u}{\partial \mathbf{n}} \, ds + \frac{1}{2\pi} \oint_C u \frac{\partial \log r}{\partial \mathbf{n}} \, ds \,, \tag{19.31}$$

with a similar expression involving the function $v$.

Consider the first line integral in Equation (19.31),

$$\frac{1}{2\pi} \oint_C \log r \, \frac{\partial u}{\partial \mathbf{n}} \, ds. \tag{19.32}$$

Since $r = \rho$ for all $z$ on $C$, this line integral becomes

$$\frac{1}{2\pi} \log \rho \oint_C \frac{\partial u}{\partial \mathbf{n}}\, ds. \tag{19.33}$$

But, by the Inside-Outside Theorem, and the fact that $u$ is harmonic, we know that

$$\oint_C \frac{\partial u}{\partial \mathbf{n}}\, ds = \int \int_D \nabla^2 u \, dxdy \, = 0. \tag{19.34}$$

So we need only worry about the second line integral in Equation (19.31), which is

$$\frac{1}{2\pi} \oint_C u \frac{\partial \log r}{\partial \mathbf{n}}\, ds. \tag{19.35}$$

We need to look closely at the term

$$\frac{\partial \log r}{\partial \mathbf{n}}.$$

First, we have

$$\frac{\partial \log r}{\partial \mathbf{n}} = \frac{1}{2} \frac{\partial \log r^2}{\partial \mathbf{n}}. \tag{19.36}$$

The function $\log r^2$ can be viewed as

$$\log r^2 = \log \Big( \mathbf{a} \cdot \mathbf{a} \Big), \tag{19.37}$$

where $\mathbf{a}$ denotes $z - w$, thought of as a vector in $\mathbb{R}^2$. Then

$$\nabla \log r^2 = \nabla \log \Big( \mathbf{a} \cdot \mathbf{a} \Big) = 2 \frac{\mathbf{a}}{||\mathbf{a}||^2}. \tag{19.38}$$

Because $C$ is a circle centered at $w$, the unit outward normal at $z$ on $C$ is

$$\mathbf{n} = \frac{\mathbf{a}}{||\mathbf{a}||}. \tag{19.39}$$

Putting all this together, we find that

$$\frac{\partial \log r}{\partial \mathbf{n}} = \frac{1}{||\mathbf{a}||} = \frac{1}{|z - w|}. \tag{19.40}$$

Therefore, Green III tells us that

$$u(w) = \frac{1}{2\pi} \oint_C \frac{u(z)}{|z - w|} ds, \tag{19.41}$$

with a similar expression involving $v$. There is one more step we must take to get to the Cauchy Integral Formula.

We can write $z - w = \rho e^{i\theta}$ for $z$ on $C$. Therefore,

$$\frac{dz}{d\theta} = \rho i e^{i\theta}. \tag{19.42}$$

The arc-length $s$ around the curve $C$ is $s = \rho\theta$, so that

$$\frac{ds}{d\theta} = \rho. \tag{19.43}$$

Therefore, we have

$$\theta = \frac{s}{\rho}, \tag{19.44}$$

and

$$z - w = \rho e^{is/\rho}. \tag{19.45}$$

Then,

$$\frac{dz}{ds} = ie^{is/\rho}, \tag{19.46}$$

or

$$ds = \frac{1}{i} e^{-i\theta} dz. \tag{19.47}$$

Substituting for $ds$ in Equation (19.41) and in the corresponding equation involving $v$, and using the fact that

$$|z - w| e^{i\theta} = z - w, \tag{19.48}$$

we obtain Cauchy's Integral Formula (19.29).

For a brief discussion of an interesting application of Maxwell's equations, see the chapter on Invisibility.

## 19.8 The Cauchy-Riemann Equations Again

Let $\mathbf{r}(t) = (x(t), y(t))$ be a curve in $\mathbb{R}^2$, which we view as a parameterized set of complex numbers; that is, we write $z(t) = x(t) + iy(t)$. Then $z'(t) = (x'(t) + iy'(t))$. Let $f(z)$ be analytic in a domain $D$ and write

$$f(z) = u(x, y) + iv(x, y),$$

and

$$g(t) = f(z(t)) = u(x(t), y(t)) + iv(x(t), y(t)).$$

Then, with $f'(z(t)) = c(t) + id(t)$ and suppressing the $t$, we have

$$g'(t) = f'(z(t))z'(t) = (cx' - dy') + i(cy' + dx'). \tag{19.49}$$

We also have

$$g'(t) = (u_x x' + u_y y') + i(v_x x' + v_y y'). \tag{19.50}$$

Comparing Equations (19.49) and (19.50), we find that

$$cx' - dy' = u_x x' + u_y y',$$

and

$$cy' + dx' = v_x x' + v_y y'.$$

Since these last two equations must hold for any curve $\mathbf{r}(t)$, they must hold when $x'(t) = 0$ for all $t$, as well as when $y'(t) = 0$ for all $t$. It follows that $c = u_x$, $d = -u_y$, $c = v_y$, and $d = v_x$, from which we can get the Cauchy-Riemann equations easily.

# Chapter 20

# Introduction to Complex Analysis (Chapter 13)

## 20.1 Introduction

The material in this chapter is taken mainly from Chapter 13 of the text. In some cases, the ordering of topics has been altered slightly.

## 20.2 Complex-valued Functions of a Complex Variable

In complex analysis the focus is on functions $w = f(z)$, where both $z$ and $w$ are complex numbers. With $z = x + iy$, for $x$ and $y$ real, it follows that

$$w = f(z) = u(x, y) + iv(x, y), \tag{20.1}$$

with both $u(x, y)$ and $v(x, y)$ real-valued functions of the two real variables $x$ and $y$. Since $z_x = 1$ and $z_y = i$, the differential $dz$ is

$$dz = dx + idy. \tag{20.2}$$

For any curve $C$ in the complex plane the line integral of $f(z)$ along $C$ is defined as

$$\int_C f(z)dz = \int_C (u + iv)(dx + idy) = \int_C udx - vdy + i\int_C vdx + udy \tag{20.3}$$

## 20.3 Differentiability

The derivative of the function $f(z)$ at the point $z$ is defined to be

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}, \tag{20.4}$$

whenever this limit exists. Note that $\Delta z = \Delta x + i\Delta y$, so that $z + \Delta z$ is obtained by moving a small distance away from $z$, by $\Delta x$ in the horizontal direction and $\Delta y$ in the vertical direction. When the limit does exist, the function $f(z)$ is said to be *differentiable* or *analytic* at the point $z$; the function is then continuous at $z$ as well.

For real-valued functions of a real variable, requiring that the function be differentiable is not a strong requirement; however, for the functions $w = f(z)$ it certainly is. What makes differentiability a strong condition is that, for the complex plane, we can move away from $z$ in infinitely many directions, unlike in the real case, where all we can do is to move left or right away from $x$. As we shall see, in order for $f(z)$ to be differentiable, the functions $u(x, y)$ and $v(x, y)$ must be related in a special way, called the Cauchy-Riemann equations.

## 20.4 The Cauchy-Riemann Equations

We can rewrite the limit in Equation (20.4) as

$$f'(z) = \lim_{\Delta x, \Delta y \to 0} \frac{(u(x + \Delta x, y + \Delta y) - u(x, y)) + i(v(x + \Delta x, y + \Delta y) - v(x, y))}{\Delta x + i\Delta y}. \tag{20.5}$$

Suppose now that we take $\Delta y = 0$, so that $\Delta z = \Delta x$. Then the limit in Equation (20.5) becomes

$$f'(z) = \lim_{\Delta x \to 0} \frac{(u(x + \Delta x, y) - u(x, y)) + i(v(x + \Delta x, y) - v(x, y))}{\Delta x}. \tag{20.6}$$

Then

$$f'(z) = u_x + iv_x. \tag{20.7}$$

On the other hand, if we take $\Delta x = 0$, so that $\Delta z = i\Delta y$, we get

$$f'(z) = \lim_{\Delta y \to 0} \frac{(u(x, y + \Delta y) - u(x, y)) + i(v(x, y + \Delta y) - v(x, y))}{i\Delta y}. \tag{20.8}$$

Then

$$f'(z) = v_y - iu_y. \tag{20.9}$$

It follows that

$$u_x = v_y, \text{ and } u_y = -v_x. \tag{20.10}$$

For example, suppose that

$$f(z) = z^2 = (x+iy)^2 = (x^2 - y^2) + i(2xy).$$

The derivative is $f'(z) = 2z$ and $u_x = v_y = 2x$, while $u_y = -v_x = -2y$.

Since the Cauchy-Riemann equations must hold if $f(z)$ is differentiable, we can use them to find functions that are not differentiable. For example, if $f(z)$ is real-valued and not constant, then $f(z)$ is not differentiable. In this case $v_x = v_y = 0$, but $u_x$ and $u_y$ cannot both be zero. Another example is the function $f(z) = \overline{z} = x - iy$. Here $u_x = 1$, but $v_y = -1$.

However, most of the real-valued differentiable functions of $x$ can be extended to complex-valued differentiable functions of $z$ simply by replacing $x$ with $z$ in the formulas. For example, $\sin z$, $\cos z$, and $e^z$ are differentiable, with derivatives $\cos z$, $-\sin z$, and $e^z$, respectively. The function $\frac{2z-3}{3z-6}$ is differentiable throughout any region that does not include the point $z = 2$, and we obtain its derivative using the usual quotient rule.

One consequence of the Cauchy-Riemann equations is that the functions $u$ and $v$ are harmonic, that is

$$u_{xx} + u_{yy} = 0, \text{ and } v_{xx} + v_{yy} = 0.$$

## 20.5 Integration

Suppose that $f(z)$ is differentiable on and inside a simple closed curve $C$, and suppose that the partial derivatives $u_x$, $u_y$, $v_x$, and $v_y$ are continuous. Using Equation (20.3) and applying Green 2-D separately to both of the integrals, we get

$$\oint_C f(z)dz = -\int\int_D (v_x + u_y)dxdy + i\int\int_D (u_x - v_y)dxdy, \tag{20.11}$$

where $D$ denotes the interior of the region whose boundary is the curve $C$. The Cauchy-Riemann equations tell us that both of the double integrals are zero. Therefore, we may conclude that $\oint_C f(z)dz = 0$ for all such simple closed curves $C$.

It is important to remember that Green 2-D is valid for regions that have holes in them; in such cases the boundary $C$ of the region consists of more than one simple closed curve, so the line integral in Green 2-D is along each of these curves separately, with the orientation such that the region remains to the left as we traverse the line.

In a course on complex analysis it is shown that this theorem holds without the assumptions that the first partial derivatives are continuous;

this is the Cauchy-Goursat Theorem. This theorem greatly improves the theory of complex analysis, as we shall see.

## 20.6 Some Examples

Consider the function $f(z) = (z-a)^n$, where $n$ is an integer. If $n$ is a non-negative integer, then $f(z)$ is differentiable everywhere and

$$\oint_C (z-a)^n dz = 0,$$

for every simple closed curve. But what happens when $n$ is negative?

Let $C$ be a simple closed curve with $z = a$ inside $C$. Using the Cauchy-Goursat Theorem, we may replace the integral around the curve $C$ with the integral around the circle centered at $a$ and having radius $\epsilon$, where $\epsilon$ is a small positive number. Then $z = a + \epsilon e^{i\theta}$ for all $z$ on the small circle, and

$$dz = i\epsilon e^{i\theta} d\theta.$$

Then

$$\oint_C (z-a)^n dz = \int_0^{2\pi} (\epsilon e^{i\theta})^n i\epsilon e^{i\theta} d\theta = i\epsilon^n \int_0^{2\pi} e^{i(n+1)\theta} d\theta.$$

Therefore, if $n+1$ is not zero, we have

$$\oint_C (z-a)^n dz = 0,$$

and if $n + 1 = 0$

$$\oint_C (z-a)^{-1} dz = 2\pi i.$$

## 20.7 Cauchy's Integral Theorem

Suppose that $C_1$ and $C_2$ are circles in the complex plane, with common center $z = a$; let the radius of $C_2$ be the smaller. Suppose that $f(z)$ is differentiable in a region containing the annulus whose boundaries are $C_1$ and $C_2$, and that $C$ is a simple closed curve in the annulus that surrounds the curve $C_2$. Then

$$\oint_{C_1} f(z)dz = \oint_{C_2} f(z)dz = \oint_C f(z)dz. \tag{20.12}$$

The proof of this result is given, almost, in the worked problem 13.12 on p. 299 of the text. The difficulty with the proof given there is that the

curve he describes as AQPABRSTBA is not a simple closed curve; this curve repeats the part AB in both directions, so crosses itself. A more rigorous proof replaces the return path BA with one very close to BA, call it B'A'. Then the result is obtained by taking the limit, as the path B'A' approaches BA.

One way to think of this theorem is that if we can morf the curve $C$ into $C_1$ or $C_2$ without passing through a point where $f(z)$ is not differentiable, then the integrals are the same. Now we use this fact to prove the Cauchy Integral Theorem.

Let $f(z)$ be differentiable on and inside a simple closed curve $C$, and let $a$ be a point inside $C$. Then Cauchy's Integral Theorem tells us that

$$f(a) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z-a} dz. \tag{20.13}$$

Again, we may replace the curve $C$ with the circle centered at $a$ and having radius $\epsilon$, for some small positive $\epsilon$. Then

$$\oint_C \frac{f(z)}{z-a} dz = \int_0^{2\pi} f(a+\epsilon e^{i\theta})(\epsilon e^{i\theta})^{-1} i\epsilon e^{i\theta} d\theta = i \int_0^{2\pi} f(a+\epsilon e^{i\theta}) d\theta.$$

Letting $\epsilon \downarrow 0$, we get

$$\oint_C \frac{f(z)}{z-a} dz = 2\pi i f(a).$$

Differentiating with respect to $a$ in Cauchy's Integral Theorem we find that

$$f'(a) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z-a)^2} dz, \tag{20.14}$$

and more generally

$$f^{(n)}(a) = \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z-a)^{n+1}} dz. \tag{20.15}$$

So not only is $f(z)$ differentiable, but it has derivatives of all orders. This is one of the main ways in which complex analysis differs from real analysis.

## 20.8 Taylor Series Expansions

When we study Taylor series expansions for real-valued functions of a real variable $x$, we find that the function $f(x) = 1/(x^2+1)$ poses a bit of a mystery. We learn that

$$1/(x^2+1) = 1 - x^2 + x^4 - x^6 + ...,$$

and that this series converges for $|x| < 1$ only. But why? The function $f(x)$ is differentiable for all $x$, so why shouldn't the Taylor series converge

for all $x$, as the Taylor series for $\sin x$ or $e^x$ do. The answer comes when we consider the complex extension, $f(z) = 1/(z^2+1)$. This function is undefined at $z = i$ and $z = -i$. The Taylor series for $f(z)$ converges in the largest circle centered at $a = 0$ that does not contain a point where $f(z)$ fails to be differentiable, so must converge only within a circle of radius one. This must apply on the real line as well.

Let $f$ be differentiable on and inside a circle $C$ centered at $a$ and let $a + h$ be inside $C$. Then

$$f(a+h) = a_0 + a_1 h + a_2 h^2 + ..., \tag{20.16}$$

where $a_n = f^{(n)}(a)/n!$. The Taylor series converges in the largest circle centered at $z = a$ that does not include a point where $f(z)$ is not differentiable.

Taking $a+h$ inside $C$ and using the Cauchy Integral Theorem, we have

$$f(a+h) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z-(a+h)} dz.$$

Writing

$$1/(z-(a+h)) = 1/((z-a)-h) = (z-a)^{-1} \frac{1}{1-h(z-a)^{-1}},$$

we have

$$1/(z-(a+h)) = (z-a)^{-1}[1 + h(z-a)^{-1} + h^2(z-a)^{-2} + ...].$$

The series converges because $z$ lies on $C$ and $a + h$ is inside, so that the absolute value of the ratio $h/(z-a)$ is less than one. Equation (20.16) follows now from Equation (20.15).

## 20.9 Laurent Series: An Example

Suppose now that $C_1$ and $C_2$ are concentric circles with common center $a$, and the radius of $C_2$ is the smaller. We assume that the function $f(z)$ is differentiable in a region containing the annulus bounded by $C_1$ and $C_2$, but perhaps not differentiable at some points inside $C_2$. We want an infinite series expansion for $f(z)$ that is valid within the annulus. We begin with an example.

### 20.9.1 Expansion Within an Annulus

Let $f(z) = (7z-2)/(z+1)z(z-2)$. Then $f(z)$ is not differentiable at $z = -1$, $z = 0$, and $z = 2$. Suppose that we want a series expansion for $f(z)$ in terms of powers of $z + 1$, valid within the annulus $1 < |z+1| < 3$.

To simplify the calculations we replace $z$ with $t = z + 1$, so that

$$f(t) = (7t - 9)/t(t - 1)(t - 3),$$

and seek a series representation of $f(t)$ in terms of powers of $t$.

Using partial fractions, we obtain

$$f(t) = -3t^{-1} + (t - 1)^{-1} + 2(t - 3)^{-1}.$$

For $(t - 1)^{-1}$ we have

$$(t - 1)^{-1} = 1/(t - 1) = t^{-1}(1/(1 - t^{-1})) = t^{-1}[1 + t^{-1} + t^{-2} + ...],$$

which converges for $|t| > 1$. For $(t - 3)^{-1}$ we have

$$(t - 3)^{-1} = \frac{-1}{3}(1/(1 - t/3)) = \frac{-1}{3}[1 + \frac{t}{3} + (\frac{t}{3})^2 + ...],$$

which converges for $|t| < 3$. Putting all this together, we get

$$f(z) = -3(z+1)^{-1}+(z+1)^{-1}[1+(z+1)^{-1}+(z+1)^{-2}+...]-\frac{2}{3}[1+\frac{1}{3}(z+1)+\frac{1}{9}(z+1)^2+...].$$

### 20.9.2 Expansion Within the Inner Circle

Suppose now that we want a Laurent series expansion of $f(z)$ in powers of $z + 1$ that is convergent within the circle centered at $z = -1$, with radius one. The function $f(z)(z + 1)$ is differentiable within that circle, and so has a Taylor series expansion there. To get the Laurent series expansion for $f(z)$ we simply move the factor $z + 1$ to the other side, multiplying it by the Taylor expansion.

In this case, we write $(t - 1)^{-1}$ as

$$(t - 1)^{-1} = \frac{1}{t - 1} = \frac{-1}{1 - t} = -[1 + t + t^2 + t^3 + ...],$$

which converges for $|t| < 1$. The series for $(t - 3)^{-1}$ remains the same as before.

## 20.10 Laurent Series Expansions

Let $C_1$ and $C_2$ be two concentric circles with common center $a$, with the radius of $C_2$ the smaller. Let $f(z)$ be differentiable in a region containing the annulus bounded by $C_1$ and $C_2$. Let $a + h$ be inside the annulus and $C$ any simple closed curve in the annulus that surrounds the inner circle $C_2$. Then

$$f(a + h) = \sum_{n=-\infty}^{\infty} a_n h^n, \tag{20.17}$$

for

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z-a)^{n+1}} dz. \tag{20.18}$$

Note that for non-negative $n$ the integral in Equation (20.18) need not be $f^{(n)}(a)$, since the function $f(z)$ need not be differentiable inside of $C_2$. This theorem is discussed in problem 13.82 of the text.

To prove this, we use the same approach as in problem 13.12 of the text. What we find is that, since the two curves form the boundary of the annulus, Cauchy's Integral Theorem becomes

$$f(a+h) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z)}{z-(a+h)} dz - \frac{1}{2\pi i} \oint_{C_2} \frac{f(z)}{z-(a+h)} dz. \tag{20.19}$$

To obtain the desired result we write the expression $\frac{1}{z-(a+h)}$ in two ways, depending on if the $z$ lies on $C_1$ or on $C_2$. For $z$ on $C_1$ we write

$$\frac{1}{z-(a+h)} = (z-a)^{-1}[1 + \frac{h}{z-a} + (\frac{h}{z-a})^2 + ...], \tag{20.20}$$

while for $z$ on $C_2$ we write

$$\frac{1}{z-(a+h)} = -h^{-1}[1 + \frac{z-a}{h} + (\frac{z-a}{h})^2 + ...]. \tag{20.21}$$

Then

$$\oint_{C_1} \frac{f(z)}{z-(a+h)} dz = \sum_{n=0}^{\infty} h^n \oint_{C_1} \frac{f(z)}{(z-a)^{n+1}} dz, \tag{20.22}$$

and

$$\oint_{C_2} \frac{f(z)}{z-(a+h)} dz = \sum_{n=-\infty}^{-1} h^n \oint_{C_2} \frac{f(z)}{(z-a)^{n+1}} dz. \tag{20.23}$$

Both integrals are equivalent to integrals over the curve $C$. The desired result follows by applying Equation (20.15).

## 20.11 Residues

Suppose now that we want to integrate $f(z)$ over the simple closed curve $C$ in the previous section. From Equation (20.18) and $n+1=0$ we see that

$$\oint_C f(z) dz = (2\pi i) a_{-1}. \tag{20.24}$$

Note that if $f(z)$ is also differentiable inside of $C_2$ then $a_{-1} = 0$ and the integral is also zero.

If $(z-a)^m f(z)$ is differentiable on and inside $C$ then the Laurent expansion becomes

$$f(z) = a_{-m}(z-a)^{-m} + a_{-m+1}(z-a)^{-m+1} + ... + \sum_{n=0}^{\infty} a_n(z-a)^n \quad (20.25)$$

If $a_{-m}$ is not zero, then $f(z)$ is said to have a *pole of order* $m$ at $z = a$. We then have

$$\oint f(z)dz = (2\pi i)a_{-1}; \quad (20.26)$$

the number $a_{-1}$ is called the *residue* of $f(z)$ at the point $z = a$. Furthermore, we have

$$a_{-1} = \lim_{z\to a} \frac{1}{(m-1)!}\frac{d^{m-1}}{dz^{m-1}}\Big((z-a)^m f(z)\Big). \quad (20.27)$$

Note that we can replace the curve $C$ with an arbitrarily small circle centered at $z = a$.

If $f(z)$ is differentiable on and inside a simple closed curve $C$, except for a finite number of poles, then $\frac{1}{2\pi i}\oint_C f(z)dz$ is the sum of the residues at these poles; this is the Residue Theorem (see problem 13.25 of the text).

For example, consider again the function $f(z) = \frac{7z-2}{(z+1)z(z-2)}$. For the annulus $1 < |z+1| < 3$ and the curve $C$ the circle of radius two centered at $z = -1$, we have

$$\oint_C f(z)dz = -4\pi i,$$

since the residues of $f(z)$ are $-3$ at the pole $z = -1$ and 1 at the pole $z = 0$, both inside $C$.

For a second example, consider the function $f(z) = \frac{z^2}{(z^2+1)(z-2)}$. The points $z = 2$, $z = i$ and $z = -i$ are poles of order one. The residue of $f(z)$ at $z = 2$ is

$$\lim_{z\to 2}(z-2)f(z) = \frac{4}{5}.$$

## 20.12 The Binomial Theorem

Let $f(z) = \frac{1}{z(z+2)^3}$. Suppose that we want to represent $f(z)$ as a Laurent series involving powers of $z$. We write

$$f(z) = \frac{1}{8z}(1 + \frac{z}{2})^{-3}.$$

We need to expand $(1+\frac{z}{2})^{-3}$ as a Taylor series involving powers of $z$.

The *binomial theorem* tells us that, for any positive integer $N$,

$$(1+x)^N = \sum_{n=0}^{N} \binom{N}{n} x^n, \tag{20.28}$$

for

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}. \tag{20.29}$$

Now if $\alpha$ is any real number, we would like to have

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n; \tag{20.30}$$

The function

$$f(z) = (1+z)^\alpha \tag{20.31}$$

is analytic in the region $|z| < 1$, and so has a Taylor-series expansion of the form

$$(1+z)^\alpha = a_0 + a_1 z + a_2 z^2 + ..., \tag{20.32}$$

where

$$a_n = f^{(n)}(0)/n! = \alpha(\alpha-1)(\alpha-2)\cdots(\alpha-(n-1))/n!. \tag{20.33}$$

This tells us how to define $\binom{\alpha}{n}$. We can also see how to do it when we write

$$\binom{N}{n} = \frac{N(N-1)(N-2)(N-(n-1))}{n!}; \tag{20.34}$$

we now write

$$\binom{\alpha}{n} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-(n-1))}{n!}. \tag{20.35}$$

Using this extended binomial theorem we have

$$(1+\frac{z}{2})^{-3} = 1 - \frac{3}{2}z + \frac{3}{2}z^2 - \frac{5}{4}z^3 + .... \tag{20.36}$$

Therefore, we have

$$f(z) = \frac{1}{z(z+2)^3} = \frac{1}{8z} - \frac{3}{16} + \frac{3}{16}z - \frac{5}{32}z^2 + ... \tag{20.37}$$

The residue of $f(z)$ at the point $z=0$ is then $\frac{1}{8}$. Since $(z+2)^3 f(z) = z^{-1}$ and the second derivative is $2z^{-3}$, the residue of $f(z)$ at the point $z=-2$ is $\frac{-1}{8}$.

## 20.13 Using Residues

Suppose now that we want to find $\int_0^{2\pi} \frac{1}{5+3\sin\theta} d\theta$. Let $z = e^{i\theta}$. Then

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{z - z^{-1}}{2i},$$

and

$$dz = ie^{i\theta} d\theta = iz d\theta.$$

The integral is then

$$\int_0^{2\pi} \frac{1}{5+3\sin\theta} d\theta = \oint_C \frac{2}{3z^2 + 10iz - 3} dz, \tag{20.38}$$

where $C$ is the circle of radius one, centered at the origin. The poles of the integrand are $z = -3i$ and $z = -\frac{i}{3}$, both of order one. Only the pole $z = -\frac{i}{3}$ lies within $C$. The residue of the integrand at the pole $z = -\frac{i}{3}$ is $\frac{1}{4i}$, so the integral has the value $\frac{\pi}{2}$.

We conclude this chapter with a quick look at several of the more important consequences of the theory developed so far.

## 20.14 Cauchy's Estimate

Again, let $f(z)$ be analytic in a region $\mathcal{R}$, let $z_0$ be in $\mathcal{R}$, and let $C$ within $\mathcal{R}$ be the circle with center $z_0$ and radius $r$. Let $M$ be the maximum value of $|f(z)|$ for $z$ on $C$. Using Equation (20.15), it is easy to show that

$$|f^{(n)}(z_0)| \leq \frac{n!M}{r^n}. \tag{20.39}$$

Note that $M$ depends on $r$ and probably changes as $r$ changes. As we shall see now, this relatively simple calculation has important consequences.

## 20.15 Liouville's Theorem

Suppose now that $f(z)$ is analytic throughout the complex plane; that is, $f(z)$ is an *entire* function. Suppose also that $f(z)$ is *bounded*, that is, there is a constant $B > 0$ such that $|f(z)| \leq B$, for all $z$. Applying Equation (20.39) for the case of $n = 1$, we get

$$|f'(z_0)| \leq \frac{B}{r},$$

where the $B$ now does not change as $r$ changes. Then we write

$$|f'(z_0)|r \leq B.$$

Unless $f'(z_0) = 0$, the left side goes to infinity, as $r \to \infty$, while the right side stays constant. Therefore, $f'(z_0) = 0$ for all $z_0$ and $f(z)$ is constant. This is Liouville's Theorem.

## 20.16 The Fundamental Theorem of Algebra

The fundamental theorem of algebra tells us that every polynomial of degree greater than zero has a (possibly non-real) root. We can prove this using Liouville's Theorem.

Suppose $P(z)$ is such a polynomial and $P(z)$ has no roots. Then $P(z)^{-1}$ is analytic everywhere in the complex plane. Since $|P(z)| \to \infty$ as $|z| \to \infty$, it follows that $P(z)^{-1}$ is a bounded function. By Liouville's Theorem, $P(z)^{-1}$ must be constant; but this is not true. So $P(z)$ must be zero for some $z$.

## 20.17 Morera's Theorem

We know that if $f(z)$ is analytic in a region $\mathcal{R}$, then $\oint_C f(z)dz = 0$ for every simple closed curve $C$ in $\mathcal{R}$. Morera's Theorem is the converse.

**Theorem 20.1** *Let $f(z)$ be continuous in a region $\mathcal{R}$ and such that $\oint_C f(z)dz = 0$ for every simple closed curve $C$ in $\mathcal{R}$. Then $f(z)$ is analytic in $\mathcal{R}$.*

**Proof:** Let $f(z) = u(x, y) + iv(x, y)$. For arbitrary fixed $z_0 = (x_0, y_0)$ and variable $z = (x, y)$ in $\mathcal{R}$, define

$$F(z) = \int_{z_0}^{z} f(w)dw.$$

Then, writing

$$F(z) = U(x, y) + iV(x, y),$$

we can easily show that

$$U(x, y) = \int_{(x_0,y_0)}^{(x,y)} udx - vdy,$$

and

$$V(x, y) = \int_{(x_0,y_0)}^{(x,y)} vdx + udy.$$

It follows then, from our previous discussions of the Green's identities, that $U_x = u$, $U_y = -v$, $V_x = v$ and $V_y = u$. Therefore, $U_x = V_y$ and $U_y = -V_x$; that is, the Cauchy-Riemann Equations are satisfied. Therefore, since these partial derivatives are continuous, we can conclude that $F(z)$ is analytic in $\mathcal{R}$. But then, so is $F'(z) = f(z)$. ∎

# Chapter 21

# The Quest for Invisibility (Chapter 5,6)

## 21.1 Invisibility: Fact and Fiction

The military uses special materials and clever design in its stealth technology to build aircraft that are nearly invisible to radar. Fictional characters have it much easier; J.K. Rowling's hero Harry Potter becomes invisible when he wraps himself in his special cloak, and the Romulans in *Star Trek* can make their entire fleet of ships invisible by selectively bending light rays. In his *Republic* Plato, another best-selling author, has his hero Socrates and his friend Glaucon discuss ethical behavior. Socrates asserts that being honest and just is a good thing in its own right. Glaucon counters by recalling the mythical shepherd Gyges, who found a magic ring that he could use to make himself invisible. Glaucon wonders if anyone would behave honestly if no one would know if you did, and there was no possibility of punishment if you did not.

As Kurt Bryan and Tanya Leise discuss in the article [7], recent research in impedance tomography suggests that it may be possible, through the use of special meta-materials with carefully designed microstructure, to render certain objects invisible to certain electromagnetic probing. This chapter is a brief sketch of the theory; for more detail, see [7].

## 21.2 The Electro-Static Theory

Suppose that $\Omega$ is the open disk with center at the origin and radius one in two-dimensional space, and $\partial\Omega$ is its boundary, the circle with radius one centered at the origin. The points of $\partial\Omega$ are denoted $(\cos\theta, \sin\theta)$ for $0 \leq \theta < 2\pi$.

Let $f(\theta)$ describe a time-independent distribution of electrical charge along the boundary. Then $f(\theta)$ induces an electro-static field $\mathbf{E}(x, y)$ within $\Omega$. We know that there is an electro-static potential function $u(x, y)$ such that $\mathbf{E}(x, y) = -\nabla u(x, y)$.

If $f$ is constant, then so are $u$ and $\mathbf{E}$. If the disk $\Omega$ is made of a perfectly homogeneous conducting material, then current will flow within $\Omega$; the current vector at $(x, y)$ is denoted $\mathbf{J}(x, y)$ and $\mathbf{J}(x, y) = \gamma\mathbf{E}(x, y)$, where $\gamma > 0$ is the constant conductance. The component of the current field normal to the boundary at any point is

$$\frac{\partial u}{\partial \mathbf{n}}(\theta) = \nabla \cdot \mathbf{n}(x, y) = -\mathbf{J} \cdot \mathbf{n},$$

where $\mathbf{n} = \mathbf{n}(\theta)$ is the unit outward normal at $\theta$. This outward component of the current will also be constant over all $\theta$.

If $f$ is not constant, then the induced potential $u(x, y)$ will vary with $(x, y)$, as will the field $\mathbf{E}(x, y)$. Finding the induced potential $u(x, y)$ from $f(\theta)$ is called the *Dirichlet problem.*

If the conductance is not constant within $\Omega$, then each point $(x, y)$ will have a direction of maximum conductance and an orthgonal direction of minimum conductance. Using these as the eigenvectors of a positive-definite matrix $S = S(x, y)$, we have

$$\nabla \cdot (S\nabla u) = 0,$$

and $\mathbf{J} = S\nabla u$.

## 21.3 Impedance Tomography

In impedance tomography we attempt to determine the potential $u(x, y)$ within $\Omega$ by first applying a current at the points of the boundary, and then measuring the outward flux of the induced electro-static field at points of the boundary. The measured outward flow is called the *Neumann data.*

When the conductivity within $\Omega$ is changed, the relationship between the applied current and the measured outward flux changes. This suggests that when there is a non-conducting region $D$ within a homogeneous $\Omega$, we can detect it by noting the change in the measured outward flux.

## 21.4 Cloaking

Suppose we want to hide a conducting object within a non-conducting region $D$. We can do this, but it will still be possible to "see" the presence of $D$ and determine its size. If $D$ is large enough to conceal an object of a certain size, then one might become suspicious. What we need to do is

to make it look like the region $D$ is smaller than it really is, or is not even there.

By solving Laplace's equation for the region between the outer boundary, where we have measured the flux, and the inner boundary of $D$, where the flux is zero, we can see how the size of $D$ is reflected in the solution obtained. The presence of $D$ distorts the potential function, and therefore the measured flux. The key to invisibility is to modify the conductivity in the region surrounding $D$ in such a way that all (or, at least, most) of the distortion takes place well inside the boundary, so that at the boundary the potential looks undistorted.

For more mathematical details and discussion of the meta-materials needed to achieve this, see [7].

# Chapter 22

# Calculus of Variations (Chapter 16)

## 22.1 Introduction

In optimization, we are usually concerned with maximizing or minimizing real-valued functions of one or several variables, possibly subject to constraints. In this chapter, we consider another type of optimization problem, maximizing or minimizing *a function of functions*. The functions themselves we shall denote by simply $y = y(x)$, instead of the more common notation $y = f(x)$, and the function of functions will be denoted $J(y)$; in the calculus of variations, such functions of functions are called *functionals*. We then want to optimize $J(y)$ over a class of *admissible* functions $y(x)$. We shall focus on the case in which $x$ is a single real variable, although there are situations in which the functions $y$ are functions of several variables.

When we attempt to minimize a function $g(x_1, ..., x_N)$, we consider what happens to $g$ when we perturb the values $x_n$ to $x_n + \Delta x_n$. In order for $\mathbf{x} = (x_1, ..., x_N)$ to minimize $g$, it is necessary that

$$g(x_1 + \Delta x_1, ..., x_N + \Delta x_N) \geq g(x_1, ..., x_N),$$

for all perturbations $\Delta x_1, ..., \Delta x_N$. For differentiable $g$, this means that the gradient of $g$ at $\mathbf{x}$ must be zero. In the calculus of variations, when we attempt to minimize $J(y)$, we need to consider what happens when we perturb the function $y$ to a nearby *admissible* function, denoted $y + \Delta y$. In order for $y$ to minimize $J(y)$, we need

$$J(y + \Delta y) \geq J(y),$$

for all $\Delta y$ that make $y + \Delta y$ admissible. We end up with something analogous to a first derivative of $J$, which is then set to zero. The result is a

differential equation, called the *Euler-Lagrange Equation*, which must be satisfied by the minimizing $y$.

## 22.2 Some Examples

In this section we present some of the more famous examples of problems from the calculus of variations.

### 22.2.1 The Shortest Distance

Among all the functions $y = y(x)$, defined for $x$ in the interval $[0, 1]$, with $y(0) = 0$ and $y(1) = 1$, the straight-line function $y(x) = x$ has the shortest length. Assuming the functions are differentiable, the formula for the length of such curves is

$$J(y) = \int_0^1 \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx. \tag{22.1}$$

Therefore, we can say that the function $y(x) = x$ minimizes $J(y)$, over all such functions.

In this example, the functional $J(y)$ involves only the first derivative of $y = y(x)$ and has the form

$$J(y) = \int f(x, y(x), y'(x))dx, \tag{22.2}$$

where $f = f(u, v, w)$ is the function of three variables

$$f(u, v, w) = \sqrt{1 + w^2}. \tag{22.3}$$

In general, the functional $J(y)$ can come from almost any function $f(u, v, w)$. In fact, if higher derivatives of $y(x)$ are involved, the function $f$ can be a function of more than three variables. In this chapter we shall confine our discussion to problems involving only the first derivative of $y(x)$.

### 22.2.2 The Brachistochrone Problem

Consider a frictionless wire connecting the two points $A = (0, 0)$ and $B = (1, 1)$; for convenience, the positive $y$-axis is downward. A metal ball rolls from point $A$ to point $B$ under the influence of gravity. What shape should the wire take in order to make the travel time of the ball the smallest? This famous problem, known as the *Brachistochrone Problem*, was posed in 1696 by Johann Bernoulli. This event is viewed as marking the beginning of the calculus of variations.

The velocity of the ball along the curve is $v = \frac{ds}{dt}$, where $s$ denotes the arc-length. Therefore,

$$dt = \frac{ds}{v} = \frac{1}{v}\sqrt{1 + \left(\frac{dy}{dx}\right)^2}dx.$$

Because the ball is falling under the influence of gravity only, the velocity it attains after falling from $(0,0)$ to $(x,y)$ is the same as it would have attained had it fallen $y$ units vertically; only the travel times are different. This is because the loss of potential energy is the same either way. The velocity attained after a vertical free fall of $y$ units is $\sqrt{2gy}$. Therefore, we have

$$dt = \frac{\sqrt{1 + \left(\frac{dy}{dx}\right)^2}dx}{\sqrt{2gy}}.$$

The travel time from $A$ to $B$ is therefore

$$J(y) = \frac{1}{\sqrt{2g}}\int_0^1 \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\frac{1}{\sqrt{y}}dx. \tag{22.4}$$

For this example, the function $f(u,v,w)$ is

$$f(u,v,w) = \frac{\sqrt{1+w^2}}{\sqrt{v}}. \tag{22.5}$$

### 22.2.3 Minimal Surface Area

Given a function $y = y(x)$ with $y(0) = 1$ and $y(1) = 0$, we imagine revolving this curve around the $x$-axis, to generate a surface of revolution. The functional $J(y)$ that we wish to minimize now is the surface area. Therefore, we have

$$J(y) = \int_0^1 y\sqrt{1 + y'(x)^2}dx. \tag{22.6}$$

Now the function $f(u,v,w)$ is

$$f(u,v,w) = v\sqrt{1+w^2}. \tag{22.7}$$

### 22.2.4 The Maximum Area

Among all curves of length $L$ connecting the points $(0,0)$ and $(1,0)$, find the one for which the area $A$ of the region bounded by the curve and the $x$-axis is maximized. The length of the curve is given by

$$L = \int_0^1 \sqrt{1 + y'(x)^2}dx, \tag{22.8}$$

and the area, assuming that $y(x) \geq 0$ for all $x$, is

$$A = \int_0^1 y(x)dx. \tag{22.9}$$

This problem is different from the previous ones, in that we seek to optimize a functional, subject to a second functional being held fixed. Such problems are called *problems with constraints*.

### 22.2.5 Maximizing Burg Entropy

The *Burg entropy* of a positive-valued function $y(x)$ on $[-\pi, \pi]$ is

$$BE(y) = \int_{-\pi}^{\pi} \log\Big(y(x)\Big)dx. \tag{22.10}$$

An important problem in signal processing is to maximize $BE(y)$, subject to

$$r_n = \int_{-\pi}^{\pi} y(x)e^{-inx}dx, \tag{22.11}$$

for $|n| \leq N$. The $r_n$ are values of the Fourier transform of the function $y(x)$.

## 22.3 Comments on Notation

The functionals $J(y)$ that we shall consider in this chapter have the form

$$J(y) = \int f(x, y(x), y'(x))dx, \tag{22.12}$$

where $f = f(u, v, w)$ is some function of three real variables. It is common practice, in the calculus of variations literature, to speak of $f = f(x, y, y')$, rather than $f(u, v, w)$. Unfortunately, this leads to potentially confusing notation, such as when $\frac{\partial f}{\partial u}$ is written as $\frac{\partial f}{\partial x}$, which is not the same thing as the total derivative of $f(x, y(x), y'(x))$,

$$\frac{d}{dx}f(x, y(x), y'(x)) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y'(x) + \frac{\partial f}{\partial y'}y''(x). \tag{22.13}$$

Using the notation of this chapter, Equation (22.13) becomes

$$\frac{d}{dx}f(x, y(x), y'(x)) = \frac{\partial f}{\partial u}(x, y(x), y'(x)) +$$

$$\frac{\partial f}{\partial v}(x, y(x), y'(x))y'(x) + \frac{\partial f}{\partial w}(x, y(x), y'(x))y''(x). \tag{22.14}$$

The common notation forces us to view $f(x, y, y')$ both as a function of three unrelated variables, $x$, $y$, and $y'$, and as $f(x, y(x), y'(x))$, a function of the single variable $x$.

For example, suppose that

$$f(u, v, w) = u^2 + v^3 + \sin w,$$

and

$$y(x) = 7x^2.$$

Then

$$f(x, y(x), y'(x)) = x^2 + (7x^2)^3 + \sin(14x), \tag{22.15}$$

$$\frac{\partial f}{\partial x}(x, y(x), y'(x)) = 2x, \tag{22.16}$$

and

$$\frac{d}{dx} f(x, y(x), y'(x)) = \frac{d}{dx}\Big(x^2 + (7x^2)^3 + \sin(14x)\Big)$$

$$= 2x + 3(7x^2)^2(14x) + 14\cos(14x). \tag{22.17}$$

## 22.4 The Euler-Lagrange Equation

In the problems we shall consider in this chapter, admissible functions are differentiable, with $y(x_1) = y_1$ and $y(x_2) = y_2$; that is, the graphs of the admissible functions pass through the end points $(x_1, y_1)$ and $(x_2, y_2)$. If $y = y(x)$ is one such function and $\eta(x)$ is a differentiable function with $\eta(x_1) = 0$ and $\eta(x_2) = 0$, then $y(x) + \epsilon\eta(x)$ is admissible, for all values of $\epsilon$. For fixed admissible function $y = y(x)$, we define

$$J(\epsilon) = J(y(x) + \epsilon\eta(x)), \tag{22.18}$$

and force $J'(\epsilon) = 0$ at $\epsilon = 0$. The tricky part is calculating $J'(\epsilon)$.

Since $J(y(x) + \epsilon\eta(x))$ has the form

$$J(y(x) + \epsilon\eta(x)) = \int_{x_1}^{x_2} f(x, y(x) + \epsilon\eta(x), y'(x) + \epsilon\eta'(x))dx, \tag{22.19}$$

we obtain $J'(\epsilon)$ by differentiating under the integral sign.

Omitting the arguments, we have

$$J'(\epsilon) = \int_{x_1}^{x_2} \frac{\partial f}{\partial v}\eta + \frac{\partial f}{\partial w}\eta' dx. \tag{22.20}$$

Using integration by parts and $\eta(x_1) = \eta(x_2) = 0$, we have

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial w}\eta' dx = -\int_{x_1}^{x_2} \frac{d}{dx}(\frac{\partial f}{\partial w})\eta dx. \tag{22.21}$$

Therefore, we have

$$J'(\epsilon) = \int_{x_1}^{x_2} \Big(\frac{\partial f}{\partial v} - \frac{d}{dx}(\frac{\partial f}{\partial w})\Big)\eta dx. \tag{22.22}$$

In order for $y = y(x)$ to be the optimal function, this integral must be zero for every appropriate choice of $\eta(x)$, when $\epsilon = 0$. It can be shown without too much trouble that this forces

$$\frac{\partial f}{\partial v} - \frac{d}{dx}(\frac{\partial f}{\partial w}) = 0. \tag{22.23}$$

Equation (22.23) is the *Euler-Lagrange Equation*.

For clarity, let us rewrite that Euler-Lagrange Equation using the arguments of the functions involved. Equation (22.23) is then

$$\frac{\partial f}{\partial v}(x, y(x), y'(x)) - \frac{d}{dx}\Big(\frac{\partial f}{\partial w}(x, y(x), y'(x))\Big) = 0. \tag{22.24}$$

## 22.5 Special Cases of the Euler-Lagrange Equation

The Euler-Lagrange Equation simplifies in certain special cases. Here we consider two cases: 1) when $f(u, v, w)$ is independent of the variable $v$, as in Equation (22.3); and 2) when $f(u, v, w)$ is independent of the variable $u$, as in Equations (22.5) and (22.7).

### 22.5.1 If $f$ is independent of $v$

If the function $f(u, v, w)$ is independent of the variable $v$ then the Euler-Lagrange Equation (22.24) becomes

$$\frac{\partial f}{\partial w}(x, y(x), y'(x)) = c, \tag{22.25}$$

for some constant $c$. If, in addition, the function $f(u, v, w)$ is a function of $w$ alone, then so is $\frac{\partial f}{\partial w}$, from which we conclude from the Euler-Lagrange Equation that $y'(x)$ is constant.

### 22.5.2 If $f$ is independent of $u$

Note that we can write

$$\frac{d}{dx}f(x,y(x),y'(x)) = \frac{\partial f}{\partial u}(x,y(x),y'(x))$$

$$+\frac{\partial f}{\partial v}(x,y(x),y'(x))y'(x) + \frac{\partial f}{\partial w}(x,y(x),y'(x))y''(x). \tag{22.26}$$

We also have

$$\frac{d}{dx}\Big(y'(x)\frac{\partial f}{\partial w}(x,y(x),y'(x))\Big) =$$

$$y'(x)\frac{d}{dx}\Big(\frac{\partial f}{\partial w}(x,y(x),y'(x))\Big) + y''(x)\frac{\partial f}{\partial w}(x,y(x),y'(x)). \tag{22.27}$$

Subtracting Equation (22.27) from Equation (22.26), we get

$$\frac{d}{dx}\Big(f(x,y(x),y'(x)) - y'(x)\frac{\partial f}{\partial w}(x,y(x),y'(x))\Big) =$$

$$\frac{\partial f}{\partial u}(x,y(x),y'(x)) + y'(x)\Big(\frac{\partial f}{\partial v} - \frac{d}{dx}\frac{\partial f}{\partial w}\Big)(x,y(x),y'(x)). \tag{22.28}$$

Now, using the Euler-Lagrange Equation, we see that Equation (22.28) reduces to

$$\frac{d}{dx}\Big(f(x,y(x),y'(x)) - y'(x)\frac{\partial f}{\partial w}(x,y(x),y'(x))\Big) = \frac{\partial f}{\partial u}(x,y(x),y'(x)). \tag{22.29}$$

If it is the case that $\frac{\partial f}{\partial u} = 0$, then equation (22.29) leads to

$$f(x,y(x),y'(x)) - y'(x)\frac{\partial f}{\partial w}(x,y(x),y'(x)) = c, \tag{22.30}$$

for some constant $c$.

## 22.6 Using the Euler-Lagrange Equation

We derive and solve the Euler-Lagrange Equation for each of the examples presented previously.

### 22.6.1 The Shortest Distance

In this case, we have

$$f(u, v, w) = \sqrt{1 + w^2}, \tag{22.31}$$

so that

$$\frac{\partial f}{\partial v} = 0,$$

and

$$\frac{\partial f}{\partial u} = 0.$$

We conclude that $y'(x)$ is constant, so $y(x)$ is a straight line.

### 22.6.2 The Brachistochrone Problem

Equation (22.5) tells us that

$$f(u, v, w) = \frac{\sqrt{1 + w^2}}{\sqrt{v}}. \tag{22.32}$$

Then, since

$$\frac{\partial f}{\partial u} = 0,$$

and

$$\frac{\partial f}{\partial w} = \frac{w}{\sqrt{1 + w^2}\sqrt{v}},$$

Equation (22.30) tells us that

$$\frac{\sqrt{1 + y'(x)^2}}{\sqrt{y(x)}} - y'(x)\frac{y'(x)}{\sqrt{1 + y'(x)^2}\sqrt{y(x)}} = c. \tag{22.33}$$

Equivalently, we have

$$\sqrt{y(x)}\sqrt{1 + y'(x)^2} = \sqrt{a}. \tag{22.34}$$

Solving for $y'(x)$, we get

$$y'(x) = \sqrt{\frac{a - y(x)}{y(x)}}. \tag{22.35}$$

Separating variables and integrating, using the substitution

$$y = a\sin^2\theta = \frac{a}{2}(1 - \cos 2\theta),$$

we obtain

$$x = 2a \int \sin^2 \theta d\theta = \frac{a}{2}(2\theta - \sin 2\theta) + k. \tag{22.36}$$

From this, we learn that the minimizing curve is a *cycloid*, that is, the path a point on a circle traces as the circle rolls.

There is an interesting connection, discussed by Simmons in [42] , between the brachistochrone problem and the refraction of light rays. Imagine a ray of light passing from the point $A = (0, a)$, with $a > 0$, to the point $B = (c, b)$, with $c > 0$ and $b < 0$. Suppose that the speed of light is $v_1$ above the $x$-axis, and $v_2 < v_1$ below the $x$-axis. The path consists of two straight lines, meeting at the point $(0, x)$. The total time for the journey is then

$$T(x) = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{b^2 + (c - x)^2}}{v_2}.$$

Fermat's Principle of Least Time says that the (apparent) path taken by the light ray will be the one for which $x$ minimizes $T(x)$. From calculus, it follows that

$$\frac{x}{v_1\sqrt{a^2 + x^2}} = \frac{c - x}{v_2\sqrt{b^2 + (c - x)^2}},$$

and from geometry, we get *Snell's Law*:

$$\frac{\sin \alpha_1}{v_1} = \frac{\sin \alpha_2}{v_2},$$

where $\alpha_1$ and $\alpha_2$ denote the angles between the upper and lower parts of the path and the vertical, respectively.

Imagine now a stratified medium consisting of many horizontal layers, each with its own speed of light. The path taken by the light would be such that $\frac{\sin \alpha}{v}$ remains constant as the ray passes from one layer to the next. In the limit of infinitely many infinitely thin layers, the path taken by the light would satisfy the equation $\frac{\sin \alpha}{v} = \text{constant}$, with

$$\sin \alpha = \frac{1}{\sqrt{1 + y'(x)^2}}.$$

As we have already seen, the velocity attained by the rolling ball is $v = \sqrt{2gy}$, so the equation to be satisfied by the path $y(x)$ is

$$\sqrt{2gy(x)}\sqrt{1 + y'(x)^2} = \text{constant},$$

which is what we obtained from the Euler-Lagrange Equation.

### 22.6.3 Minimizing the Surface Area

For the problem of minimizing the surface area of a surface of revolution, the function $f(u, v, w)$ is

$$f(u, v, w) = v\sqrt{1 + w^2}. \tag{22.37}$$

Once again, $\frac{\partial f}{\partial u} = 0$, so we have

$$\frac{y(x)y'(x)^2}{\sqrt{1 + y'(x)^2}} - y(x)\sqrt{1 + y'(x)^2} = c. \tag{22.38}$$

It follows that

$$y(x) = b \cosh \frac{x - a}{b}, \tag{22.39}$$

for appropriate $a$ and $b$.

It is important to note that being a solution of the Euler-Lagrange Equation is a necessary condition for a differentiable function to be a solution to the original optimization problem, but it is not a sufficient condition. The optimal solution may not be a differentiable one, or there may be no optimal solution. In the case of minimum surface area, there may not be any function of the form in Equation (22.39) passing through the two given end points; see Chapter IV of Bliss [2] for details.

## 22.7 Problems with Constraints

We turn now to the problem of optimizing one functional, subject to a second functional being held constant. The basic technique is similar to ordinary optimization subject to constraints: we use Lagrange multipliers. We begin with a classic example.

### 22.7.1 The Isoperimetric Problem

A classic problem in the calculus of variations is the *Isoperimetric Problem*: find the curve of a fixed length that encloses the largest area. For concreteness, suppose the curve connects the two points $(0, 0)$ and $(1, 0)$ and is the graph of a function $y(x)$. The problem then is to maximize the area integral

$$\int_0^1 y(x)dx, \tag{22.40}$$

subject to the perimeter being held fixed, that is,

$$\int_0^1 \sqrt{1 + y'(x)^2}dx = P. \tag{22.41}$$

With

$$f(x, y(x), y'(x)) = y(x) + \lambda\sqrt{1 + y'(x)^2},$$

the Euler-Lagrange Equation becomes

$$\frac{d}{dx}\Big(\frac{\lambda y'(x)}{\sqrt{1 + y'(x)^2}}\Big) - 1 = 0, \tag{22.42}$$

or

$$\frac{y'(x)}{\sqrt{1 + y'(x)^2}} = \frac{x - a}{\lambda}. \tag{22.43}$$

Using the substitution $t = \frac{x-a}{\lambda}$ and integrating, we find that

$$(x - a)^2 + (y - b)^2 = \lambda^2, \tag{22.44}$$

which is the equation of a circle. So the optimal function $y(x)$ is a portion of a circle.

What happens if the assigned perimeter $P$ is greater than $\frac{\pi}{2}$, the length of the semicircle connecting $(0, 0)$ and $(1, 0)$? In this case, the desired curve is not the graph of a function of $x$, but a parameterized curve of the form $(x(t), y(t))$, for, say, $t$ in the interval $[0, 1]$. Now we have one independent variable, $t$, but two dependent ones, $x$ and $y$. We need a generalization of the Euler-Lagrange Equation to the multivariate case.

### 22.7.2 Burg Entropy

According to the Euler-Lagrange Equation for this case, we have

$$\frac{1}{y(x)} + \sum_{n=-N}^{N} \lambda_n e^{-ixn}, \tag{22.45}$$

or

$$y(x) = 1/\sum_{n=-N}^{N} a_n e^{inx}. \tag{22.46}$$

The *spectral factorization* theorem [37] tells us that if the denominator is positive for all $x$, then it can be written as

$$\sum_{n=-N}^{N} a_n e^{inx} = |\sum_{m=0}^{N} b_m e^{imx}|^2. \tag{22.47}$$

With a bit more work (see [10]), it can be shown that the desired coefficients $b_m$ are the solution to the system of equations

$$\sum_{m=0}^{N} r_{m-k} b_m = 0, \tag{22.48}$$

for $k = 1, 2, ..., N$ and

$$\sum_{m=0}^{N} r_m b_m = 1. \tag{22.49}$$

## 22.8 The Multivariate Case

Suppose that the integral to be optimized is

$$J(x, y) = \int_a^b f(t, x(t), x'(t), y(t), y'(t))dt, \tag{22.50}$$

where $f(u, v, w, s, r)$ is a real-valued function of five variables. In such cases, the Euler-Lagrange Equation is replaced by the two equations

$$\begin{aligned} \frac{d}{dt}\Big(\frac{\partial f}{\partial w}\Big) - \frac{\partial f}{\partial v} &= 0, \\ \frac{d}{dt}\Big(\frac{\partial f}{\partial r}\Big) - \frac{\partial f}{\partial s} &= 0. \end{aligned} \tag{22.51}$$

We apply this now to the problem of maximum area for a fixed perimeter.

We know from Green's Theorem in two dimensions that the area $A$ enclosed by a curve $C$ is given by the integral

$$A = \frac{1}{2}\oint_C (xdy - ydx) = \frac{1}{2}\int_0^1 (x(t)y'(t) - y(t)x'(t))dt. \tag{22.52}$$

The perimeter $P$ of the curve is

$$P = \int_0^1 \sqrt{x'(t)^2 + y'(t)^2}dt. \tag{22.53}$$

So the problem is to maximize the integral in Equation (22.52), subject to the integral in Equation (22.53) being held constant.

The problem is solved by using a Lagrange multiplier. We write

$$J(x, y) = \int_0^1 \Big(x(t)y'(t) - y(t)x'(t) + \lambda\sqrt{x'(t)^2 + y'(t)^2}\Big)dt. \tag{22.54}$$

The generalized Euler-Lagrange Equations are

$$\frac{d}{dt}\Big(\frac{1}{2}x(t) + \frac{\lambda y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\Big) + \frac{1}{2}x'(t) = 0, \tag{22.55}$$

and

$$\frac{d}{dt}\Big(-\frac{1}{2}y(t) + \frac{\lambda x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\Big) - \frac{1}{2}y'(t) = 0. \tag{22.56}$$

It follows that

$$y(t) + \frac{\lambda x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}} = c, \tag{22.57}$$

and

$$x(t) + \frac{\lambda y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}} = d. \tag{22.58}$$

Therefore,

$$(x-d)^2 + (y-c)^2 = \lambda^2. \tag{22.59}$$

The optimal curve is then a portion of a circle.

## 22.9 Finite Constraints

Let $x$, $y$ and $z$ be functions of the independent variable $t$, with $\dot{x} = x'(t)$. Suppose that we want to minimize the functional

$$J(x,y,z) = \int_a^b f(x,\dot{x},y,\dot{y},z,\dot{z})dt,$$

subject to the constraint

$$G(x,y,z) = 0.$$

Here we suppose that the points $(x(t), y(t), z(t))$ describe a curve in space and that the condition $G(x(t), y(t), z(t)) = 0$ restricts the curve to the surface $G(x, y, z) = 0$. Such a problem is said to be one of *finite constraints*. In this section we illustrate this type of problem by considering the geodesic problem.

### 22.9.1 The Geodesic Problem

The space curve $(x(t), y(t), z(t))$, defined for $a \leq t \leq b$, lies on the surface described by $G(x, y, z) = 0$ if $G(x(t), y(t), z(t)) = 0$ for all $t$ in $[a, b]$. The *geodesic problem* is to find the curve of shortest length lying on the surface and connecting points $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$. The functional to be minimized is the arc length

$$J = \int_a^b \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}dt, \tag{22.60}$$

where $\dot{x} = \frac{dx}{dt}$. Here the function $f$ is

$$f(x,\dot{x},y,\dot{y},z,\dot{z}) = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}.$$

We assume that the equation $G(x, y, z) = 0$ can be rewritten as

$$z = g(x, y),$$

that is, we assume that we can solve for the variable $z$, and that the function $g$ has continuous second partial derivatives. We may not be able to do this for the entire surface, as the equation of a sphere $G(x, y, z) = x^2 + y^2 + z^2 - r^2 = 0$ illustrates, but we can usually solve for $z$, or one of the other variables, on part of the surface, as, for example, on the upper or lower hemisphere.

We then have

$$\dot{z} = g_x\dot{x} + g_y\dot{y} = g_x(x(t), y(t))\dot{x}(t) + g_y(x(t), y(t))\dot{y}(t), \tag{22.61}$$

where $g_x = \frac{\partial g}{\partial x}$.

Substituting for $z$ in Equation (22.60), we see that the problem is now to minimize the functional

$$J = \int_a^b \sqrt{\dot{x}^2 + \dot{y}^2 + (g_x\dot{x} + g_y\dot{y})^2}dt, \tag{22.62}$$

which we write as

$$J = \int_a^b F(x, \dot{x}, y, \dot{y})dt. \tag{22.63}$$

The Euler-Lagrange Equations are then

$$\frac{\partial F}{\partial x} - \frac{d}{dt}(\frac{\partial F}{\partial \dot{x}}) = 0, \tag{22.64}$$

and

$$\frac{\partial F}{\partial y} - \frac{d}{dt}(\frac{\partial F}{\partial \dot{y}}) = 0. \tag{22.65}$$

We want to rewrite the Euler-Lagrange equations.

**Lemma 22.1** *We have*

$$\frac{\partial \dot{z}}{\partial x} = \frac{d}{dt}(g_x).$$

**Proof:** From Equation (22.61) we have

$$\frac{\partial \dot{z}}{\partial x} = \frac{\partial}{\partial x}(g_x\dot{x} + g_y\dot{y}) = g_{xx}\dot{x} + g_{yx}\dot{y}.$$

We also have

$$\frac{d}{dt}(g_x) = \frac{d}{dt}(g_x(x(t), y(t)) = g_{xx}\dot{x} + g_{xy}\dot{y}.$$

Since $g_{xy} = g_{yx}$, the assertion of the lemma follows. ∎

From the Lemma we have both

$$\frac{\partial \dot{z}}{\partial x} = \frac{d}{dt}(g_x), \tag{22.66}$$

and

$$\frac{\partial \dot{z}}{\partial y} = \frac{d}{dt}(g_y). \tag{22.67}$$

Using

$$\begin{aligned}\frac{\partial F}{\partial x} &= \frac{\partial f}{\partial \dot{z}}\frac{\partial(g_x\dot{x} + g_y\dot{y})}{\partial x} \\ &= \frac{\partial f}{\partial \dot{z}}\frac{\partial}{\partial x}(\frac{dg}{dt}) = \frac{\partial f}{\partial \dot{z}}\frac{\partial \dot{z}}{\partial x}\end{aligned}$$

and

$$\frac{\partial F}{\partial y} = \frac{\partial f}{\partial \dot{z}}\frac{\partial \dot{z}}{\partial y},$$

we can rewrite the Euler-Lagrange Equations as

$$\frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + g_x\frac{d}{dt}(\frac{\partial f}{\partial \dot{z}}) = 0, \tag{22.68}$$

and

$$\frac{d}{dt}(\frac{\partial f}{\partial \dot{y}}) + g_y\frac{d}{dt}(\frac{\partial f}{\partial \dot{z}}) = 0. \tag{22.69}$$

To see why this is the case, we reason as follows. First

$$\begin{aligned}\frac{\partial F}{\partial \dot{x}} &= \frac{\partial f}{\partial \dot{x}} + \frac{\partial f}{\partial \dot{z}}\frac{\partial \dot{z}}{\partial \dot{x}} \\ &= \frac{\partial f}{\partial \dot{x}} + \frac{\partial f}{\partial \dot{z}}g_x,\end{aligned}$$

so that

$$\begin{aligned}\frac{d}{dt}(\frac{\partial F}{\partial \dot{x}}) &= \frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + \frac{d}{dt}(\frac{\partial f}{\partial \dot{z}}g_x) \\ &= \frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + \frac{d}{dt}(\frac{\partial f}{\partial \dot{z}})g_x + \frac{\partial f}{\partial \dot{z}}\frac{d}{dt}(g_x) \\ &= \frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + \frac{d}{dt}(\frac{\partial f}{\partial \dot{z}})g_x + \frac{\partial f}{\partial \dot{z}}\frac{\partial \dot{z}}{\partial x}.\end{aligned}$$

Therefore,

$$\frac{d}{dt}(\frac{\partial F}{\partial \dot{x}}) = \frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + \frac{d}{dt}(\frac{\partial f}{\partial \dot{z}})g_x + \frac{\partial F}{\partial x},$$

so that

$$0 = \frac{d}{dt}(\frac{\partial F}{\partial \dot{x}}) - \frac{\partial F}{\partial x} = \frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) + \frac{d}{dt}(\frac{\partial f}{\partial \dot{z}})g_x. \tag{22.70}$$

Let the function $\lambda(t)$ be defined by

$$\frac{d}{dt}(\frac{\partial f}{\partial \dot{z}}) = \lambda(t)G_z.$$

From $G(x, y, z) = 0$ and $z = g(x, y)$, we have

$$H(x, y) = G(x, y, g(x, y)) = 0.$$

Then we have

$$H_x = G_x + G_z g_x = 0,$$

so that

$$g_x = -\frac{G_x}{G_z};$$

similarly, we have

$$g_y = -\frac{G_y}{G_z}.$$

Then the Euler-Lagrange Equations become

$$\frac{d}{dt}(\frac{\partial f}{\partial \dot{x}}) = \lambda(t)G_x, \tag{22.71}$$

and

$$\frac{d}{dt}(\frac{\partial f}{\partial \dot{y}}) = \lambda(t)G_y. \tag{22.72}$$

Eliminating $\lambda(t)$ and extending the result to include $z$ as well, we have

$$\frac{\frac{d}{dt}(\frac{\partial f}{\partial \dot{x}})}{G_x} = \frac{\frac{d}{dt}(\frac{\partial f}{\partial \dot{y}})}{G_y} = \frac{\frac{d}{dt}(\frac{\partial f}{\partial \dot{z}})}{G_z}. \tag{22.73}$$

Notice that we could obtain the same result by calculating the Euler-Lagrange Equation for the functional

$$\int_a^b f(\dot{x}, \dot{y}, \dot{z}) + \lambda(t)G(x(t), y(t), z(t))dt. \tag{22.74}$$

### 22.9.2 An Example

Let the surface be a sphere, with equation

$$0 = G(x, y, z) = x^2 + y^2 + z^2 - r^2.$$

Then Equation (22.73) becomes

$$\frac{f\ddot{x} - \dot{x}\dot{f}}{2xf^2} = \frac{f\ddot{y} - \dot{y}\dot{f}}{2yf^2} = \frac{f\ddot{z} - \dot{z}\dot{f}}{2zf^2}.$$

We can rewrite these equations as

$$\frac{\ddot{x}y - x\ddot{y}}{\dot{x}y - x\dot{y}} = \frac{y\ddot{z} - z\ddot{y}}{y\dot{z} - z\dot{y}} = \frac{\dot{f}}{f}.$$

The numerators are the derivatives, with respect to $t$, of the denominators, which leads to

$$\log|x\dot{y} - y\dot{x}| = \log|y\dot{z} - z\dot{y}| + c_1.$$

Therefore,

$$x\dot{y} - y\dot{x} = c_1(y\dot{z} - z\dot{y}).$$

Rewriting, we obtain

$$\frac{\dot{x} + c_1\dot{z}}{x + c_1 z} = \frac{\dot{y}}{y},$$

or

$$x + c_1 z = c_2 y,$$

which is a plane through the origin. The geodesics on the sphere are great circles, that is, the intersection of the sphere with a plane through the origin.

## 22.10 Hamilton's Principle and the Lagrangian

### 22.10.1 Generalized Coordinates

Suppose there are $J$ particles at positions $r_j(t) = (x_j(t), y_j(t), z_j(t))$, with masses $m_j$, for $j = 1, 2, ..., J$. Assume that there is a potential function $V(x_1, y_1, z_1, ..., x_J, y_J, z_J)$ such that the force acting on the $j$th particle is

$$F_j = -(\frac{\partial V}{\partial x_j}, \frac{\partial V}{\partial y_j}, \frac{\partial V}{\partial z_j}).$$

The kinetic energy is then

$$T = \frac{1}{2}\sum_{j=1}^{J} m_j\Big((\dot{x}_j)^2 + (\dot{y}_j)^2 + (\dot{z}_j)^2\Big).$$

Suppose also that the positions of the particles are constrained by the conditions

$$\phi_i(x_1, y_1, z_1, ..., x_J, y_J, z_J) = 0,$$

for $i = 1, ..., I$. Then there are $N = 3J - I$ *generalized coordinates* $q_1, ..., q_N$ describing the behavior of the particles.

For example, suppose that there is one particle moving on the surface of a sphere with radius $R$. Then the constraint is that

$$x^2 + y^2 + z^2 = R^2.$$

The generalized coordinates can be chosen to be the two angles describing position on the surface, or latitude and longitude, say.

We then have

$$\dot{x}_j = \sum_{n=1}^{N} \frac{\partial x_j}{\partial q_n} \dot{q}_n,$$

with similar expressions for the other time derivatives.

### 22.10.2 Homogeneity and Euler's Theorem

A function $f(u, v, w)$ is said to be *n-homogeneous* if

$$f(tu, tv, tw) = t^n f(u, v, w),$$

for any scalar $t$. The kinetic energy $T$ is 2-homogeneous in the variables $\dot{q}_n$.

**Lemma 22.2** *Let $f(u, v, w)$ be n-homogeneous. Then*

$$\frac{\partial f}{\partial u}(au, av, aw) = a^{n-1} \frac{\partial f}{\partial u}(u, v, w). \tag{22.75}$$

**Proof:** We write

$$\frac{\partial f}{\partial u}(au, av, aw) = \lim_{\Delta \to 0} \frac{f(au + a\Delta, av, aw) - f(au, av, aw)}{a\Delta}$$

$$= \frac{a^n}{a} \frac{\partial f}{\partial u}(u, v, w) = a^{n-1} \frac{\partial f}{\partial u}(u, v, w).$$

∎

**Theorem 22.1 (Euler's Theorem)** *Let $f(u, v, w)$ be n-homogeneous. Then*

$$u \frac{\partial f}{\partial u}(u, v, w) + v \frac{\partial f}{\partial v}(u, v, w) + w \frac{\partial f}{\partial w}(u, v, w) = n f(u, v, w). \tag{22.76}$$

**Proof:** Define $g(a) = f(au, av, aw)$, so that

$$g'(a) = u\frac{\partial f}{\partial u}(au, av, aw) + v\frac{\partial f}{\partial v}(au, av, aw) + w\frac{\partial f}{\partial w}(au, av, aw).$$

Using Equation (22.75) we have

$$g'(a) = a^{n-1}\Big(u\frac{\partial f}{\partial u}(u, v, w) + v\frac{\partial f}{\partial v}(u, v, w) + w\frac{\partial f}{\partial w}(u, v, w)\Big).$$

But we also know that

$$g(a) = a^n f(u, v, w),$$

so that

$$g'(a) = na^{n-1} f(u, v, w).$$

It follows that

$$u\frac{\partial f}{\partial u}(u, v, w) + v\frac{\partial f}{\partial v}(u, v, w) + w\frac{\partial f}{\partial w}(u, v, w) = nf(u, v, w).$$

∎

Since the kinetic energy $T$ is 2-homogeneous in the variables $\dot{q}_n$, it follows that

$$2T = \sum_{n=1}^{N} \dot{q}_n \frac{\partial T}{\partial \dot{q}_n}. \tag{22.77}$$

### 22.10.3 Hamilton's Principle

The *Lagrangian* is defined to be

$$L(q_1, ..., q_N, \dot{q}_1, ..., \dot{q}_N) = T - V.$$

Hamilton's principle is then that the paths taken by the particles are such that the integral

$$\int_{t_1}^{t_2} L(t)dt = \int_{t_1}^{t_2} T(t) - V(t)dt$$

is minimized. Consequently, the paths must satisfy the Euler-Lagrange equations

$$\frac{\partial L}{\partial q_n} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_n} = 0,$$

for each $n$. Since the variable $t$ does not appear explicitly, we know that

$$\sum_{n=1}^{N} \dot{q}_n \frac{\partial L}{\partial \dot{q}_n} - L = E,$$

for some constant $E$.

Noting that

$$\frac{\partial L}{\dot{q}_n} = \frac{\partial T}{\dot{q}_n},$$

since $V$ does not depend on the variables $\dot{q}_n$, and using Equation (22.77), we find that

$$E = 2T - L = 2T - (T - V) = T + V,$$

so that the sum of the kinetic and potential energies is constant.

## 22.11 Sturm-Liouville Differential Equations

We have seen how optimizing a functional can lead to a differential equation that must be solved. If we are given a differential equation to solve, it can be helpful to know if it is the Euler-Lagrange equation for some functional.

For example, the Sturm-Liouville differential equations have the form

$$\frac{d}{dx}\Big(p(x)\frac{dy}{dx}\Big) + \Big(q(x) + \lambda r(x)\Big)y = 0.$$

This differential equation is the Euler-Lagrange equation for the constrained problem of minimizing the functional

$$\int_{x_1}^{x_2} \Big(p(x)(y'(x))^2 - q(x)(y(x))^2\Big)dx,$$

subject to

$$\int_{x_1}^{x_2} r(x)(y(x))^2 dx = 1.$$

We have more to say about these differential equations elsewhere in these notes.

## 22.12 Exercises

**Exercise 22.1** *Suppose that the cycloid in the brachistochrone problem connects the starting point* $(0, 0)$ *with the point* $(\pi a, -2a)$*, where* $a > 0$*. Show that the time required for the ball to reach the point* $(\pi a, -2a)$ *is* $\pi\sqrt{\frac{a}{g}}$.

**Exercise 22.2** *Show that, for the situation in the previous exercise, the time required for the ball to reach* $(\pi a, -2a)$ *is again* $\pi\sqrt{\frac{a}{g}}$*, if the ball begins rolling at any intermediate point along the cycloid. This is the tautochrone property of the cycloid.*

# Chapter 23

# Sturm-Liouville Problems (Chapter 10,11)

## 23.1 Recalling Some Matrix Theory

In this chapter we stress the similarities between special types of linear differential operators and Hermitian matrices. We begin with a review of the relevant linear algebra.

Every linear operator $T$ on the vector space $\mathbb{C}^N$ of $N$-dimensional complex column vectors is multiplication by an $N$ by $N$ matrix; that is, there is a complex matrix $A$ such that $T(x) = Ax$ for all $x$ in $\mathbb{C}^N$. The space $\mathbb{C}^N$ is an inner-product space under the usual inner product, or dot product, $\langle x, y\rangle$ given by

$$\langle x, y\rangle = \sum_{n=1}^{N} x_n \overline{y_n}. \tag{23.1}$$

Note that the inner product can be written as

$$\langle x, y\rangle = y^\dagger x. \tag{23.2}$$

We call a matrix $A$ "real" if all its entries are real numbers. A matrix $A$ is Hermitian if $A^\dagger = A$, where $A^\dagger$ denotes the conjugate transpose of $A$. If $A$ is real and Hermitian then $A^T = A$, so $A$ is symmetric.

We have defined what it means for $A$ to be real and to be Hermitian in terms of the entries of $A$; if we are to extend these notions to linear differential operators we will need to define these notions differently. It is easy to see that $A$ is real if and only if $\langle Au, v\rangle$ is a real number, for every real $u$ and $v$ in $\mathbb{C}^N$, and $A$ is Hermitian if and only if

$$\langle Au, v\rangle = \langle u, Av\rangle, \tag{23.3}$$

for every $u$ and $v$ in $\mathbb{C}^N$. These definitions we will be able to extend later.

The Hermitian matrices have the nicest properties and ones we wish to extend to linear differential operators. A non-zero vector $u$ in $\mathbb{C}^N$ is an eigenvector of $A$ with associated eigenvalue $\lambda$ if $Au = \lambda u$.

**Proposition 23.1** *If $A$ is Hermitian, then all its eigenvalues are real.*

**Proof:** We have

$$\langle Au, u\rangle = \langle \lambda u, u\rangle = \lambda\langle u, u\rangle,$$

and

$$\langle Au, u\rangle = \langle u, Au\rangle = \overline{\lambda}\langle u, u\rangle.$$

Since $\langle u, u\rangle$ is not zero, we may conclude that $\overline{\lambda} = \lambda$, or that $\lambda$ is a real number. ∎

**Proposition 23.2** *If $A$ is Hermitian and $Au^m = \lambda_m u^m$ and $Au^n = \lambda_n u^n$, with $\lambda_m \neq \lambda_n$, then $\langle u^m, u^n\rangle = 0$, so $u^m$ and $u^n$ are orthogonal.*

**Proof:** We have

$$\langle Au^m, u^n\rangle = \lambda_m\langle u^m, u^n\rangle,$$

and

$$\langle Au^m, u^n\rangle = \langle u^m, Au^n\rangle = \lambda_n\langle u^m, u^n\rangle.$$

Since $\lambda_m \neq \lambda_n$, it follows that $\langle u^m, u^n\rangle = 0$. ∎

When we change the inner product on $\mathbb{C}^N$ the Hermitian matrices may no longer be the ones we focus on. For any inner product on $\mathbb{C}^N$ we say that a matrix $B$ is *self-adjoint* if

$$\langle Bu, v\rangle = \langle u, Bv\rangle, \tag{23.4}$$

for all $u$ and $v$ in $\mathbb{C}^N$. For example, suppose that $Q$ is a positive-definite $N$ by $N$ matrix, which means that $Q = C^2$, where $C$ is a Hermitian, invertible matrix. We then define the $Q$-inner product to be

$$\langle u, v\rangle_Q = v^\dagger Qu = (Cv)^\dagger Cu. \tag{23.5}$$

We say that a matrix $B$ is self-adjoint with respect to the $Q$-inner product if

$$\langle Bu, v\rangle_Q = \langle u, Bv\rangle_Q, \tag{23.6}$$

or, equivalently,

$$v^\dagger QBu = (Bv)^\dagger Qu = v^\dagger B^\dagger Qu, \tag{23.7}$$

for all $u$ and $v$ in $\mathbb{C}^N$. This means that $QB = B^\dagger Q$ or that the matrix $QB$ is Hermitian. If $QB = BQ$, so that $B$ and $Q$ commute, then $B^\dagger = B$ and

$B$ is Hermitian; in general, however, $B$ being self-adjoint for the $Q$-inner product is different from $B$ being Hermitian.

For a general linear operator $T$ on an inner-product space we shall say that $T$ is *self-adjoint* for the given inner product if

$$\langle Tu, v\rangle = \langle u, Tv\rangle, \tag{23.8}$$

for all $u$ and $v$ in the space.

## 23.2 The Sturm-Liouville Form

We begin with the second-order linear homogeneous ordinary differential equation in standard form,

$$y''(x) + P(x)y'(x) + Q(x)y(x) = 0. \tag{23.9}$$

Let $F(x) = \int P$ denote an anti-derivative of $P(x)$, that is, $F'(x) = P(x)$, and let $S(x) = \exp(F(x))$. Then

$$S(x)y''(x) + S(x)F'(x)y'(x) + S(x)Q(x)y(x) = 0, \tag{23.10}$$

so that

$$\frac{d}{dx}(S(x)y'(x)) + S(x)Q(x)y(x) = 0, \tag{23.11}$$

or

$$\frac{d}{dx}(p(x)y'(x)) + g(x)y(x) = 0. \tag{23.12}$$

This is the Sturm-Liouville form for the differential equation in Equation (23.9).

We shall be particularly interested in differential equations having the Sturm-Liouville form

$$\frac{d}{dx}(p(x)y'(x)) - w(x)q(x)y(x) + \lambda w(x)y(x) = 0, \tag{23.13}$$

where $w(x) > 0$ and $\lambda$ is a constant. Rewriting Equation (23.13) as

$$-\frac{1}{w(x)}\Big(\frac{d}{dx}(p(x)y'(x))\Big) + q(x)y(x) = \lambda y(x) \tag{23.14}$$

suggests an analogy with the linear algebra eigenvalue problem

$$Au = \lambda u, \tag{23.15}$$

where $A$ is a square matrix, $\lambda$ is an eigenvalue of $A$, and $u \neq 0$ is an associated eigenvector. It also suggests that we study the linear differential operator

$$(Ly)(x) = -\frac{1}{w(x)}\Big(\frac{d}{dx}(p(x)y'(x))\Big) + q(x)y(x) \tag{23.16}$$

to see if we can carry the analogy with linear algebra further.

## 23.3 Inner Products and Self-Adjoint Differential Operators

For the moment, let $V_0$ be the vector space of complex-valued integrable functions $f(x)$, defined for $a \leq x \leq b$, for which

$$\int_a^b |f(x)|^2 dx < \infty.$$

For any $f$ and $g$ in $V_0$ the inner product of $f$ and $g$ is then

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)}dx. \tag{23.17}$$

Let $V_1$ be the subspace of functions $y(x)$ in $V_0$ that are twice continuously differentiable. Now let $V$ be the subspace of $V_1$ consisting of all $y(x)$ with $y(a) = y(b) = 0$. A linear operator $T$ on $V$ is said to be *self-adjoint* with respect to the inner product in Equation (23.17) if

$$\int_a^b (Tf)(x)\overline{g(x)}dx = \int_a^b f(x)\overline{(Tg)(x)}dx, \tag{23.18}$$

for all $f(x)$ and $g(x)$ in $V$.

### 23.3.1 An Example of a Self-Adjoint Operator

The linear differential operator $Sy = iy'$ is self-adjoint. Using integration by parts, we have

$$\langle Sf, g \rangle = i\int_a^b f'(x)\overline{g(x)}dx = i[f(x)\overline{g(x)}|_a^b - \int_a^b f(x)\overline{g'(x)}dx$$

$$= \overline{i\int_a^b g'(x)\overline{f(x)}dx} = \overline{\langle Sg, f \rangle} = \langle f, Sg \rangle.$$

### 23.3.2 Another Example

The linear differential operator

$$Ty = y''$$

is defined for the subspace $V$.

**Proposition 23.3** *The operator $Ty = y''$ is self-adjoint on $V$.*

**Proof:** Note that $T = -S^2$. Therefore, we have

$$\langle Tf, g\rangle = -\langle S^2 f, g\rangle = -\langle Sf, Sg\rangle$$

$$= -\langle f, S^2 g\rangle = \langle f, Tg\rangle.$$

∎

It is useful to note that

$$\langle Ty, y\rangle = -\int_a^b |y'(x)|^2 dx \leq 0,$$

for all $y(x)$ in $V$, which prompts us to say that the differential operator $(-T)y = S^2 y = -y''$ is *non-negative definite*. We then expect all eigenvalues of $-T$ to be non-negative. We know, in particular, that solutions of

$$-y''(x) = \lambda y(x),$$

with $y(0) = y(1) = 0$ are $y_m(x) = \sin(m\pi x)$, and the eigenvalues are $\lambda_m = m^2\pi^2$.

### 23.3.3 The Sturm-Liouville Operator

We turn now to the differential operator $L$ given by Equation (23.16). We take $V_0$ to be all complex-valued integrable functions $f(x)$ with

$$\int_a^b |f(x)|^2 w(x) dx < \infty.$$

We let the inner product of any $f(x)$ and $g(x)$ in $V_0$ be

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}w(x)dx. \tag{23.19}$$

Let $V_1$ be all functions in $V_0$ that are twice continuously differentiable, and $V$ all the functions $y(x)$ in $V_1$ with $y(a) = y(b) = 0$. We then have the following result.

**Theorem 23.1** *The operator $L$ given by Equation (23.16) is self-adjoint on the inner product space $V$.*

**Proof:** From

$$(pyz' - pzy')' = (pz')'y - (py')'z$$

we have

$$(Ly)z - y(Lz) = \frac{1}{w(x)}\frac{d}{dx}(pyz' - py'z).$$

Therefore,

$$\int_a^b \Big((Ly)z - y(Lz)\Big)w(x)dx = (pyz' - py'z)|_a^b = 0.$$

Therefore, $L$ is self-adjoint on $V$. ∎

It is interesting to note that

$$\langle Ly, y\rangle = \int_a^b p(y')^2 dx + \int_a^b qy^2 dx,$$

so that, if we have $p(x) \geq 0$ and $q(x) \geq 0$, then the operator $L$ is non-negative-definite and we expect all its eigenvalues to be non-negative.

A square matrix $Q$ is *non-negative definite* if and only if it has the form $Q = C^2$, for some Hermitian matrix $C$; the non-negative definite matrices are therefore analogous to the non-negative real numbers in that each is a square. As we just saw, the differential operator $Ly = -y''$ is self-adjoint and non-negative definite. By analogy with the matrix case, we would expect to be able to write the operator $L$ as $L = C^2$, where $C$ is some self-adjoint linear differential operator. In fact, this is true for the operator $Cy = Ty = iy'$.

## 23.4 Orthogonality

Once again, let $V$ be the space of all twice continuously differentiable functions $y(x)$ on $[a, b]$ with $y(a) = y(b) = 0$. Let $\lambda_m$ and $\lambda_n$ be distinct eigenvalues of the linear differential operator $L$ given by Equation (23.16), with associated eigenfunctions $u_m(x)$ and $u_n(x)$, respectively. Let the inner product on $V$ be given by Equation (23.19).

**Theorem 23.2** *The eigenfunctions $u_m(x)$ and $u_n(x)$ are orthogonal.*

**Proof:** We have

$$\frac{d}{dx}(p(x)u_m'(x)) - w(x)q(x)u_m(x) = -\lambda_m u_m(x)w(x),$$

and

$$\frac{d}{dx}(p(x)u_n'(x)) - w(x)q(x)u_n(x) = -\lambda_n u_n(x)w(x),$$

so that

$$u_n(x)\frac{d}{dx}(p(x)u_m'(x)) - w(x)q(x)u_m(x)u_n(x) = -\lambda_m u_m(x)u_n(x)w(x)$$

and

$$u_m(x)\frac{d}{dx}(p(x)u_n'(x)) - w(x)q(x)u_m(x)u_n(x) = -\lambda_n u_m(x)u_n(x)w(x).$$

Subtracting one equation from the other, we get

$$u_n(x)\frac{d}{dx}(p(x)u'_m(x)) - u_m(x)\frac{d}{dx}(p(x)u'_n(x)) = (\lambda_n - \lambda_m)u_m(x)u_n(x)w(x).$$

The left side of the previous equation can be written as

$$u_n(x)\frac{d}{dx}(p(x)u'_m(x)) - u_m(x)\frac{d}{dx}(p(x)u'_n(x))$$

$$= \frac{d}{dx}\Big(p(x)u_n(x)u'_m(x) - p(x)u_m(x)u'_n(x)\Big).$$

Therefore,

$$(\lambda_n - \lambda_m)\int_a^b u_m(x)u_n(x)w(x)dx =$$

$$\Big(p(x)u_n(x)u'_m(x) - p(x)u_m(x)u'_n(x)\Big)|_a^b = 0. \tag{23.20}$$

Since $\lambda_m \neq \lambda_n$, it follows that

$$\int_a^b u_m(x)u_n(x)w(x)dx = 0.$$

∎

Note that it is not necessary to have $u_m(a) = u_m(b) = 0$ for all $m$ in order for the right side of Equation (23.20) to be zero; it is enough to have

$$p(a)u_m(a) = p(b)u_m(b) = 0.$$

We shall make use of this fact in our discussion of Bessel's and Legendre's equations.

## 23.5 Normal Form of Sturm-Liouville Equations

We can put an equation in the Sturm-Liouville form into normal form by first writing it in standard form. There is a better way, though. With the change of variable from $x$ to $\mu$, where

$$\mu(x) = \int_a^x \frac{1}{p(t)}dt,$$

and

$$\mu'(x) = 1/p(x),$$

we can show that

$$\frac{dy}{dx} = \frac{1}{p(x)}\frac{dy}{d\mu}$$

and

$$\frac{d^2y}{dx^2} = \frac{1}{p^2}\frac{d^2y}{d\mu^2} - \frac{p'(x)}{p(x)}\frac{dy}{d\mu}.$$

It follows that

$$\frac{d^2y}{d\mu^2} + q_1(\mu)y = 0. \tag{23.21}$$

For that reason, we study equations of the form

$$y'' + q(x)y = 0. \tag{23.22}$$

## 23.6 Examples

In this section we present several examples. We shall study these in more detail later in these notes.

### 23.6.1 Wave Equations

Separating the variables to solve wave equations leads to important ordinary differential equations.

#### The Homogeneous Vibrating String

The wave equation for the homogeneous vibrating string is

$$T\frac{\partial^2 u}{\partial x^2} = m\frac{\partial^2 u}{\partial t^2}, \tag{23.23}$$

where $T$ is the constant tension and $m$ the constant mass density. Separating the variables leads to the differential equation

$$-y''(x) = \lambda y(x). \tag{23.24}$$

#### The Non-homogeneous Vibrating String

When the mass density $m(x)$ varies with $x$, the resulting wave equation becomes

$$T\frac{\partial^2 u}{\partial x^2} = m(x)\frac{\partial^2 u}{\partial t^2}. \tag{23.25}$$

Separating the variables leads to the differential equation

$$-\frac{T}{m(x)}y''(x) = \lambda y(x). \tag{23.26}$$

**The Vibrating Hanging Chain**

In the hanging chain problem, considered in more detail later, the tension is not constant along the chain, since at each point it depends on the weight of the part of the chain below. The wave equation becomes

$$\frac{\partial^2 u}{\partial t^2} = g\frac{\partial}{\partial x}\Big(x\frac{\partial u}{\partial x}\Big). \tag{23.27}$$

Separating the variables leads to the differential equation

$$-g\frac{d}{dx}\Big(x\frac{dy}{dx}\Big) = \lambda y(x). \tag{23.28}$$

Note that all three of these differential equations have the form

$$Ly = \lambda y,$$

for $L$ given by Equation (23.16).

If we make the change of variable

$$z = 2\sqrt{\frac{\lambda x}{g}},$$

the differential equation in (23.28) becomes

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + (z^2 - 0^2)y = 0. \tag{23.29}$$

As we shall see shortly, this is a special case of Bessel's Equation, with $\nu = 0$.

### 23.6.2 Bessel's Equations

For each non-negative constant $\nu$ the associated Bessel's Equation is

$$x^2y''(x) + xy'(x) + (x^2 - \nu^2)y(x) = 0. \tag{23.30}$$

Note that the differential equation in Equation (23.28) has the form $Ly = \lambda y$, but Equation (23.29) was obtained by a change of variable that absorbed the $\lambda$ into the $z$, so we do not expect this form of the equation to be in eigenvalue form. However, we can rewrite Equation (23.30) as

$$-\frac{1}{x}\frac{d}{dx}\Big(xy'(x)\Big) + \frac{\nu^2}{x^2}y(x) = y(x), \tag{23.31}$$

which is in the form of a Sturm-Liouville eigenvalue problem, with $w(x) = x = p(x)$, $q(x) = \frac{\nu^2}{x^2}$, and $\lambda = 1$. As we shall discuss again in the chapter

on Bessel's Equations, we can use this fact to obtain a family of orthogonal eigenfunctions.

Let us fix $\nu$ and denote by $J_\nu(x)$ a solution of Equation (23.30). Then $J_\nu(x)$ solves the eigenvalue problem in Equation (23.31), for $\lambda = 1$. A little calculation shows that for any $a$ the function $u(x) = J_\nu(ax)$ satisfies the eigenvalue problem

$$-\frac{1}{x}\frac{d}{dx}\Big(xy'(x)\Big) + \frac{\nu^2}{x^2}y(x) = a^2 y(x). \tag{23.32}$$

Let $\gamma_m > 0$ be the positive roots of $J_\nu(x)$ and define $y_m(x) = J_\nu(\gamma_m x)$ for each $m$. Then we have

$$-\frac{1}{x}\frac{d}{dx}\Big(xy_m'(x)\Big) + \frac{\nu^2}{x^2}y_m(x) = \gamma_m^2 y_m(x), \tag{23.33}$$

and $y_m(1) = 0$ for each $m$. We have the following result.

**Theorem 23.3** *Let $\gamma_m$ and $\gamma_n$ be distinct positive zeros of $J_\nu(x)$. Then*

$$\int_0^1 y_m(x)y_n(x)xdx = 0.$$

**Proof:** The proof is quite similar to the proof of Theorem 23.2. The main point is that now

$$\Big(xy_n(x)y_m'(x) - xy_m(x)y_n'(x)\Big)|_0^1 = 0$$

because $y_m(1) = 0$ for all $m$ and the function $w(x) = x$ is zero when $x = 0$.

### 23.6.3 Legendre's Equations

Legendre's equations have the form

$$(1-x^2)y''(x) - 2xy'(x) + p(p+1)y(x) = 0, \tag{23.34}$$

where $p$ is a constant. When $p = n$ is a non-negative integer, there is a solution $P_n(x)$ that is a polynomial of degree $n$, containing only even or odd powers, as $n$ is either even or odd; $P_n(x)$ is called the $n$th Legendre polynomial. Since the differential equation in (23.34) can be written as

$$-\frac{d}{dx}\Big((1-x^2)y'(x)\Big) = p(p+1)y(x), \tag{23.35}$$

it is a Sturm-Liouville eigenvalue problem with $w(x) = 1$, $p(x) = (1-x^2)$ and $q(x) = 0$. The polynomials $P_n(x)$ are eigenfunctions of the Legendre differential operator $T$ given by

$$(Ty)(x) = -\frac{d}{dx}\Big((1-x^2)y'(x)\Big), \tag{23.36}$$

but we have not imposed any explicit boundary conditions. Nevertheless, we have the following orthogonality theorem.

**Theorem 23.4** *For $m \neq n$ we have*

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0.$$

**Proof:** In this case, Equation (23.20) becomes

$$(\lambda_n - \lambda_m)\int_{-1}^{1} P_m(x)P_n(x)dx =$$

$$\Big((1-x^2)[P_n(x)P_m'(x) - P_m(x)P_n'(x)]\Big)|_{-1}^{1} = 0, \tag{23.37}$$

which holds not because we have imposed end-point conditions on the $P_n(x)$, but because $p(x) = 1 - x^2$ is zero at both ends. ∎

### 23.6.4 Other Famous Examples

Well known examples of Sturm-Liouville problems also include

- **Chebyshev:**

$$\frac{d}{dx}\Big(\sqrt{1-x^2}\frac{dy}{dx}\Big) + \lambda(1-x^2)^{-1/2}y = 0;$$

- **Hermite:**

$$\frac{d}{dx}\Big(e^{-x^2}\frac{dy}{dx}\Big) + \lambda e^{-x^2}y = 0;$$

  and

- **Laguerre:**

$$\frac{d}{dx}\Big(xe^{-x}\frac{dy}{dx}\Big) + \lambda e^{-x}y = 0.$$

**Exercise 23.1** *For each of the three differential equations just listed, see if you can determine the interval over which their eigenfunctions will be orthogonal.*

# Chapter 24

# Series Solutions for Differential Equations (Chapter 10,11)

## 24.1 First-Order Linear Equations

There are only a few linear equations that can be solved exactly in closed form. For the others, we need different approaches. One such is to find a series representation for the solution. We begin with two simple examples.

### 24.1.1 An Example

Consider the differential equation

$$y' = y. \tag{24.1}$$

We look for a solution of Equation (24.1) of the form

$$y(x) = a_0 + a_1 x + a_2 x^2 + ....$$

Writing

$$y'(x) = a_1 + 2a_2 x + 3a_3 x^2 + ...,$$

and inserting these series into the equation $y' - y = 0$, we have

$$0 = (a_0 - a_1) + (a_1 - 2a_2)x + (2a_2 - 3a_3)x^2 + ....$$

Each coefficient must be zero, from which we determine that

$$a_n = a_0/n!.$$

The solutions then are

$$y(x) = a_0 \sum_{n=0}^{\infty} \frac{x^n}{n!} = a_0 e^x.$$

### 24.1.2 Another Example: The Binomial Theorem

Consider now the differential equation

$$(1+x)y' - py = 0, \tag{24.2}$$

with $y(0) = 1$. Writing

$$y(x) = a_0 + a_1 x + a_2 x^2 + ...,$$

we find that

$$0 = (a_1 - p) + (2a_2 - (p-1)a_1)x + (3a_3 - (p-2)a_2)x^2 + ....$$

Setting each coefficient to zero, we find that

$$y(x) = \sum_{n=0}^{\infty} \frac{p(p-1)(p-2)...(p-n+1)}{n!} x^n.$$

The function $y(x) = (1+x)^p$ can be shown to be the unique solution of the original differential equation. Therefore, we have

$$(1+x)^p = \sum_{n=0}^{\infty} \frac{p(p-1)(p-2)...(p-n+1)}{n!} x^n; \tag{24.3}$$

this is the *Binomial Theorem*, with the series converging for $|x| < 1$.

## 24.2 Second-Order Problems

We turn now to the second-order problem

$$y''(x) + P(x)y'(x) + Q(x)y = 0. \tag{24.4}$$

If both $P(x)$ and $Q(x)$ have Taylor series expansions that converge in a neighborhood of $x = x_0$, we say that $x_0$ is an *ordinary point* for the differential equation. In that case, we expect to find a Taylor series representation for the solution that converges in a neighborhood of $x_0$.

If $x_0$ is not an ordinary point, but both $(x-x_0)P(x)$ and $(x-x_0)^2Q(x)$ have Taylor series expansions that converge in a neighborhood of $x_0$, we say that $x_0$ is a *regular singular point* of the differential equation. In such cases, we seek a *Frobenius series* solution.

## 24.3 Ordinary Points

We consider several examples of equations for which $x = 0$ is an ordinary point.

### 24.3.1 The Wave Equation

When we separate variables in the vibrating string problem we find that we have to solve the equation

$$y'' + y = 0. \tag{24.5}$$

Writing the solution as

$$y(x) = \sum_{n=0}^{\infty} a_n x^n,$$

we find that

$$y(x) = a_0\Big(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + ...\Big) + a_1\Big(x - \frac{x^3}{3!} + \frac{x^5}{5!} - ...\Big)$$

so that

$$y(x) = a_0 \cos x + a_1 \sin x.$$

### 24.3.2 Legendre's Equations

Legendre's Equations have the form

$$(1 - x^2)y'' - 2xy' + p(p + 1)y = 0. \tag{24.6}$$

Writing

$$y(x) = \sum_{n=0}^{\infty} a_n x^n,$$

we find that

$$y(x) = a_0\Big(1 - \frac{p(p+1)}{2!}x^2 + \frac{p(p-2)(p+1)(p+3)}{4!}x^4 - ...\Big)$$

$$+a_1\Big(x - \frac{(p-1)(p+2)}{3!}x^3 + \frac{(p-1)(p-3)(p+2)(p+4)}{5!}x^5 - ...\Big).$$

If $p = n$ is a positive even integer, the first series terminates, and if $p = n$ is an odd positive integer, the second series terminates. In either case, we get the Legendre polynomial solutions, denoted $P_n(x)$.

### 24.3.3 Hermite's Equations

Hermite's Equations have the form

$$y'' - 2xy' + 2py = 0. \tag{24.7}$$

The solutions of equation (24.7) are

$$y(x) = a_0 y_1(x) + a_1 y_2(x),$$

where

$$y_1(x) = 1 - \frac{2p}{2!}x^2 + \frac{2^2 p(p-2)}{4!}x^4 - \frac{2^3 p(p-2)(p-4)}{6!}x^6 + ...,$$

and

$$y_2(x) = x - \frac{2(p-1)}{3!}x^3 + \frac{2^2(p-1)(p-3)}{5!}x^5 - ...$$

If $p = n$ is a non-negative integer, one of these series terminates and gives the Hermite polynomial solution $H_n(x)$.

## 24.4 Regular Singular Points

We turn now to the case of regular singular points.

### 24.4.1 Motivation

We motivate the Frobenius series approach by considering Euler's differential equation,

$$x^2 y'' + pxy' + qy = 0, \tag{24.8}$$

where both $p$ and $q$ are constants and $x > 0$. Equation (24.8) can be written as

$$y'' + \frac{p}{x}y' + \frac{q}{x^2}y = 0,$$

from which we see that $x = 0$ is a regular singular point.

Changing variables to $z = \log x$, we obtain

$$\frac{d^2y}{dz^2} + (p-1)\frac{dy}{dz} + qy = 0. \tag{24.9}$$

We seek a solution of the form $y(z) = e^{mz}$. Inserting this guess into Equation (24.9), we find that we must have

$$m^2 + (p-1)m + q = 0;$$

this is the *indicial equation*. If the roots $m = m_1$ and $m = m_2$ are distinct, the solutions are $e^{m_1 z}$ and $e^{m_2 z}$. If $m_1 = m_2$, then the solutions are $e^{m_1 z}$ and $ze^{m_1 z}$. Reverting back to the original variables, we find that the solutions are either $y(x) = x^{m_1}$ and $y(x) = x^{m_2}$, or $y(x) = x^{m_1}$ and $y(x) = x^{m_1}\log x$.

### 24.4.2 Frobenius Series

When $p$ is replaced by $\sum_{n=0}^{\infty} p_n x^n$ and $q$ is replaced by $\sum_{n=0}^{\infty} q_n x^n$, we expect solutions to have the form

$$y(x) = x^m \sum_{n=0}^{\infty} a_n x^n,$$

or

$$y(x) = x^m \log x \sum_{n=0}^{\infty} a_n x^n,$$

where $m$ is a root of an indicial equation. This is the Frobenius series approach.

A Frobenius series associated with the singular point $x_0 = 0$ has the form

$$y(x) = x^m\big(a_0 + a_1 x + a_2 x^2 + ...\big), \tag{24.10}$$

where $m$ is to be determined, and $a_0 \neq 0$. Since $xP(x)$ and $x^2Q(x)$ are analytic, we can write

$$xP(x) = p_0 + p_1 x + p_2 x^2 + ..., \tag{24.11}$$

and

$$x^2Q(x) = q_0 + q_1 x + q_2 x^2 + ..., \tag{24.12}$$

with convergence for $|x| < R$. Inserting these expressions into the differential equation, and performing a bit of algebra, we arrive at

$$\sum_{n=0}^{\infty} \Bigg\{ a_n\big[(m+n)(m+n-1) + (m+n)p_0 + q_0\big] + \sum_{k=0}^{n-1} a_k\big[(m+k)p_{n-k} + q_{n-k}\big] \Bigg\} x^n = 0. \tag{24.13}$$

Setting each coefficient to zero, we obtain a recursive algorithm for finding the $a_n$. To start with, we have

$$a_0\big[m(m-1) + mp_0 + q_0\big] = 0. \tag{24.14}$$

Since $a_0 \neq 0$, we must have

$$m(m-1) + mp_0 + q_0 = 0; \tag{24.15}$$

this is called the *Indicial Equation*. We solve the quadratic Equation (24.15) for $m = m_1$ and $m = m_2$.

### 24.4.3 Bessel Functions

Applying these results to Bessel's Equation, we see that $P(x) = \frac{1}{x}$, $Q(x) = 1 - \frac{\nu^2}{x^2}$, and so $p_0 = 1$ and $q_0 = -\nu^2$. The Indicial Equation (24.15) is now

$$m^2 - \nu^2 = 0, \tag{24.16}$$

with solutions $m_1 = \nu$, and $m_2 = -\nu$. The recursive algorithm for finding the $a_n$ is

$$a_n = -a_{n-2}/n(2\nu + n). \tag{24.17}$$

Since $a_0 \neq 0$ and $a_{-1} = 0$, it follows that the solution for $m = \nu$ is

$$y = a_0 x^\nu \left[1 - \frac{x^2}{2^2(\nu+1)} + \frac{x^4}{2^4 2!(\nu+1)(\nu+2)} - ...\right]. \tag{24.18}$$

Setting $a_0 = 1/2^\nu \nu!$, we get the $\nu$th Bessel function,

$$J_\nu(x) = \sum_{n=0}^{\infty} (-1)^n \left(\frac{x}{2}\right)^{2n+\nu} /n!(\nu+n)!. \tag{24.19}$$

The most important Bessel functions are $J_0(x)$ and $J_1(x)$.

There is a potential problem in Equation (24.19). Notice that we have not required that $\nu$ be a non-negative integer, so the term $(\nu + n)!$ may be undefined. This leads to consideration of the gamma function, which extends the factorial function beyond non-negative integers. Equation (24.19) should be written

$$J_\nu(x) = \sum_{n=0}^{\infty} (-1)^n \left(\frac{x}{2}\right)^{2n+\nu} /n!\Gamma(\nu+n+1). \tag{24.20}$$

# Chapter 25

# Bessel's Equations (Chapter 9,10,11)

For each non-negative constant $\nu$, the associated *Bessel Equation* is

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - \nu^2)y = 0, \tag{25.1}$$

which can also be written in the form

$$y'' + P(x)y' + Q(x)y = 0, \tag{25.2}$$

with $P(x) = \frac{1}{x}$ and $Q(x) = 1 - \frac{\nu^2}{x^2}$.

Solutions of Equation (25.1) are *Bessel functions*. These functions first arose in Daniel Bernoulli's study of the oscillations of a hanging chain, and now play important roles in many areas of applied mathematics [42].

We begin this note with Bernoulli's problem, to see how Bessel's Equation becomes involved. We then consider Frobenius-series solutions to second-order linear differential equations with regular singular points; Bessel's Equation is one of these. Once we obtain the Frobenius-series solution of Equation (25.1), we discover that it involves terms of the form $p!$, for (possibly) non-integer $p$. This leads to the *Gamma Function*, which extends the factorial function to such non-integer arguments.

The Gamma Function, defined for $x > 0$ by the integral

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt, \tag{25.3}$$

is a *higher transcendental* function that cannot be evaluated by purely algebraic means, and can only be approximated by numerical techniques. With clever changes of variable, a large number of challenging integration problems can be rewritten and solved in terms of the gamma function.

We prepare for our discussion of Bernoulli's hanging chain problem by recalling some important points in the derivation of the one-dimensional wave equation for the vibrating string problem.

## 25.1 The Vibrating String Problem

In the vibrating string problem, the string is fixed at end-points $(0,0)$ and $(1,0)$. The position of the string at time $t$ is given by $y(x,t)$, where $x$ is the horizontal spatial variable. It is assumed that the string has a constant mass density, $m$. Consider the small piece of the string corresponding to the interval $[x, x+\Delta x]$. Its mass is $m\Delta x$, and so, from Newton's equating of force with mass times acceleration, we have that the force $f$ on the small piece of string is related to acceleration by

$$f \approx m(\Delta x)\frac{\partial^2 y}{\partial t^2}. \tag{25.4}$$

In this problem, the force is not gravitational, but comes from the tension applied to the string; we denote by $T(x)$ the tension in the string at $x$. This tensile force acts along the tangent to the string at every point. Therefore, the force acting on the left end-point of the small piece is directed to the left and is given by $-T(x)\sin(\theta(x))$; at the right end-point it is $T(x+\Delta x)\sin(\theta(x+\Delta x))$, where $\theta(x)$ is the angle the tangent line at $x$ makes with the horizontal. For small-amplitude oscillations of the string, the angles are near zero and the sine can be replaced by the tangent. Since $\tan(\theta(x)) = \frac{\partial y}{\partial x}(x)$, we can write the net force on the small piece of string as

$$f \approx T(x+\Delta x)\frac{\partial y}{\partial x}(x+\Delta x) - T(x)\frac{\partial y}{\partial x}(x). \tag{25.5}$$

Equating the two expressions for $f$ in Equations (25.4) and (25.5) and dividing by $\Delta x$, we obtain

$$\frac{T(x+\Delta x)\frac{\partial y}{\partial x}(x+\Delta x) - T(x)\frac{\partial y}{\partial x}(x)}{\Delta x} \approx m\frac{\partial^2 y}{\partial t^2}. \tag{25.6}$$

Taking limits, as $\Delta x \to 0$, we arrive at the *Wave Equation*

$$\frac{\partial}{\partial x}\Big(T(x)\frac{\partial y}{\partial x}(x)\Big) = m\frac{\partial^2 y}{\partial t^2}. \tag{25.7}$$

For the vibrating string problem, we also assume that the tension function is constant, that is, $T(x) = T$, for all $x$. Then we can write Equation (25.7) as the more familiar

$$T\frac{\partial^2 y}{\partial x^2} = m\frac{\partial^2 y}{\partial t^2}. \tag{25.8}$$

We could have introduced the assumption of constant tension earlier in this discussion, but we shall need the wave equation for variable tension Equation (25.7) when we consider the hanging chain problem.

## 25.2 The Hanging Chain Problem

Imagine a flexible chain hanging vertically. Assume that the chain has a constant mass density $m$. Let the origin $(0, 0)$ be the bottom of the chain, with the positive $x$-axis running vertically, up through the chain. The positive $y$-axis extends horizontally to the left, from the bottom of the chain. As before, the function $y(x, t)$ denotes the position of each point on the chain at time $t$. We are interested in the oscillation of the hanging chain. This is the vibrating string problem turned on its side, except that now the tension is not constant.

### 25.2.1 The Wave Equation for the Hanging Chain

The tension at the point $x$ along the chain is due to the weight of the portion of the chain below the point $x$, which is then $T(x) = mgx$. Applying Equation (25.7), we have

$$\frac{\partial}{\partial x}\Big(mgx\frac{\partial y}{\partial x}(x)\Big) = m\frac{\partial^2 y}{\partial t^2}. \tag{25.9}$$

As we normally do at this stage, we separate the variables, to find potential solutions.

### 25.2.2 Separating the Variables

We consider possible solutions having the form

$$y(x, t) = u(x)v(t). \tag{25.10}$$

Inserting this $y(x, t)$ into Equation (25.9), and doing a bit of algebra, we arrive at

$$gxu''(x) + gu'(x) + \lambda u(x) = 0, \tag{25.11}$$

and

$$v''(t) + \lambda v(t) = 0, \tag{25.12}$$

where $\lambda$ is the separation constant. It is Equation (25.11), which can also be written as

$$\frac{d}{dx}(gxu'(x)) + \lambda u(x) = 0, \tag{25.13}$$

that interests us here.

### 25.2.3 Obtaining Bessel's Equation

With a bit more work, using the change of variable $z = 2\sqrt{\frac{\lambda}{g}}\sqrt{x}$ and the Chain Rule (no pun intended!), we find that we can rewrite Equation (25.11) as

$$z^2\frac{d^2u}{dz^2} + z\frac{du}{dz} + (z^2 - 0^2)u = 0, \tag{25.14}$$

which is Bessel's Equation (25.1), with the parameter value $\nu = 0$.

## 25.3 Solving Bessel's Equations

Second-order linear differential equations with the form

$$y''(x) + P(x)y'(x) + Q(x)y(x) = 0, \tag{25.15}$$

with neither $P(x)$ nor $Q(x)$ analytic at $x = x_0$, but with both $(x - x_0)P(x)$ and $(x - x_0)^2Q(x)$ analytic, are said to be equations with *regular singular points*. Writing Equation (25.1) as

$$y''(x) + \frac{1}{x}y'(x) + (1 - \frac{\nu^2}{x^2})y(x) = 0, \tag{25.16}$$

we see that Bessel's Equation is such a regular singular point equation, with the singular point $x_0 = 0$. Solutions to such equations can be found using the technique of Frobenius series.

### 25.3.1 Frobenius-series solutions

A Frobenius series associated with the singular point $x_0 = 0$ has the form

$$y(x) = x^m\big(a_0 + a_1x + a_2x^2 + ...\big), \tag{25.17}$$

where $m$ is to be determined, and $a_0 \neq 0$. Since $xP(x)$ and $x^2Q(x)$ are analytic, we can write

$$xP(x) = p_0 + p_1x + p_2x^2 + ..., \tag{25.18}$$

and

$$x^2Q(x) = q_0 + q_1x + q_2x^2 + ..., \tag{25.19}$$

with convergence for $|x| < R$. Inserting these expressions into the differential equation, and performing a bit of algebra, we arrive at

$$\sum_{n=0}^{\infty}\Big\{a_n\big[(m+n)(m+n-1) + (m+n)p_0 + q_0\big] +$$

$$\sum_{k=0}^{n-1} a_k\big[(m+k)p_{n-k}+q_{n-k}\big]\Bigg\}x^n = 0. \tag{25.20}$$

Setting each coefficient to zero, we obtain a recursive algorithm for finding the $a_n$. To start with, we have

$$a_0\big[m(m-1)+mp_0+q_0\big] = 0. \tag{25.21}$$

Since $a_0 \neq 0$, we must have

$$m(m-1)+mp_0+q_0 = 0; \tag{25.22}$$

this is called the *Indicial Equation*. We solve the quadratic Equation (25.22) for $m = m_1$ and $m = m_2$.

### 25.3.2 Bessel Functions

Applying these results to Bessel's Equation (25.1), we see that $P(x) = \frac{1}{x}$, $Q(x) = 1 - \frac{\nu^2}{x^2}$, and so $p_0 = 1$ and $q_0 = -\nu^2$. The Indicial Equation (25.22) is now

$$m^2 - \nu^2 = 0, \tag{25.23}$$

with solutions $m_1 = \nu$, and $m_2 = -\nu$. The recursive algorithm for finding the $a_n$ is

$$a_n = -a_{n-2}/n(2\nu+n). \tag{25.24}$$

Since $a_0 \neq 0$ and $a_{-1} = 0$, it follows that the solution for $m = \nu$ is

$$y = a_0 x^\nu \left[1 - \frac{x^2}{2^2(\nu+1)} + \frac{x^4}{2^4 2!(\nu+1)(\nu+2)} - ...\right]. \tag{25.25}$$

Setting $a_0 = 1/2^\nu \nu!$, we get the $\nu$th Bessel function,

$$J_\nu(x) = \sum_{n=0}^{\infty}(-1)^n\Big(\frac{x}{2}\Big)^{2n+\nu}/n!(\nu+n)!. \tag{25.26}$$

The most important Bessel functions are $J_0(x)$ and $J_1(x)$.

**We have a Problem!** So far, we have allowed $\nu$ to be any real number. What, then, do we mean by $\nu!$ and $(n+\nu)!$? To answer this question, we need to investigate the gamma function.

## 25.4 Bessel Functions of the Second Kind

If $\nu$ is not an integer, then $J_\nu(x)$ and $J_{-\nu}(x)$ are linearly independent and the complete solution of Equation (25.1) is

$$y(x) = AJ_\nu(x) + BJ_{-\nu}(x). \tag{25.27}$$

If $\nu = n$ is an integer, then

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

For $n = 0, 1, ...$, the Bessel function of the second kind, of order $n$, is

$$Y_n(x) = \lim_{\nu \to n} \frac{J_\nu(x)\cos\nu\pi - J_{-\nu}(x)}{\sin\nu\pi}. \tag{25.28}$$

The general solution of Equation (25.1), for $\nu = n$, is then

$$y(x) = AJ_n(x) + BY_n(x).$$

## 25.5 Hankel Functions

The Hankel functions of the first and second kind are

$$H_n^{(1)}(x) = J_n(x) + iY_n(x), \tag{25.29}$$

and

$$H_n^{(2)}(x) = J_n(x) - iY_n(x). \tag{25.30}$$

## 25.6 The Gamma Function

We want to define $\nu!$ for $\nu$ not a non-negative integer. The Gamma Function is the way to do this.

### 25.6.1 Extending the Factorial Function

As we said earlier, the Gamma Function is defined for $x > 0$ by

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt. \tag{25.31}$$

Using integration by parts, it is easy to show that

$$\Gamma(x+1) = x\Gamma(x). \tag{25.32}$$

Using Equation (25.32) and the fact that

$$\Gamma(1) = \int_0^\infty e^{-t} dt = 1, \tag{25.33}$$

we obtain

$$\Gamma(n+1) = n!, \tag{25.34}$$

for $n = 0, 1, 2, ....$

### 25.6.2 Extending $\Gamma(x)$ to negative $x$

We can use

$$\Gamma(x) = \frac{\Gamma(x+1)}{x} \tag{25.35}$$

to extend $\Gamma(x)$ to any $x < 0$, with the exception of the negative integers, at which $\Gamma(x)$ is unbounded.

### 25.6.3 An Example

We have

$$\Gamma(\frac{1}{2}) = \int_0^\infty e^{-t} t^{-1/2} dt. \tag{25.36}$$

Therefore, using $t = u^2$, we have

$$\Gamma(\frac{1}{2}) = 2\int_0^\infty e^{-u^2} du. \tag{25.37}$$

Squaring, we get

$$\Gamma(\frac{1}{2})^2 = 4\int_0^\infty \int_0^\infty e^{-u^2} e^{-v^2} du dv. \tag{25.38}$$

In polar coordinates, this becomes

$$\Gamma(\frac{1}{2})^2 = 4\int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2} r dr d\theta$$

$$= 2\int_0^{\frac{\pi}{2}} 1 d\theta = \pi. \tag{25.39}$$

Consequently, we have

$$\Gamma(\frac{1}{2}) = \sqrt{\pi}. \tag{25.40}$$

## 25.7 Representing the Bessel Functions

There are several equivalent ways to represent the Bessel functions.

### 25.7.1 Taylor Series

The Bessel function of the first kind and order $n$, $J_n(x)$, is sometimes defined by the infinite series

$$J_n(x) = \sum_{m=0}^{\infty} \frac{(-1)^m (x/2)^{n+2m}}{m!\Gamma(n+m+1)}, \tag{25.41}$$

for $n = 0, 1, \ldots$ . The series converges for all $x$. From Equation (25.41) we have

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 4^2} - \frac{x^6}{2^2 4^2 6^2} + \ldots, \tag{25.42}$$

from which it follows immediately that $J_0(-x) = J_0(x)$.

### 25.7.2 Generating Function

For each fixed $x$, the function of the complex variable $z$ given by

$$f(z) = \exp(\frac{x}{2}(z - 1/z))$$

has the Laurent series expansion

$$\exp(\frac{x}{2}(z - 1/z)) = \sum_{n=-\infty}^{\infty} J_n(x)\, z^n. \tag{25.43}$$

Using Cauchy's formula for the coefficients of a Laurent series, we find that

$$J_n(x) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z^{n+1}} dz, \tag{25.44}$$

for any simple closed curve $C$ surrounding the essential singularity $z = 0$.

### 25.7.3 An Integral Representation

Now we take as $C$ the circle of radius one about the origin, so that $z = e^{i\theta}$, for $\theta$ in the interval $[0, 2\pi]$. Rewriting Equation (25.44), we get

$$J_n(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{i(x \sin\theta - n\theta)} d\theta. \tag{25.45}$$

By symmetry, we can also write

$$J_n(x) = \frac{1}{\pi}\int_0^{\pi} \cos(x\sin\theta - n\theta)d\theta. \tag{25.46}$$

From Equation (25.45) we have

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{i(x\sin\theta)}d\theta, \tag{25.47}$$

or, equivalently,

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{i(x\cos\theta)}d\theta. \tag{25.48}$$

## 25.8 Fourier Transforms and Bessel Functions

Bessel functions are closely related to Fourier transforms.

### 25.8.1 The Case of Two Dimensions

Let $f(x, y)$ be a complex-valued function of the two real variables $x$ and $y$. Then its Fourier transform is the function $F(\alpha, \beta)$ of two real variables defined by

$$F(\alpha,\beta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{i\alpha x}e^{i\beta y}dxdy. \tag{25.49}$$

The Fourier Inversion Formula then gives

$$f(x,y) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(\alpha,\beta)e^{-i\alpha x}e^{-i\beta y}d\alpha d\beta. \tag{25.50}$$

### 25.8.2 The Case of Radial Functions

Suppose that we express $f(x, y)$ in polar coordinates $r$ and $\theta$, with $x = r\cos\theta$ and $y = r\sin\theta$. The function $f(x, y)$ is said to be *radial* if, when expressed in polar coordinates, it is independent of $\theta$. Another way to say this is there is some function $g(t)$ such that

$$f(x,y) = g(\sqrt{x^2+y^2}) = g(r), \tag{25.51}$$

for each $x$ and $y$; that is, $f$ is constant on circles centered at the origin.

When $f(x, y)$ is radial, so if $F(\alpha, \beta)$, as we now prove. We begin by expressing $F(\alpha, \beta)$ in polar coordinates as well, with $\alpha = \rho\cos\omega$ and $\beta = \rho\sin\omega$. Then

$$F(\rho\cos\omega, \rho\sin\omega) = \int_0^{\infty}\int_0^{2\pi} g(r)e^{ir\rho\cos(\theta-\omega)}d\theta r dr$$

$$= \int_0^\infty \Big( \int_0^{2\pi} e^{ir\rho\cos(\theta-\omega)} d\theta \Big) g(r) r dr.$$

By making the variable substitution of $\gamma = \theta - \omega$, it is easy to show that the inner integral,

$$\int_0^{2\pi} e^{ir\rho\cos(\theta-\omega)} d\theta,$$

is actually independent of $\omega$, which tells us that $F$ is radial; we then write $F(\rho\cos\omega, \rho\sin\omega) = H(\rho)$.

From Equation (25.48) we know that

$$2\pi J_0(r\rho) = \int_0^{2\pi} e^{ir\rho\cos(\theta)} d\theta \tag{25.52}$$

We then have

$$H(\rho) = 2\pi \int_0^\infty r g(r) J_0(r\rho) dr. \tag{25.53}$$

There are several things to notice here.

### 25.8.3 The Hankel Transform

First, note that when $f(x, y)$ is radial, its two-dimensional Fourier transform is also radial, but $H(\rho)$ is not the one-dimensional Fourier transform of $g(r)$. The integral in Equation (25.53) tells us that $\frac{1}{2\pi}H(\rho)$ is the *Hankel transform* of $g(r)$. Because of the similarity between Equations (25.49) and (25.50), we also have

$$g(r) = \frac{1}{2\pi} \int_0^\infty \rho H(\rho) J_0(r\rho) d\rho. \tag{25.54}$$

For any function $s(x)$ of a single real variable, its *Hankel transform* is

$$T(\gamma) = \int_0^\infty x s(x) J_0(\gamma x) dx. \tag{25.55}$$

The inversion formula is

$$s(x) = \frac{1}{2\pi} \int_0^\infty \gamma T(\gamma) J_0(\gamma x) d\gamma. \tag{25.56}$$

## 25.9 An Application of the Bessel Functions in Astronomy

In remote sensing applications, it is often the case that what we measure is the Fourier transform of what we really want. This is the case in medical imaging, for example, in both x-ray tomography and magnetic-resonance imaging. It is also often the case in astronomy. Consider the problem of determining the size of a distant star.

We model the star as a distance disk of uniform brightness. Viewed as a function of two variables, it is the function that, in polar coordinates, can be written as $f(r, \theta) = g(r)$, that is, it is a radial function that is a function of $r$ only, and independent of $\theta$. The function $g(r)$ is, say, one for $0 \leq r \leq R$, where $R$ is the radius of the star, and zero, otherwise. From the theory of Fourier transform pairs in two-dimensions, we know that the two-dimensional Fourier transform of $f$ is also a radial function; it is the function

$$H(\rho) = 2\pi \int_0^R r J_0(r\rho) dr,$$

where $J_0$ is the zero-th order Bessel function of the first kind. From the theory of Bessel functions, we learn that

$$\frac{d}{dx}[xJ_1(x)] = xJ_0(x),$$

so that

$$H(\rho) = \frac{2\pi}{\rho} R J_1(R\rho).$$

When the star is viewed through a telescope, the image is blurred by the atmosphere. It is commonly assumed that the atmosphere performs a convolution filtering on the light from the star, and that this filter is random and varies somewhat from one observation to another. Therefore, at each observation, it is not $H(\rho)$, but $H(\rho)G(\rho)$ that is measured, where $G(\rho)$ is the filter transfer function operating at that particular time.

Suppose we observe the star $N$ times, for each $n = 1, 2, ..., N$ measuring values of the function $H(\rho)G_n(\rho)$. If we then average over the various measurements, we can safely say that the first zero we observe in our measurements is the first zero of $H(\rho)$, that is, the first zero of $J_1(R\rho)$. The first zero of $J_1(x)$ is known to be about 3.8317, so knowing this, we can determine $R$. Actually, it is not truly $R$ that we are measuring, since we also need to involve the distance $D$ to the star, known by other means. What we are measuring is the perceived radius, in other words, half the subtended angle. Combining this with our knowledge of $D$, we get $R$.

## 25.10 Orthogonality of Bessel Functions

As we have seen previously, the orthogonality of trigonometric functions plays an important role in Fourier series. A similar notion of orthogonality holds for Bessel functions. We begin with the following theorem.

**Theorem 25.1** *Let $u(x)$ be a non-trivial solution of $u''(x)+q(x)u(x)=0$. If*

$$\int_1^\infty q(x)dx = \infty\,,$$

*then $u(x)$ has infinitely many zeros on the positive $x$-axis.*

Bessel's Equation

$$x^2y''(x) + xy'(x) + (x^2-\nu^2)y(x) = 0, \tag{25.57}$$

can be written in *normal form* as

$$y''(x) + \left(1 + \frac{1-4\nu^2}{4x^2}\right)y(x) = 0, \tag{25.58}$$

and, as $x \to \infty$,

$$q(x) = 1 + \frac{1-4\nu^2}{4x^2} \to 1,$$

so, according to the theorem, every non-trivial solution of Bessel's Equation has infinitely many positive zeros.

Now consider the following theorem, which is a consequence of the Sturm Comparison Theorem discussed elsewhere in these notes.

**Theorem 25.2** *Let $y_\nu(x)$ be a non-trivial solution of Bessel's Equation*

$$x^2y''(x) + xy'(x) + (x^2-\nu^2)y(x) = 0,$$

*for $x>0$. If $0 \le \nu < \frac{1}{2}$, then every interval of length $\pi$ contains at least one zero of $y_\nu(x)$; if $\nu = \frac{1}{2}$, then the distance between successive zeros of $y_\nu(x)$ is precisely $\pi$; and if $\nu > \frac{1}{2}$, then every interval of length $\pi$ contains at most one zero of $y_\nu(x)$.*

It follows from these two theorems that, for each fixed $\nu$, the function $y_\nu(x)$ has an infinite number of positive zeros, say $\lambda_1 < \lambda_2 < ...$, with $\lambda_n \to \infty$.

For fixed $\nu$, let $y_n(x) = y_\nu(\lambda_n x)$. As we saw earlier, we have the following orthogonality theorem.

**Theorem 25.3** *For $m \neq n$, $\int_0^1 xy_m(x)y_n(x)dx = 0$.*

**Proof:** Let $u(x) = y_m(x)$ and $v(x) = y_n(x)$. Then we have

$$u'' + \frac{1}{x}u' + (\lambda_m^2 - \frac{\nu^2}{x^2})u = 0,$$

and

$$v'' + \frac{1}{x}v' + (\lambda_n^2 - \frac{\nu^2}{x^2})v = 0.$$

Multiplying on both sides by $x$ and subtracting one equation from the other, we get

$$x(uv'' - vu'') + (uv' - vu') = (\lambda_m^2 - \lambda_n^2)xuv.$$

Since

$$\frac{d}{dx}\Big(x(uv' - vu')\Big) = x(uv'' - vu'') + (uv' - vu'),$$

it follows, by integrating both sides over the interval $[0, 1]$, that

$$x(uv' - vu')|_0^1 = (\lambda_m^2 - \lambda_n^2)\int_0^1 xu(x)v(x)dx.$$

But

$$x(uv' - vu')|_0^1 = u(1)v'(1) - v(1)u'(1) = 0.$$

∎

# Chapter 26

# Legendre's Equations (Chapter 10,11)

## 26.1 Legendre's Equations

In this chapter we shall be interested in Legendre's equations of the form

$$(1-x^2)y''(x) - 2xy'(x) + n(n+1)y(x) = 0, \tag{26.1}$$

where $n$ is a non-negative integer. In this case, there is a solution $P_n(x)$ that is a polynomial of degree $n$, containing only even or odd powers, as $n$ is either even or odd; $P_n(x)$ is called the $n$th Legendre polynomial. Since the differential equation in (26.1) can be written as

$$-\frac{d}{dx}\Big((1-x^2)y'(x)\Big) = n(n+1)y(x), \tag{26.2}$$

it is a Sturm-Liouville eigenvalue problem with $w(x) = 1$, $p(x) = (1-x^2)$ and $q(x) = 0$. The polynomials $P_n(x)$ are eigenfunctions of the Legendre differential operator $T$ given by

$$(Ty)(x) = -\frac{d}{dx}\Big((1-x^2)y'(x)\Big), \tag{26.3}$$

but we have not imposed any explicit boundary conditions. Nevertheless, we have the following orthogonality theorem.

**Theorem 26.1** *For $m \neq n$ we have*

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0.$$

**Proof:** In this case, Equation (23.20) becomes

$$(\lambda_n - \lambda_m) \int_{-1}^{1} P_m(x) P_n(x) dx =$$

$$\Big( (1-x^2)[P_n(x)P'_m(x) - P_m(x)P'_n(x)] \Big) |_{-1}^{1} = 0, \tag{26.4}$$

which holds not because we have imposed end-point conditions on the $P_n(x)$, but because $p(x) = 1 - x^2$ is zero at both ends. ∎

From the orthogonality we can conclude that

$$\int_{-1}^{1} P_N(x) Q(x) dx = 0,$$

for any polynomial $Q(x)$ of degree at most $N-1$. This is true because $Q(x)$ can be written as a linear combination of the Legendre polynomials $P_n(x)$, for $n = 0, 1, ..., N-1$.

Using orthogonality we can prove the following theorem:

**Theorem 26.2** *All the $N$ roots of $P_N(x)$ lie in the interval $[-1, 1]$.*

**Proof:** Let $\{x_n | n = 1, 2, ..., N\}$ be the roots of $P_N(x)$. For fixed $n$ let

$$Q_n(x) = \prod_{m \neq n} (x - x_m),$$

so that $P_N(x) = c(x - x_n)Q_n(x)$, for some constant $c$. Then

$$0 = \int_{-1}^{1} P_N(x) Q_n(x) dx = c \int_{-1}^{1} (x - x_n) \prod_{m \neq n} (x - x_m)^2 dx,$$

from which we conclude that $(x - x_n)$ does not have constant sign on the interval $[-1, 1]$. ∎

Now that we know that all $N$ roots of $P_N(x)$ are real, we can use orthogonality again to prove that all the roots are distinct.

**Theorem 26.3** *All the roots of $P_N(x)$ are distinct.*

**Proof:** Suppose that $x_1 = x_2$. Then we can write

$$P_N(x) = c(x - x_1)^2 \prod_{m=3}^{N} (x - x_m) = (x - x_1)^2 Q(x),$$

where $Q(x)$ is a polynomial of degree $N-2$. Therefore,

$$\int_{-1}^{1} P_N(x) Q(x) dx = 0,$$

by orthogonality. But

$$\int_{-1}^{1} P_N(x)Q(x)dx = \int_{-1}^{1} (x - x_1)^2 Q(x)^2 dx,$$

which cannot equal zero, since the integrand is a non-negative polynomial.

## 26.2 Rodrigues' Formula

There is a simple formula, called *Rodrigues' Formula*, for generating the successive Legendre polynomials:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n. \tag{26.5}$$

Using Equation (26.5), we find that

$$P_0(x) = 1,$$

$$P_1(x) = x,$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1),$$

and so on.

**Exercise 26.1** *Calculate* $P_3(x)$.

## 26.3 A Recursive Formula for $P_n(x)$

While Rodrigues' formula is simple to write down, it is not simple to apply. With $n = 0$, we get $P_0(x) = 1$. Then

$$P_1(x) = \frac{1}{2}\frac{d}{dx}[(x^2 - 1)] = \frac{1}{2}(2x) = x,$$

$$P_2(x) = \frac{1}{8}\frac{d^2}{dx^2}[(x^2 - 1)^2] = \frac{1}{8}[12x^2 - 4] = \frac{3}{2}x^2 - \frac{1}{2}.$$

The others follow in the same way, although, as I am sure you will discover, the calculations become increasingly tedious. One approach that simplifies the calculation is to use the binomial theorem to expand $(x^2 - 1)^n$ before differentiating. Only the arithmetic involving the coefficients remains a nuisance then.

They say that necessity is the mother of invention, but I think avoiding tedium can also be a strong incentive. Finding a shortcut is not necessarily a way to save time; in the time spent finding the shortcut, you could probably have solved the original problem three times over. It is easy to make

mistakes in long calculations, though, and a shortcut that requires fewer calculations can be helpful. In a later section we shall see a standard recursive formula that allows us to compute $P_{n+1}(x)$ from $P_n(x)$ and $P_{n-1}(x)$. In this section we try to find our own simplification of Rodrigues' formula.

Here is one idea. Convince yourself that the $n+1$-st derivative of a product of $f(x)$ and $g(x)$ can be written

$$(fg)^{(n+1)} = \sum_{k=0}^{n+1} \binom{n+1}{k} f^{(k)} g^{(n+1-k)},$$

where

$$\binom{n+1}{k} = \frac{(n+1)!}{k!(n+1-k)!}.$$

Now we find $[(x^2-1)^{n+1}]^{(n+1)}$ by defining $f(x) = (x^2-1)^n$ and $g(x) = x^2 - 1$.

Since $g^{(n+1-k)} = 0$, except for $k = n+1, n$, and $n-1$, the sum above has only three terms. Two of the three terms involve $P_n(x)$ and $P_n'(x)$, which we would already have found. The third term involves the anti-derivative of $P_n(x)$. We can easily calculate this anti-derivative, except for the constant. See if you can figure out what the constant must be.

## 26.4 A Generating Function Approach

For each fixed $x$ and variable $t$, the function

$$\frac{1}{\sqrt{1-2xt+t^2}} = P_0(x) + P_1(x)t + P_2(x)t^2 + P_3(x)t^3 + ... + P_n(x)t^n + ....$$

This function is called *the generating function* for the Legendre polynomials.

**Exercise 26.2** *Use the generating function and the Taylor expansion of* $\log(t+1)$ *around* $t = 0$ *to prove that*

$$\int_{-1}^{1} P_n(x)P_n(x)dx = \frac{2}{2n+1}.$$

**Exercise 26.3** *Use the generating function to*

- *(a) verify that* $P_n(1) = 1$ *and* $P_n(-1) = (-1)^n$, *and*
- *(b) show that* $P_{2n+1}(0) = 0$ *and*

$$P_{2n}(0) = \frac{(-1)^n}{2^n n!}(1 \cdot 3 \cdot 5 \cdots (2n-1)).$$

## 26.5 A Two-Term Recursive Formula for $P_n(x)$

Using the generating function, we can show that

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x), \tag{26.6}$$

for $n = 1, 2, ....$ This is a *two-term recursive formula* for the $P_n(x)$.

**Exercise 26.4** *Use the recursive formula to compute $P_n(x)$ for $n = 3, 4,$ and 5.*

## 26.6 Legendre Series

Just as Fourier series deals with the representation of a function on a finite interval as a (possibly infinite) sum of sines and cosines, Legendre series involves the representation of a function $f(x)$ on the interval $[-1, 1]$ as

$$f(x) = a_0P_0(x) + a_1P_1(x) + \cdots a_nP_n(x) + \cdots. \tag{26.7}$$

**Exercise 26.5** *Use the orthogonality of the $P_n(x)$ over the interval $[-1, 1]$ to show that*

$$a_n = \left(n + \frac{1}{2}\right)\int_{-1}^{1} f(x)P_n(x)dx. \tag{26.8}$$

## 26.7 Best Approximation by Polynomials

Suppose that we want to approximate a function $f(x)$ by a polynomial of degree $n$, over the interval $[-1, 1]$. Which polynomial is the best? Since much attention is paid to the Taylor expansion of a function, you might guess that the first $n+1$ terms of the Taylor series for $f(x)$ might be what we want to use, but this is not necessarily the case.

First of all, we need to be clear about what we mean by *best*. Let us agree that we want to find the polynomial

$$p(x) = b_0 + b_1x + b_2x^2 + \cdots b_nx^n$$

that minimizes

$$\int_{-1}^{1} \big(f(x) - p(x)\big)^2 dx. \tag{26.9}$$

It is helpful to note that any polynomial of degree $n$ can be written as

$$p(x) = c_0P_0(x) + c_1P_1(x) + ... + c_nP_n(x), \tag{26.10}$$

for some coefficients $c_0, ..., c_n$. For example,

$$p(x) = 3x^2 + 4x + 7 = 8P_0(x) + 4P_1(x) + 2P_2(x).$$

**Exercise 26.6** *Show that the choice of coefficients in Equation (26.10) for which the distance in Equation (26.9) is minimized is $c_m = a_m$, for the $a_n$ given in Equation (26.8).*

## 26.8 Legendre's Equations and Potential Theory

Potential theory is the name given to the study of Laplace's Equation, $\nabla^2 U = 0$, where $U = U(x, y, z, t)$ and the Laplacian operator is with respect to the spatial variables only. Steady-state solutions of the heat equation, $a^2\nabla^2 U = \frac{\partial U}{\partial t}$ satisfy Laplace's equation. An important problem in potential theory is to find a function $U$ satisfying Laplace's equation in the interior of a solid and taking specified values on its boundary. For example, take a sphere where the temperatures at each point on its surface do not change with time. Now find the steady-sate distribution of temperature inside the sphere.

It is natural to select coordinates that are easily related to the solid in question. Because the solid here is a sphere, spherical coordinates are the best choice. When the Laplacian is translated into spherical coordinates and Laplace's equation is solved by separation of variables, one of the equations that results is easily transformed into Legendre's equation.

## 26.9 Legendre Polynomials and Gaussian Quadrature

A *quadrature method* is a way of estimating the integral of a function from finitely many of its values. For example, the two-point trapezoidal method estimates the integral $\int_a^b f(x)dx$ as

$$\int_a^b f(x)dx \approx \frac{1}{2(b-a)}f(a) + \frac{1}{2(b-a)}f(b).$$

The Legendre polynomials play an important role in one such method, known as Gaussian Quadrature.

### 26.9.1 The Basic Formula

Suppose that we are given the points $(x_n, f(x_n))$, for $n = 1, 2, ..., N$, and we want to use these values to estimate the integral $\int_a^b f(x)dx$. One way is to use

$$\int_a^b f(x)dx \approx \sum_{n=1}^{N} c_n f(x_n). \tag{26.11}$$

If we select the $c_n$ so that the formula in Equation (26.11) is exact for the functions 1, $x$, ..., $x^{N-1}$, then the formula will provide the exact value of the integral for any polynomial $f(x)$ of degree less than $N$. Remarkably, we can do better than this if we are allowed to select the $x_n$ as well as the $c_n$.

### 26.9.2 Lagrange Interpolation

Let $x_n$, $n = 1, 2, ..., N$ be arbitrary points in $[a, b]$. Then the Lagrange polynomials $L_n(x)$, $n = 1, 2, ..., N$, are

$$L_n(x) = \prod_{m \neq n} \frac{(x - x_m)}{(x_n - x_m)}.$$

Then $L_n(x_n) = 1$ and $L_n(x_m) = 0$, for $m \neq n$. The polynomial

$$P(x) = \sum_{n=1}^{N} f(x_n) L_n(x)$$

interpolates $f(x)$ at the $N$ points $x_n$, since $P(x_n) = f(x_n)$ for $n = 1, 2, ..., N$.

### 26.9.3 Using the Legendre Polynomials

Let $N$ be given, and let $x_n$, $n = 1, 2, ..., N$ be the $N$ roots of the Legendre polynomial $P_N(x)$. We know that all these roots lie in the interval $[-1, 1]$. For each $n$ let $c_n = \int_{-1}^{1} L_n(x)dx$. Let $P(x)$ be any polynomial of degree less than $2N$. We show that

$$\int_{-1}^{1} P(x)dx = \sum_{n=1}^{N} c_n P(x_n);$$

that is, the quadrature method provides the correct answer, not just for polynomials of degree less than $N$, but for polynomials of degree less than $2N$.

Divide $P(x)$ by $P_N(x)$ to get

$$P(x) = Q(x)P_N(x) + R(x),$$

where both $Q(x)$ and $R(x)$ are polynomials of degree less than $N$. Then

$$\int_{-1}^{1} P(x)dx = \int_{-1}^{1} Q(x)P_N(x)dx + \int_{-1}^{1} R(x)dx = \int_{-1}^{1} R(x)dx,$$

since $P_N(x)$ is orthogonal to all polynomials of degree less than $N$. Since

$$\sum_{n=1}^{N} R(x_n) L_n(x)$$

is a polynomial of degree at most $N-1$ that interpolates $R(x)$ at $N$ points, we must have

$$\sum_{n=1}^{N} R(x_n)L_n(x) = R(x).$$

In addition,

$$P(x_n) = Q(x_n)P_N(x_n) + R(x_n) = R(x_n),$$

so that

$$\sum_{n=1}^{N} c_n R(x_n) = \int_{-1}^{1} R(x)dx = \int_{-1}^{1} P(x)dx = \sum_{n=1}^{N} c_n P(x_n).$$

# Chapter 27

# Hermite's Equations and Quantum Mechanics (Chapter 10,11)

## 27.1 The Schrödinger Wave Function

In quantum mechanics, the behavior of a particle with mass $m$ subject to a potential $V(x,t)$ satisfies the Schrödinger Equation

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = -\frac{\hbar}{2m}\frac{\partial^2\psi(x,t)}{\partial x^2} + V(x,t)\psi(x,t), \tag{27.1}$$

where $\hbar$ is Planck's constant. Here the $x$ is one-dimensional, but extensions to higher dimensions are also possible.

When the solution $\psi(x,t)$ is selected so that

$$|\psi(x,t)| \to 0,$$

as $|x| \to \infty$, and

$$\int_{-\infty}^{\infty} |\psi(x,t)|^2 dx = 1,$$

then, for each fixed $t$, the function $|\psi(x,t)|^2$ is a probability density function governing the position of the particle. In other words, the probability of finding the particle in the interval $[a,b]$ at time $t$ is

$$\int_a^b |\psi(x,t)|^2 dx.$$

An important special case is that of time-independent potentials.

## 27.2 Time-Independent Potentials

We say that $V(x,t)$ is time-independent if $V(x,t) = V(x)$, for all $t$. We then attempt to solve Equation (27.1) by separating the variables; we take $\psi(x,t) = f(t)g(x)$ and insert this product into Equation (27.1).

The time function is easily shown to be

$$f(t) = e^{-Et/\hbar},$$

where $E$ is defined to be the energy. The function $g(x)$ satisfies the *time-independent Schrödinger Equation*

$$-\frac{\hbar}{2m}g''(x) + V(x)g(x) = Eg(x). \tag{27.2}$$

An important special case is the harmonic oscillator.

## 27.3 The Harmonic Oscillator

The case of the *harmonic oscillator* corresponds to the potential $V(x) = \frac{1}{2}kx^2$.

### 27.3.1 The Classical Spring Problem

To motivate the development of the harmonic oscillator in quantum mechanics, it is helpful to recall the classical spring problem. In this problem a mass $m$ slides back and forth along a frictionless surface, with position $x(t)$ at time $t$. It is connected to a fixed structure by a spring with spring constant $k > 0$. The restoring force acting on the mass at any time is $-kx$, with $x = 0$ the equilibrium position of the mass. The equation of motion is

$$mx''(t) = -kx(t),$$

and the solution is

$$x(t) = x(0)\cos\sqrt{\frac{k}{m}}t.$$

The period of oscillation is $T = 2\pi\sqrt{\frac{m}{k}}$ and the frequency of oscillation is $\nu = \frac{1}{T} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, from which we obtain the equation

$$k = 4\pi^2 m\nu^2.$$

The potential energy is $\frac{1}{2}kx^2$, while the kinetic energy is $\frac{1}{2}m\dot{x}^2$. The sum of the kinetic and potential energies is the total energy, $E(t)$. Since $E'(t) = 0$, the energy is constant.

### 27.3.2 Back to the Harmonic Oscillator

When the potential function is $V(x) = \frac{1}{2}kx^2$, Equation (27.2) becomes

$$\frac{\hbar}{2m}g''(x) + (E - \frac{1}{2}kx^2)g(x) = 0, \tag{27.3}$$

where $k = m\omega^2$, for $\omega = 2\pi\nu$. With $u = \sqrt{\frac{m\omega}{\hbar}}$ and $\epsilon = \frac{2E}{\hbar\omega}$, we have

$$\frac{d^2g}{du^2} + (\epsilon - u^2)g = 0. \tag{27.4}$$

Equation (27.4) is equivalent to

$$w''(x) + (2p + 1 - x^2)w(x) = 0,$$

which can be transformed into Hermite's Equation

$$y'' - 2xy' + 2py = 0,$$

by writing $y(x) = w(x)e^{x^2/2}$.

In order for the solutions of Equation (27.3) to be physically admissible solutions, it is necessary that $p$ be a non-negative integer, which means that

$$E = \hbar\omega(n + \frac{1}{2}),$$

for some non-negative integer $n$; this gives the *quantized energy levels* for the harmonic oscillator.

## 27.4 Dirac's Equation

Einstein's theory of special relativity tells us that there are four variables, not just three, that have length for their units of measurement: the familiar three-dimensional spatial coordinates, and $ct$, where $c$ is the speed of light and $t$ is time. Looked at this way, Schrödinger's Equation (27.1), extended to three spatial dimensions, is peculiar, in that it treats the variable $ct$ differently from the others. There is only a first partial derivative in $t$, but second partial derivatives in the other variables. In 1930 the British mathematician Paul Dirac presented his relativistically correct version of Schrödinger's Equation.

Dirac's Equation, a version of which is inscribed on the wall of Westminster Abbey, is the following:

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar c}{i}\Big(\alpha_1\frac{\partial\psi}{\partial x_1} + \alpha_2\frac{\partial\psi}{\partial x_2} + \alpha_3\frac{\partial\psi}{\partial x_3}\Big) + \alpha_4 mc^2\psi. \tag{27.5}$$

Here the $\alpha_i$ are the Dirac matrices.

This equation agreed remarkably well with experimental data on the behavior of electrons in electric and magnetic fields, but it also seemed to allow for nonsensical solutions, such as spinning electrons with negative energy. The next year, Dirac realized that what the equation was calling for was *anti-matter*, a particle with the same mass as the electron, but with a positive charge. In the summer of 1932 Carl Anderson, working at Cal Tech, presented clear evidence for the existence of such a particle, which we now call the *positron*. What seemed like the height of science fiction in 1930 has become commonplace today.

When a positron collides with an electron their masses vanish and two gamma ray photons of pure energy are produced. These photons then move off in opposite directions. In positron emission tomography (PET) certain positron-emitting chemicals, such as glucose with radioactive fluorine chemically attached, are injected into the patient. When the PET scanner detects two photons arriving at the two ends of a line segment at (almost) the same time, called *coincidence detection*, it concludes that a positron was emitted somewhere along that line. This is repeated thousands of times. Once all this data has been collected, the mathematicians take over and use these clues to reconstruct an image of where the glucose is in the body. It is this image that the doctor sees.

# Chapter 28

# Array Processing (Chapter 8)

In radar and sonar, the field $u(\mathbf{s}, t)$ being sampled is usually viewed as a discrete or continuous superposition of planewave solutions with various amplitudes, frequencies, and wavevectors. We sample the field at various spatial locations $\mathbf{s}_m$, $m = 1, ..., M$, for $t$ in some finite interval of time. We simplify the situation a bit now by assuming that all the planewave solutions are associated with the same frequency, $\omega$. If not, we perform an FFT on the functions of time received at each sensor location $\mathbf{s}_m$ and keep only the value associated with the desired frequency $\omega$.

In the continuous superposition model, the field is

$$u(\mathbf{s}, t) = e^{i\omega t} \int f(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{s}} d\mathbf{k}.$$

Our measurements at the sensor locations $\mathbf{s}_m$ give us the values

$$F(\mathbf{s}_m) = \int f(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{s}_m} d\mathbf{k},$$

for $m = 1, ..., M$. The data are then Fourier transform values of the complex function $f(\mathbf{k})$; $f(\mathbf{k})$ is defined for all three-dimensional real vectors $\mathbf{k}$, but is zero, in theory, at least, for those $\mathbf{k}$ whose squared length $||\mathbf{k}||^2$ is not equal to $\omega^2/c^2$. Our goal is then to estimate $f(\mathbf{k})$ from finitely many values of its Fourier transform. Since each $\mathbf{k}$ is a normal vector for its planewave field component, determining the value of $f(\mathbf{k})$ will tell us the strength of the planewave component coming from the direction $\mathbf{k}$.

The collection of sensors at the spatial locations $\mathbf{s}_m$, $m = 1, ..., M$, is called *an array*, and the size of the array, in units of the wavelength $\lambda = 2\pi c/\omega$, is called the *aperture* of the array. Generally, the larger the

aperture the better, but what is a large aperture for one value of $\omega$ will be a smaller aperture for a lower frequency.

In some applications the sensor locations are essentially arbitrary, while in others their locations are carefully chosen. Sometimes, the sensors are collinear, as in sonar towed arrays. Let's look more closely at the collinear case.

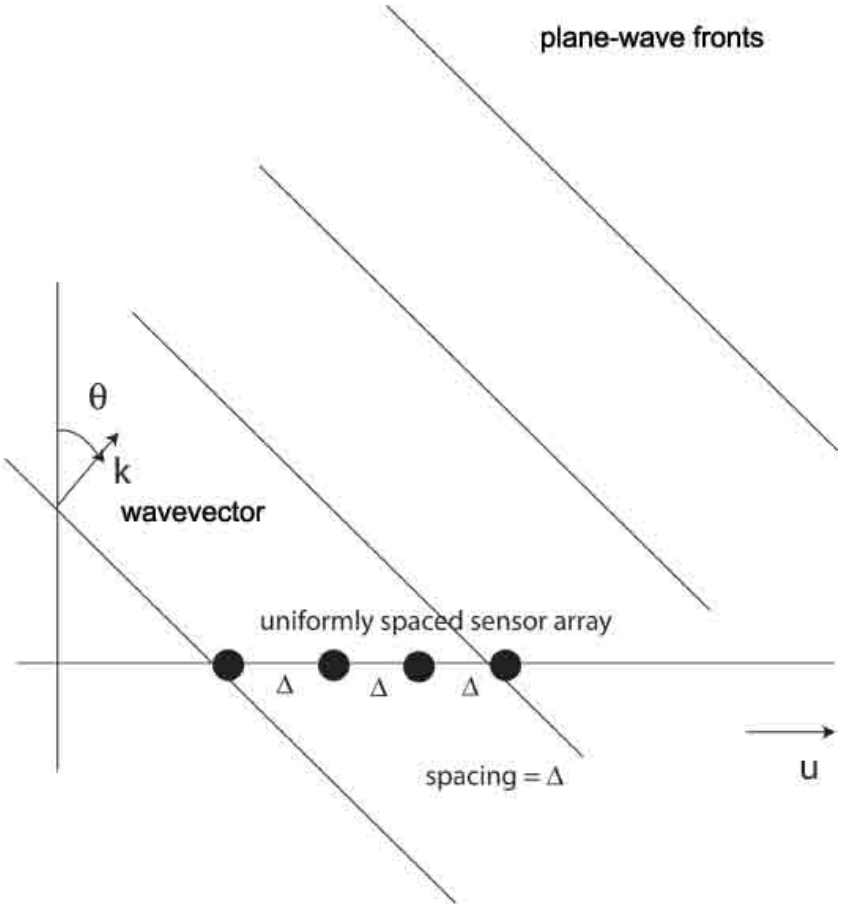



Figure 28.1: A uniform line array sensing a planewave field.

We assume now that the sensors are equispaced along the $x$-axis, at locations $(m\Delta, 0, 0)$, $m = 1, ..., M$, where $\Delta > 0$ is the sensor spacing; such an arrangement is called a *uniform line array*. This setup is illustrated in

Figure 28.1. Our data is then

$$F_m = F(\mathbf{s}_m) = F((m\Delta, 0, 0)) = \int f(\mathbf{k}) e^{im\Delta \mathbf{k}\cdot(1,0,0)} d\mathbf{k}.$$

Since $\mathbf{k} \cdot (1, 0, 0) = \frac{\omega}{c} \cos\theta$, for $\theta$ the angle between the vector $\mathbf{k}$ and the $x$-axis, we see that there is some ambiguity now; we cannot distinguish the cone of vectors that have the same $\theta$. It is common then to assume that the wavevectors $\mathbf{k}$ have no $z$-component and that $\theta$ is the angle between two vectors in the $x, y$-plane, the so-called *angle of arrival*. The *wavenumber* variable $k = \frac{\omega}{c} \cos\theta$ lies in the interval $[-\frac{\omega}{c}, \frac{\omega}{c}]$, and we imagine that $f(\mathbf{k})$ is now $f(k)$, defined for $|k| \leq \frac{\omega}{c}$. The Fourier transform of $f(k)$ is $F(s)$, a function of a single real variable $s$. Our data is then viewed as the values $F(m\Delta)$, for $m = 1, ..., M$. Since the function $f(k)$ is zero for $|k| > \frac{\omega}{c}$, the Nyquist spacing in $s$ is $\frac{\pi c}{\omega}$, which is $\frac{\lambda}{2}$, where $\lambda = \frac{2\pi c}{\omega}$ is the wavelength.

To avoid aliasing, which now means mistaking one direction of arrival for another, we need to select $\Delta \leq \frac{\lambda}{2}$. When we have oversampled, so that $\Delta < \frac{\lambda}{2}$, the interval $[-\frac{\omega}{c}, \frac{\omega}{c}]$, the so-called *visible region*, is strictly smaller than the interval $[-\frac{\pi}{\Delta}, \frac{\pi}{\Delta}]$. If the model of propagation is accurate, all the signal component planewaves will correspond to wavenumbers $k$ in the visible region and the background noise will also appear as a superposition of such propagating planewaves. In practice, there can be components in the noise that appear to come from wavenumbers $k$ outside of the visible region; this means these components of the noise are not due to distant sources propagating as planewaves, but, perhaps, to sources that are in the *near field*, or localized around individual sensors, or coming from the electronics within the sensors.

Using the relation $\lambda\omega = 2\pi c$, we can calculate the Nyquist spacing for any particular case of planewave array processing. For electromagnetic waves the propagation speed is the speed of light, which we shall take here to be $c = 3 \times 10^8$ meters per second. The wavelength $\lambda$ for gamma rays is around one Angstrom, which is $10^{-10}$ meters; for x-rays it is about one millimicron, or $10^{-9}$ meters. The visible spectrum has wavelengths that are a little less than one micron, that is, $10^{-6}$ meters. Shortwave radio has wavelength around one millimeter; broadcast radio has a $\lambda$ running from about 10 meters to 1000 meters, while the so-called long radio waves can have wavelengths several thousand meters long. At the one extreme it is impractical (if not physically impossible) to place individual sensors at the Nyquist spacing of fractions of microns, while at the other end, managing to place the sensors far enough apart is the challenge.

The wavelengths used in primitive early radar at the start of World War II were several meters long. Since resolution is proportional to aperture, which, in turn, is the length of the array, in units of wavelength, antennae for such radar needed to be quite large. The general feeling at the time was that the side with the shortest wavelength would win the war. The cavity

magnetron, invented during the war by British scientists, made possible 10 cm wavelength radar, which could then easily be mounted on planes.

In ocean acoustics it is usually assumed that the speed of propagation of sound is around 1500 meters per second, although deviations from this *ambient sound speed* are significant and since they are caused by such things as temperature differences in the ocean, can be used to estimate these differences. At around the frequency $\omega = 50$ Hz, we find sound generated by man-made machinery, such as motors in vessels, with higher frequency harmonics sometimes present also; at other frequencies the main sources of acoustic energy may be wind-driven waves or whales. The wavelength for 50 Hz is $\lambda = 30$ meters; sonar will typically operate both above and below this wavelength. It is sometimes the case that the array of sensors is fixed in place, so what may be Nyquist spacing for 50 Hz will be oversampling for 20 Hz.

We have focused here exclusively on planewave propagation, which results when the source is far enough way from the sensors and the speed of propagation is constant. In many important applications these conditions are violated, and different versions of the wave equation are needed, which have different solutions. For example, sonar signal processing in environments such as shallow channels, in which some of the sound reaches the sensors only after interacting with the ocean floor or the surface, requires more complicated parameterized models for solutions of the appropriate wave equation. Lack of information about the depth and nature of the bottom can also cause errors in the signal processing. In some cases it is possible to use acoustic energy from known sources to determine the needed information.

Array signal processing can be done in *passive* or *active* mode. In passive mode the energy is either reflected off of or originates at the object of interest: the moon reflects sunlight, while ships generate their own noise. In the active mode the object of interest does not generate or reflect enough energy by itself, so the energy is generated by the party doing the sensing: active sonar is sometimes used to locate quiet vessels, while radar is used to locate planes in the sky or to map the surface of the earth. In the February 2003 issue of *Harper's Magazine* there is an article on scientific apocalypse, dealing with the search for near-earth asteroids. These objects are initially detected by passive optical observation, as small dots of reflected sunlight; once detected, they are then imaged by active radar to determine their size, shape, rotation and such.

# Chapter 29

# Matched Field Processing (Chapter 10,11,12)

Previously we considered the array processing problem in the context of planewave propagation. When the environment is more complicated, the wave equation must be modified to reflect the physics of the situation and the signal processing modified to incorporate that physics. A good example of such modification is provided by acoustic signal processing in shallow water, the topic of this chapter.

## 29.1 The Shallow-Water Case

In the shallow-water situation the acoustic energy from the source interacts with the surface and with the bottom of the channel, prior to being received by the sensors. The nature of this interaction is described by the wave equation in cylindrical coordinates. The deviation from the ambient pressure is the function $p(t, \mathbf{s}) = p(t, r, z, \theta)$, where $\mathbf{s} = (r, z, \theta)$ is the spatial vector variable, $r$ is the range, $z$ the depth, and $\theta$ the bearing angle in the horizontal. We assume a single frequency, $\omega$, so that

$$p(t, \mathbf{s}) = e^{i\omega t} g(r, z, \theta).$$

We shall assume cylindrical symmetry to remove the $\theta$ dependence; in many applications the bearing is essentially known or limited by the environment or can be determined by other means. The sensors are usually positioned in a vertical array in the channel, with the top of the array taken to be the origin of the coordinate system and positive $z$ taken to mean positive depth below the surface. We shall also assume that there is a single source of acoustic energy located at range $r_s$ and depth $z_s$.

To simplify a bit, we assume here that the sound speed $c = c(z)$ does not change with range, but only with depth, and that the channel has constant depth and density. Then, the Helmholtz equation for the function $g(r, z)$ is

$$\nabla^2 g(r, z) + [\omega/c(z)]^2 g(r, z) = 0.$$

The Laplacian is

$$\nabla^2 g(r, z) = g_{rr}(r, z) + \frac{1}{r} g_r(r, z) + g_{zz}(r, z).$$

We separate the variables once again, writing

$$g(r, z) = f(r)u(z).$$

Then, the range function $f(r)$ must satisfy the differential equation

$$f''(r) + \frac{1}{r} f'(r) = -\alpha f(r),$$

and the depth function $u(z)$ satisfies the differential equation

$$u''(z) + k(z)^2 u(z) = \alpha u(z),$$

where $\alpha$ is a separation constant and

$$k(z)^2 = [\omega/c(z)]^2.$$

Taking $\lambda^2 = \alpha$, the range equation becomes

$$f''(r) + \frac{1}{r} f'(r) + \lambda^2 f(r) = 0,$$

which is Bessel's equation, with Hankel-function solutions. The depth equation becomes

$$u''(z) + (k(z)^2 - \lambda^2)u(z) = 0,$$

which is of Sturm-Liouville type. The boundary conditions pertaining to the surface and the channel bottom will determine the values of $\lambda$ for which a solution exists.

To illustrate the way in which the boundary conditions become involved, we consider two examples.

## 29.2 The Homogeneous-Layer Model

We assume now that the channel consists of a single homogeneous layer of water of constant density, constant depth $d$, and constant sound speed $c$. We impose the following boundary conditions:

a. Pressure-release surface: $u(0) = 0$.

b. Rigid bottom: $u'(d) = 0$.

With $\gamma^2 = (k^2 - \lambda^2)$, we get $\cos(\gamma d) = 0$, so the permissible values of $\lambda$ are

$$\lambda_m = (k^2 - [(2m-1)\pi/2d]^2)^{1/2},\ m = 1, 2, ....$$

The normalized solutions of the depth equation are now

$$u_m(z) = \sqrt{2/d}\sin(\gamma_m z),$$

where

$$\gamma_m = \sqrt{k^2 - \lambda_m^2} = (2m-1)\pi/2d,\ m = 1, 2, ....$$

For each $m$ the corresponding function of the range satisfies the differential equation

$$f''(r) + \frac{1}{r}f'(r) + \lambda_m^2 f(r),$$

which has solution $H_0^{(1)}(\lambda_m r)$, where $H_0^{(1)}$ is the zeroth order Hankel-function solution of Bessel's equation. The asymptotic form for this function is

$$\pi i H_0^{(1)}(\lambda_m r) = \sqrt{2\pi/\lambda_m r}\exp(-i(\lambda_m r + \frac{\pi}{4})).$$

It is this asymptotic form that is used in practice. Note that when $\lambda_m$ is complex with a negative imaginary part, there will be a decaying exponential in this solution, so this term will be omitted in the signal processing.

Having found the range and depth functions, we write $g(r, z)$ as a superposition of these elementary products, called the *modes*:

$$g(r, z) = \sum_{m=1}^{M} A_m H_0^{(1)}(\lambda_m r) u_m(z),$$

where $M$ is the number of propagating modes free of decaying exponentials. The $A_m$ can be found from the original Helmholtz equation; they are

$$A_m = (i/4)u_m(z_s),$$

where $z_s$ is the depth of the source of the acoustic energy. Notice that the depth of the source also determines the strength of each mode in this superposition; this is described by saying that the source has *excited* certain modes and not others.

The eigenvalues $\lambda_m$ of the depth equation will be complex when

$$k = \frac{\omega}{c} < \frac{(2m-1)\pi}{2d}.$$

If $\omega$ is below the *cut-off frequency* $\frac{\pi c}{2d}$, then all the $\lambda_m$ are complex and there are no propagating modes ($M = 0$). The number of propagating modes is

$$M = \frac{1}{2} + \frac{\omega d}{\pi c},$$

which is $\frac{1}{2}$ plus the depth of the channel in units of half-wavelengths.

This model for shallow-water propagation is helpful in revealing a number of the important aspects of modal propagation, but is of limited practical utility. A more useful and realistic model is the *Pekeris waveguide*.

## 29.3 The Pekeris Waveguide

Now we assume that the water column has constant depth $d$, sound speed $c$, and density $b$. Beneath the water is an infinite half-space with sound speed $c' > c$, and density $b'$. Figure 29.1 illustrates the situation.

Using the new depth variable $v = \frac{\omega z}{c}$, the depth equation becomes

$$u''(v) + \lambda^2 u(v) = 0, \text{ for } 0 \le v \le \frac{\omega d}{c},$$

and

$$u''(v) + ((\frac{c}{c'})^2 - 1 + \lambda^2)u(v) = 0, \text{ for } \frac{\omega d}{c} < v.$$

To have a solution, $\lambda$ must satisfy the equation

$$\tan(\lambda \omega d/c) = -(\lambda b/b')/\sqrt{1 - (\frac{c}{c'})^2 - \lambda^2},$$

with

$$1 - (\frac{c}{c'})^2 - \lambda^2 \ge 0.$$

The *trapped modes* are those whose corresponding $\lambda$ satisfies

$$1 \ge 1 - \lambda^2 \ge (\frac{c}{c'})^2.$$

The eigenfunctions are

$$u_m(v) = \sin(\lambda_m v), \text{ for } 0 \le v \le \frac{\omega d}{c}$$

and

$$u_m(v) = \exp\left( - v\sqrt{1 - (\frac{c}{c'})^2 - \lambda^2}\right), \text{ for } \frac{\omega d}{c} < v.$$

Although the Pekeris model has its uses, it still may not be realistic enough in some cases and more complicated propagation models will be needed.

## 29.4 The General Normal-Mode Model

Regardless of the model by which the modal functions are determined, the general *normal-mode expansion* for the range-independent case is

$$g(r,z) = \sum_{m=1}^{M} u_m(z) s_m(r, z_s),$$

where $M$ is the number of propagating modes and $s_m(r, z_s)$ is the *modal amplitude* containing all the information about the source of the sound.

### 29.4.1 Matched-Field Processing

In planewave array processing we write the acoustic field as a superposition of planewave fields and try to find the corresponding amplitudes. This can be done using a matched filter, although high-resolution methods can also be used. In the matched-filter approach, we fix a wavevector and then match the data with the vector that describes what we would have received at the sensors had there been but a single planewave present corresponding to that fixed wavevector; we then repeat for other fixed wavevectors. In more complicated acoustic environments, such as normal-mode propagation in shallow water, we write the acoustic field as a superposition of fields due to sources of acoustic energy at individual points in range and depth and then seek the corresponding amplitudes. Once again, this can be done using a matched filter.

In matched-field processing we fix a particular range and depth and compute what we would have received at the sensors had the acoustic field been generated solely by a single source at that location. We then match the data with this computed vector. We repeat this process for many different choices of range and depth, obtaining a function of $r$ and $z$ showing the likely locations of actual sources. As in the planewave case, high-resolution nonlinear methods can also be used.

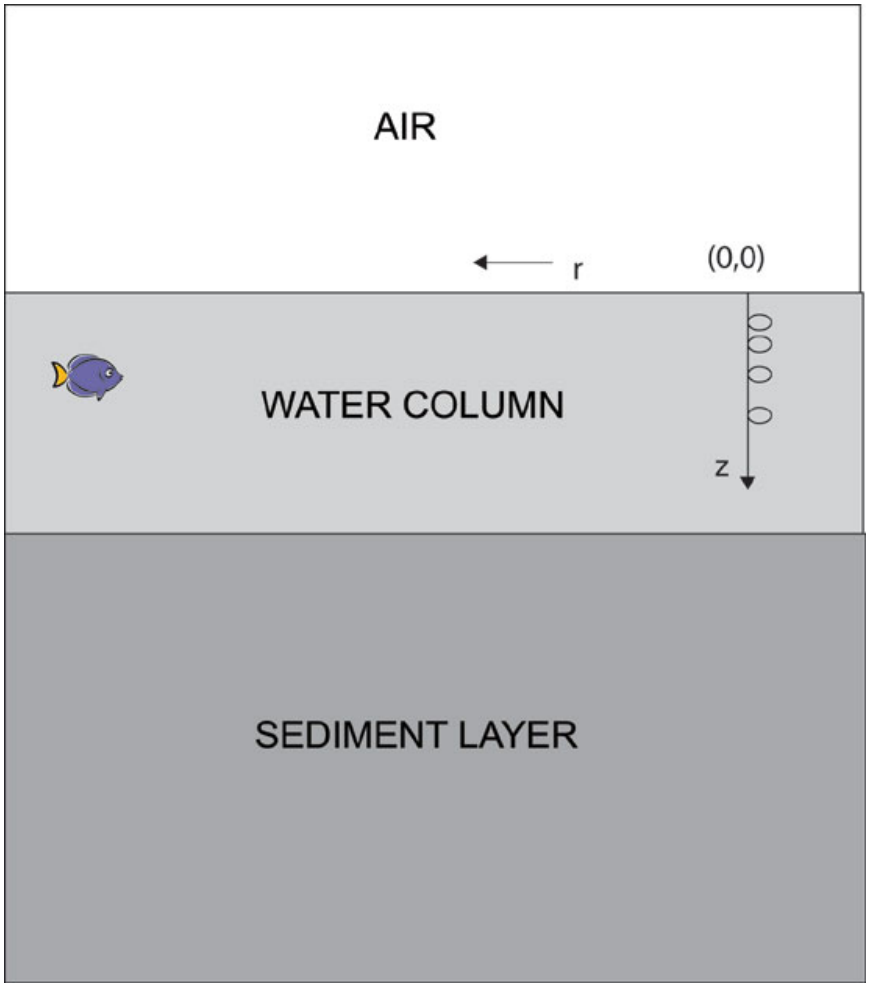



Figure 29.1: The Pekeris Model.

# Part III

# Appendices

# Chapter 30

# Inner Products and Orthogonality

## 30.1 The Complex Vector Dot Product

An *inner product* is a generalization of the notion of the dot product between two complex vectors.

### 30.1.1 The Two-Dimensional Case

Let $\mathbf{u} = (a, b)$ and $\mathbf{v} = (c, d)$ be two vectors in two-dimensional space. Let $\mathbf{u}$ make the angle $\alpha > 0$ with the positive $x$-axis and $\mathbf{v}$ the angle $\beta > 0$. Let $||\mathbf{u}|| = \sqrt{a^2 + b^2}$ denote the length of the vector $\mathbf{u}$. Then $a = ||\mathbf{u}|| \cos\alpha$, $b = ||\mathbf{u}|| \sin\alpha$, $c = ||\mathbf{v}|| \cos\beta$ and $d = ||\mathbf{v}|| \sin\beta$. So $\mathbf{u} \cdot \mathbf{v} = ac + bd = ||\mathbf{u}||||\mathbf{v}||(\cos\alpha\cos\beta + \sin\alpha\sin\beta = ||\mathbf{u}||\,||\mathbf{v}||\cos(\alpha - \beta)$. Therefore, we have

$$\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}||\,||\mathbf{v}||\cos\theta, \tag{30.1}$$

where $\theta = \alpha - \beta$ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Cauchy's inequality is

$$|\mathbf{u} \cdot \mathbf{v}| \leq ||\mathbf{u}||\,||\mathbf{v}||,$$

with equality if and only if $\mathbf{u}$ and $\mathbf{v}$ are parallel. From Equation (30.1) we know that the dot product $\mathbf{u} \cdot \mathbf{v}$ is zero if and only if the angle between these two vectors is a right angle; we say then that $\mathbf{u}$ and $\mathbf{v}$ are mutually *orthogonal*.

Cauchy's inequality extends to complex vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u} \cdot \mathbf{v} = \sum_{n=1}^{N} u_n \overline{v_n}, \tag{30.2}$$

and Cauchy's Inequality still holds.

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

$$0 \leq ||e^{-i\theta}\mathbf{u} - t\mathbf{v}||^2 = (e^{-i\theta}\mathbf{u} - t\mathbf{v}) \cdot (e^{-i\theta}\mathbf{u} - t\mathbf{v})$$

$$= ||\mathbf{u}||^2 - t[(e^{-i\theta}\mathbf{u}) \cdot \mathbf{v} + \mathbf{v} \cdot (e^{-i\theta}\mathbf{u})] + t^2||\mathbf{v}||^2$$

$$= ||\mathbf{u}||^2 - t[(e^{-i\theta}\mathbf{u}) \cdot \mathbf{v} + \overline{(e^{-i\theta}\mathbf{u}) \cdot \mathbf{v}}] + t^2||\mathbf{v}||^2$$

$$= ||\mathbf{u}||^2 - 2Re(te^{-i\theta}(\mathbf{u} \cdot \mathbf{v})) + t^2||\mathbf{v}||^2$$

$$= ||\mathbf{u}||^2 - 2Re(t|\mathbf{u} \cdot \mathbf{v}|) + t^2||\mathbf{v}||^2 = ||\mathbf{u}||^2 - 2t|\mathbf{u} \cdot \mathbf{v}| + t^2||\mathbf{v}||^2.$$

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

A careful examination of the proof just presented shows that we did not explicitly use the definition of the complex vector dot product, but only some of its properties. This suggested to mathematicians the possibility of abstracting these properties and using them to define a more general concept, an *inner product*, between objects more general than complex vectors, such as infinite sequences, random variables, and matrices. Such an inner product can then be used to define the *norm* of these objects and thereby a distance between such objects. Once we have an inner product defined, we also have available the notions of orthogonality and best approximation.

### 30.1.2 Orthogonality

Consider the problem of writing the two-dimensional real vector $(3, -2)$ as a linear combination of the vectors $(1, 1)$ and $(1, -1)$; that is, we want to find constants $a$ and $b$ so that $(3, -2) = a(1, 1) + b(1, -1)$. One way to do this, of course, is to compare the components: $3 = a + b$ and $-2 = a - b$; we can then solve this simple system for the $a$ and $b$. In higher dimensions this way of doing it becomes harder, however. A second way is to make use of the dot product and orthogonality.

The dot product of two vectors $(x, y)$ and $(w, z)$ in $R^2$ is $(x, y) \cdot (w, z) = xw+yz$. If the dot product is zero then the vectors are said to be *orthogonal*; the two vectors $(1, 1)$ and $(1, -1)$ are orthogonal. We take the dot product of both sides of $(3, -2) = a(1, 1) + b(1, -1)$ with $(1, 1)$ to get

$$1 = (3, -2) \cdot (1, 1) = a(1, 1) \cdot (1, 1) + b(1, -1) \cdot (1, 1) = a(1, 1) \cdot (1, 1) + 0 = 2a,$$

so we see that $a = \frac{1}{2}$. Similarly, taking the dot product of both sides with $(1, -1)$ gives

$$5 = (3, -2) \cdot (1, -1) = a(1, 1) \cdot (1, -1) + b(1, -1) \cdot (1, -1) = 2b,$$

so $b = \frac{5}{2}$. Therefore, $(3, -2) = \frac{1}{2}(1, 1) + \frac{5}{2}(1, -1)$. The beauty of this approach is that it does not get much harder as we go to higher dimensions.

Since the cosine of the angle $\theta$ between vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\cos\theta = \mathbf{u} \cdot \mathbf{v}/||\mathbf{u}||\,||\mathbf{v}||,$$

where $||\mathbf{u}||^2 = \mathbf{u} \cdot \mathbf{u}$, the projection of vector $\mathbf{v}$ on to the line through the origin parallel to $\mathbf{u}$ is

$$\text{Proj}_{\mathbf{u}}(\mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}}\mathbf{u}.$$

Therefore, the vector $\mathbf{v}$ can be written as

$$\mathbf{v} = \text{Proj}_{\mathbf{u}}(\mathbf{v}) + (\mathbf{v} - \text{Proj}_{\mathbf{u}}(\mathbf{v})),$$

where the first term on the right is parallel to $\mathbf{u}$ and the second one is orthogonal to $\mathbf{u}$.

How do we find vectors that are mutually orthogonal? Suppose we begin with $(1, 1)$. Take a second vector, say $(1, 2)$, that is not parallel to $(1, 1)$ and write it as we did $\mathbf{v}$ earlier, that is, as a sum of two vectors, one parallel to $(1, 1)$ and the second orthogonal to $(1, 1)$. The projection of $(1, 2)$ onto the line parallel to $(1, 1)$ passing through the origin is

$$\frac{(1, 1) \cdot (1, 2)}{(1, 1) \cdot (1, 1)}(1, 1) = \frac{3}{2}(1, 1) = (\frac{3}{2}, \frac{3}{2})$$

so

$$(1, 2) = (\frac{3}{2}, \frac{3}{2}) + ((1, 2) - (\frac{3}{2}, \frac{3}{2})) = (\frac{3}{2}, \frac{3}{2}) + (-\frac{1}{2}, \frac{1}{2}).$$

The vectors $(-\frac{1}{2}, \frac{1}{2}) = -\frac{1}{2}(1, -1)$ and, therefore, $(1, -1)$ are then orthogonal to $(1, 1)$. This approach is the basis for the *Gram-Schmidt* method for constructing a set of mutually orthogonal vectors.

## 30.2 Generalizing the Dot Product: Inner Products

The proof of Cauchy's Inequality rests not on the actual definition of the complex vector dot product, but rather on four of its most basic properties. We use these properties to extend the concept of the complex vector dot product to that of *inner product*. Later in this chapter we shall give several examples of inner products, applied to a variety of mathematical objects,

including infinite sequences, functions, random variables, and matrices. For now, let us denote our mathematical objects by $\mathbf{u}$ and $\mathbf{v}$ and the inner product between them as $\langle \mathbf{u}, \mathbf{v} \rangle$. The objects will then be said to be members of an *inner-product space*. We are interested in inner products because they provide a notion of orthogonality, which is fundamental to best approximation and optimal estimation.

### 30.2.1 Defining an Inner Product and Norm

The four basic properties that will serve to define an inner product are:

- **1:** $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{0}$;
- **2:** $\langle \mathbf{v}, \mathbf{u} \rangle = \overline{\langle \mathbf{u}, \mathbf{v} \rangle}$;
- **3:** $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$;
- **4:** $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$ for any complex number $c$.

The inner product is the basic ingredient in Hilbert space theory. Using the inner product, we define the *norm* of $\mathbf{u}$ to be

$$||\mathbf{u}|| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$$

and the distance between $\mathbf{u}$ and $\mathbf{v}$ to be $||\mathbf{u} - \mathbf{v}||$.

**The Cauchy-Schwarz inequality:** Because these four properties were all we needed to prove the Cauchy inequality for the complex vector dot product, we obtain the same inequality whenever we have an inner product. This more general inequality is the Cauchy-Schwarz inequality:

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

or

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq ||\mathbf{u}|| \, ||\mathbf{v}||,$$

with equality if and only if there is a scalar $c$ such that $\mathbf{v} = c\mathbf{u}$. We say that the vectors $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. We turn now to some examples.

### 30.2.2 Some Examples of Inner Products

Here are several examples of inner products.

- **Inner product of infinite sequences:** Let $\mathbf{u} = \{u_n\}$ and $\mathbf{v} = \{v_n\}$ be infinite sequences of complex numbers. The inner product is then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum u_n \overline{v_n},$$

and

$$||\mathbf{u}|| = \sqrt{\sum |u_n|^2}.$$

The sums are assumed to be finite; the index of summation $n$ is singly or doubly infinite, depending on the context. The Cauchy-Schwarz inequality says that

$$|\sum u_n \overline{v_n}| \leq \sqrt{\sum |u_n|^2} \sqrt{\sum |v_n|^2}.$$

- **Inner product of functions:** Now suppose that $\mathbf{u} = f(x)$ and $\mathbf{v} = g(x)$. Then,

$$\langle \mathbf{u}, \mathbf{v} \rangle = \int f(x) \overline{g(x)} dx$$

and

$$||\mathbf{u}|| = \sqrt{\int |f(x)|^2 dx}.$$

The integrals are assumed to be finite; the limits of integration depend on the support of the functions involved. The Cauchy-Schwarz inequality now says that

$$|\int f(x) \overline{g(x)} dx| \leq \sqrt{\int |f(x)|^2 dx} \sqrt{\int |g(x)|^2 dx}.$$

- **Inner product of random variables:** Now suppose that $\mathbf{u} = X$ and $\mathbf{v} = Y$ are random variables. Then,

$$\langle \mathbf{u}, \mathbf{v} \rangle = E(X\overline{Y})$$

and

$$||\mathbf{u}|| = \sqrt{E(|X|^2)},$$

which is the standard deviation of $X$ if the mean of $X$ is zero. The expected values are assumed to be finite. The Cauchy-Schwarz inequality now says that

$$|E(X\overline{Y})| \leq \sqrt{E(|X|^2)} \sqrt{E(|Y|^2)}.$$

If $E(X) = 0$ and $E(Y) = 0$, the random variables $X$ and $Y$ are orthogonal if and only if they are *uncorrelated*.

- **Inner product of complex matrices:** Now suppose that $\mathbf{u} = A$ and $\mathbf{v} = B$ are complex matrices. Then,

$$\langle \mathbf{u}, \mathbf{v} \rangle = \text{trace}(B^\dagger A)$$

and

$$||\mathbf{u}|| = \sqrt{\text{trace}(A^\dagger A)},$$

where the trace of a square matrix is the sum of the entries on the main diagonal. As we shall see later, this inner product is simply the complex vector dot product of the vectorized versions of the matrices involved. The Cauchy-Schwarz inequality now says that

$$|\text{trace}(B^\dagger A)| \leq \sqrt{\text{trace}(A^\dagger A)}\sqrt{\text{trace}(B^\dagger B)}.$$

- **Weighted inner product of complex vectors:** Let $\mathbf{u}$ and $\mathbf{v}$ be complex vectors and let $Q$ be a Hermitian positive-definite matrix; that is, $Q^\dagger = Q$ and $\mathbf{u}^\dagger Q\mathbf{u} > 0$ for all nonzero vectors $\mathbf{u}$. The inner product is then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{v}^\dagger Q\mathbf{u}$$

and

$$||\mathbf{u}|| = \sqrt{\mathbf{u}^\dagger Q\mathbf{u}}.$$

We know from the eigenvector decomposition of $Q$ that $Q = C^\dagger C$ for some matrix $C$. Therefore, the inner product is simply the complex vector dot product of the vectors $C\mathbf{u}$ and $C\mathbf{v}$. The Cauchy-Schwarz inequality says that

$$|\mathbf{v}^\dagger Q\mathbf{u}| \leq \sqrt{\mathbf{u}^\dagger Q\mathbf{u}}\sqrt{\mathbf{v}^\dagger Q\mathbf{v}}.$$

- **Weighted inner product of functions:** Now suppose that $\mathbf{u} = f(x)$ and $\mathbf{v} = g(x)$ and $w(x) > 0$. Then define

$$\langle \mathbf{u}, \mathbf{v} \rangle = \int f(x)\overline{g(x)}w(x)dx$$

and

$$||\mathbf{u}|| = \sqrt{\int |f(x)|^2 w(x)dx}.$$

The integrals are assumed to be finite; the limits of integration depend on the support of the functions involved. This inner product is simply the inner product of the functions $f(x)\sqrt{w(x)}$ and $g(x)\sqrt{w(x)}$. The Cauchy-Schwarz inequality now says that

$$|\int f(x)\overline{g(x)}w(x)dx| \leq \sqrt{\int |f(x)|^2 w(x)dx}\sqrt{\int |g(x)|^2 w(x)dx}.$$

Once we have an inner product defined, we can speak about orthogonality and best approximation. Important in that regard is the orthogonality principle.

## 30.3 Best Approximation and the Orthogonality Principle

Imagine that you are standing and looking down at the floor. The point $B$ on the floor that is closest to $N$, the tip of your nose, is the unique point on the floor such that the vector from $B$ to any other point $A$ on the floor is perpendicular to the vector from $N$ to $B$; that is, $\langle BN, BA\rangle = 0$. This is a simple illustration of the *orthogonality principle*. Whenever we have an inner product defined we can speak of orthogonality and apply the orthogonality principle to find best approximations. For notational simplicity, we shall consider only real inner product spaces.

### 30.3.1 Best Approximation

Let $\mathbf{u}$ and $\mathbf{v}^1, ..., \mathbf{v}^N$ be members of a real inner-product space. For all choices of scalars $a_1, ..., a_N$, we can compute the distance from $\mathbf{u}$ to the member $a_1\mathbf{v}^1 + ...a_N\mathbf{v}^N$. Then, we minimize this distance over all choices of the scalars; let $b_1, ..., b_N$ be this best choice.

The distance squared from $\mathbf{u}$ to $a_1\mathbf{v}^1 + ...a_N\mathbf{v}^N$ is

$$||\mathbf{u} - (a_1\mathbf{v}^1 + ...a_N\mathbf{v}^N)||^2 = \langle \mathbf{u} - (a_1\mathbf{v}^1 + ...a_N\mathbf{v}^N), \mathbf{u} - (a_1\mathbf{v}^1 + ...a_N\mathbf{v}^N)\rangle,$$

$$= ||\mathbf{u}||^2 - 2\langle \mathbf{u}, \sum_{n=1}^{N} a_n\mathbf{v}^n\rangle + \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m \langle \mathbf{v}^n, \mathbf{v}^m\rangle.$$

Setting the partial derivative with respect to $a_n$ equal to zero, we have

$$\langle \mathbf{u}, \mathbf{v}^n\rangle = \sum_{m=1}^{N} b_m \langle \mathbf{v}^m, \mathbf{v}^n\rangle.$$

With $\mathbf{b} = (b_1, ..., b_N)^T$,

$$\mathbf{d} = (\langle \mathbf{u}, \mathbf{v}^1\rangle, ..., \langle \mathbf{u}, \mathbf{v}^N\rangle)^T$$

and $V$ the matrix with entries

$$V_{mn} = \langle \mathbf{v}^m, \mathbf{v}^n\rangle,$$

we find that we must solve the system of equations $V\mathbf{b} = \mathbf{d}$. When the vectors $\mathbf{v}^n$ are mutually orthogonal and each has norm equal to one, then $V = I$, the identity matrix, and the desired vector $\mathbf{b}$ is simply $\mathbf{d}$.

### 30.3.2 The Orthogonality Principle

The *orthogonality principle* provides another way to view the calculation of the best approximation: let the best approximation of $\mathbf{u}$ be the vector

$$b_1\mathbf{v}^1 + ...b_N\mathbf{v}^N.$$

Then

$$\langle \mathbf{u} - (b_1\mathbf{v}^1 + ... + b_N\mathbf{v}^N), \mathbf{v}^n \rangle = 0,$$

for $n = 1, 2, ..., N$. This leads directly to the system of equations

$$\mathbf{d} = V\mathbf{b},$$

which, as we just saw, provides the optimal coefficients.

To see why the orthogonality principle is valid, fix a value of $n$ and consider the problem of minimizing the distance

$$||\mathbf{u} - (b_1\mathbf{v}^1 + ...b_N\mathbf{v}^N + \alpha\mathbf{v}^n)||$$

as a function of $\alpha$. Writing the norm squared in terms of the inner product, expanding the terms, and differentiating with respect to $\alpha$, we find that the minimum occurs when

$$\alpha = \langle \mathbf{u} - (b_1\mathbf{v}^1 + ... + b_N\mathbf{v}^N), \mathbf{v}^n \rangle.$$

But we already know that the minimum occurs when $\alpha = 0$. This completes the proof of the orthogonality principle.

## 30.4 Gram-Schmidt Orthogonalization

We have seen that the best approximation is easily calculated if the vectors $\mathbf{v}^n$ are mutually orthogonal. But how to we get such a mutually orthogonal set, in general? The Gram-Schmidt Orthogonalization Method is one way to proceed.

Let $\{\mathbf{v}^1, ..., \mathbf{v}^N\}$ be a linearly independent set of vectors in the space $R^M$, where $N \leq M$. The Gram-Schmidt method uses the $\mathbf{v}^n$ to create an orthogonal basis $\{\mathbf{u}^1, ..., \mathbf{u}^N\}$ for the span of the $\mathbf{v}^n$. Begin by taking $\mathbf{u}^1 = \mathbf{v}^1$. For $j = 2, ..., N$, let

$$\mathbf{u}^j = \mathbf{v}^j - \frac{\mathbf{u}^1 \cdot \mathbf{v}^j}{\mathbf{u}^1 \cdot \mathbf{u}^1}\mathbf{u}^1 - ... - \frac{\mathbf{u}^{j-1} \cdot \mathbf{v}^j}{\mathbf{u}^{j-1} \cdot \mathbf{u}^{j-1}}\mathbf{u}^{j-1}. \tag{30.3}$$

One obvious problem with this approach is that the calculations become increasingly complicated and lengthy as the $j$ increases. In many of the important examples of orthogonal functions that we study in connection with Sturm-Liouville problems, there is a two-term recursive formula that enables us to generate the next orthogonal function from the two previous ones.

# Chapter 31

# Chaos

## 31.1 The Discrete Logistics Equation

Up to now, our study of differential equations has focused on linear ones, either ordinary or partial. These constitute only a portion of the differential equations of interest to applied mathematics. In this chapter we look briefly at a non-linear ordinary differential equation, the *logistics equation*, and at its discrete version. We then consider the iterative sequence, or *dynamical system*, associated with this discrete logistics equation, and its relation to chaos theory. The best introduction to chaos theory is the book by James Gleick [18].

To illustrate the role of iteration in chaos theory, consider the simplest differential equation describing population dynamics:

$$p'(t) = ap(t), \tag{31.1}$$

with exponential solutions. More realistic models impose limits to growth, and may take the form

$$p'(t) = a(L - p(t))p(t), \tag{31.2}$$

where $L$ is an asymptotic limit for $p(t)$. The solution to Equation (31.2) is

$$p(t) = \frac{p(0)L}{p(0) + (L - p(0))\exp(-aLt)}. \tag{31.3}$$

Discrete versions of the limited-population problem then have the form

$$x_{k+1} - x_k = a(L - x_k)x_k, \tag{31.4}$$

which, for $z_k = \frac{a}{1+aL}x_k$, can be written as

$$z_{k+1} = r(1 - z_k)z_k; \tag{31.5}$$

we shall assume that $r = 1 + aL > 1$. With $Tz = r(1-z)z = f(z)$ and $z_{k+1} = Tz_k$, we are interested in the behavior of the sequence, as a function of $r$.

Figure 31.1 shows the graphs of the functions $y = f(x) = r(1-x)x$, for a fixed value of $r$, and $y = x$. The maximum of $f(x)$ occurs at $x = 0.5$ and the maximum value is $f(0.5) = \frac{r}{4}$. In the Figure, $r$ looks to be about 3.8. Figure 31.2 displays the iterations for several values of $r$, called $\lambda$ in the figures.

## 31.2 Fixed Points

The operator $T$ has a fixed point at $z_* = 0$, for every value of $r$, and another fixed point at $z_* = 1 - \frac{1}{r}$, if $r > 1$. From the Mean-Value Theorem we know that

$$z_* - z_{k+1} = f(z_*) - f(z_k) = f'(c_k)(z_* - z_k), \tag{31.6}$$

for some $c_k$ between $z_*$ and $z_k$. If $z_k$ is sufficiently close to $z_*$, then $c_k$ will be even closer to $z_*$ and $f'(c_k)$ can be approximated by $f'(z_*)$.

In order for $f(x)$ to be a mapping from $[0,1]$ to $[0,1]$ it is necessary and sufficient that $r \leq 4$. For $r > 4$, it is interesting to ask for which starting points $z_0$ does the sequence of iterates remain within $[0,1]$.

## 31.3 Stability

A fixed point $z_*$ of $f(z)$ is said to be *stable* if $|f'(z_*)| < 1$, where $f'(z_*) = r(1-2z_*)$. Since we are assuming that $r > 1$, the fixed point $z_* = 0$ is unstable. The point $z_* = 1 - \frac{1}{r}$ is stable if $1 < r < 3$. When $z_*$ is a stable fixed point, and $z_k$ is sufficiently close to $z_*$, we have

$$|z_* - z_{k+1}| < |z_* - z_k|, \tag{31.7}$$

so we get closer to $z_*$ with each iterative step. Such a fixed point is *attractive*. In fact, if $r = 2$, $z_* = 1 - \frac{1}{r} = \frac{1}{2}$ is *superstable* and convergence is quite rapid, since $f'(\frac{1}{2}) = 0$. We can see from Figure 31.3 that, for $1 < r < 3$, the iterative sequence $\{z_k\}$ has the single limit point $z_* = 1 - \frac{1}{r}$.

What happens beyond $r = 3$ is more interesting. For $r > 3$ the fixed point $z_* = 1 - \frac{1}{r}$ is no longer attracting, so all the fixed points are *repelling*. What can the sequence $\{z_k\}$ do in such a case? As we see from Figure 31.3 and the close-up in Figure 31.5, for values of $r$ from 3 to about 3.45, the sequence $\{z_k\}$ eventually oscillates between two subsequential limits; the sequence is said to have *period two*. Then period doubling occurs. For values of $r$ from about 3.45 to about 3.54, the sequence $\{z_k\}$ has period four, that is, the sequence eventually oscillates among four subsequential

limit points. Then, as $r$ continues to increase, period doubling happens again, and again, and again, each time for smaller increases in $r$ than for the previous doubling. Remarkably, there is a value of $r$ prior to which infinitely many period doublings have taken place and after which *chaos* ensues.

## 31.4 Periodicity

For $1 < r < 3$ the fixed point $z_* = 1 - \frac{1}{r}$ is stable and is an *attracting* fixed point. For $r > 3$, the fixed point $z_*$ is no longer attracting; if $z_k$ is near $z_*$ then $z_{k+1}$ will be farther away.

Using the change of variable $x = -rz + \frac{r}{2}$, the iteration in Equation (31.5) becomes

$$x_{k+1} = x_k^2 + (\frac{r}{2} - \frac{r^2}{4}), \tag{31.8}$$

and the fixed points become $x_* = \frac{r}{2}$ and $x_* = 1 - \frac{r}{2}$.

For $r = 3.835$ there is a starting point $x_0$ for which the iterates are periodic with period three, which implies, according to the results of Li and Yorke, that there are periodic orbits with period $n$, for all positive integers $n$ [33].

## 31.5 Sensitivity to the Starting Value

Using Equation (31.8), the iteration for $r = 4$ can be written as

$$x_{k+1} = x_k^2 - 2. \tag{31.9}$$

In [8] Burger and Starbird illustrate the sensitivity of this iterative scheme to the choice of $x_0$. The numbers in the first column of Figure 31.6 were generated by Excel using Equation (31.9) and starting value $x_0 = 0.5$. To form the second column, the authors retyped the first twelve entries of the first column, exactly as shown on the page, and then let Excel proceed to calculate the remaining ones. Obviously, the two columns become quite different, as the iterations proceed. Why? The answer lies in sensitivity of the iteration to initial conditions.

When Excel generated the first column, it kept more digits at each step than it displayed. Therefore, Excel used more digits to calculate the thirteenth item in the first column than just what is displayed as the twelfth entry. When the twelfth entry, exactly as displayed, was used to generate the thirteenth entry of the second column, those extra digits were not available to Excel. This slight difference, beginning in the *tenth* decimal place, was enough to cause the observed difference in the two tables.

For $r > 4$ the set of starting points in $[0, 1]$ for which the sequence of iterates never leaves $[0, 1]$ is a Cantor set, which is a fractal. The book by Devaney [12] gives a rigorous treatment of these topics; Young's book [48] contains a more elementary discussion of some of the same notions.

## 31.6 Plotting the Iterates

Figure 31.4 shows the values of $z_1$ (red), $z_2$ (yellow), $z_3$ (green), $z_4$ (blue), and $z_5$ (violet), for each $z_0$ in the interval $[0, 1]$, for the four values $r = 1, 2, 3, 4$. For $r = 1$, we have $z_{k+1} = z_k - z_k^2 < z_k$, for all $z_k > 0$, so that the only limit is zero. For $r = 2$, $z_* = 0.5$ is the only attractive fixed point and is the limit, for all $z_0$ in $(0, 1)$. For $r = 3$ we se the beginnings of instability, while by $r = 4$ chaos reigns.

## 31.7 Filled Julia Sets

The $x_k$ in the iteration in Equation (31.8) are real numbers, but the iteration can be applied to complex numbers as well. For each fixed complex number $c$, consider the iterative sequence beginning at $z_0 = 0$, with

$$z_{k+1} = z_k^2 + c. \tag{31.10}$$

We say that the sequence $\{z_k\}$ is *bounded* if there is a constant $B$ such that $|z_k| \leq B$, for all $k$. We want to know for which $c$ the sequence generated by Equation (31.10) is bounded.

In Figure 31.7 those $c$ for which the iterative sequence $\{z_k\}$ is bounded are in the black *Mandelbrot set*, and those $c$ for which the sequence is not bounded are in the white set. It is not apparent from the figure, but when we zoom in, we find the entire figure repeated on a smaller scale. As we continue to zoom in, the figure reappears again and again, each time smaller than before.

There is a theorem that tells us that if $|z_k| \geq 1 + \sqrt{2}$ for some $k$, then the sequence is not bounded. Therefore, if $c$ is in the white set, we will know this for certain after we have computed finitely many iterates. Such sets are sometimes called *recursively enumerable*. However, there does not appear to be an algorithm that will tell us when $c$ is in the black set. The situation is described by saying that the black set, often called the *Mandelbrot set*, is *non-recursive*.

Previously, we were interested in what happens as we change $c$, but start the iteration at $z_0 = 0$ each time. We could modify the problem slightly, using only a single value of $c$, but then starting at arbitrary points $z_0$. Those $z_0$ for which the sequence is bounded form the new black set, called the *filled Julia set* associated with the function $f(z) = z^2 + c$.

For much more on this subject and related ideas, see the book by Roger Penrose [38].

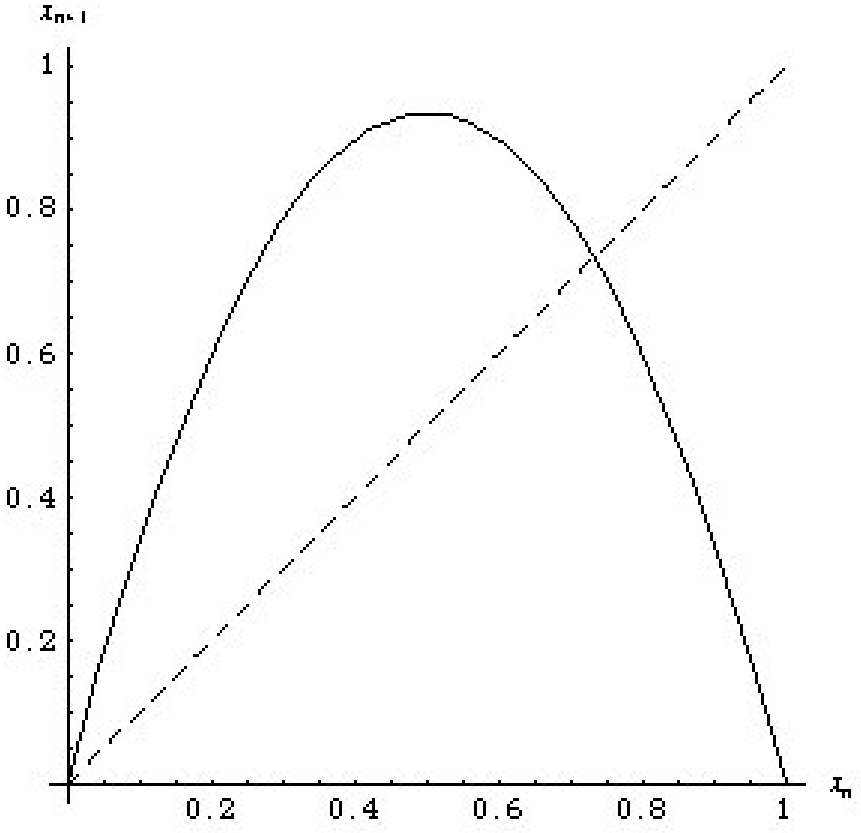



Figure 31.1: The graphs of $y = r(1-x)x$ and $y = x$.

## 31.8 The Newton-Raphson Algorithm

The well known Newton-Raphson (NR) iterative algorithm is used to find a root of a function $g : R \to R$.

**Algorithm 31.1 (Newton-Raphson)** *Let $x^0 \in R$ be arbitrary. Having calculated $x^k$, let*

$$x_{k+1} = x^k - g(x^k)/g'(x^k). \tag{31.11}$$

The operator $T$ is now the ordinary function

$$Tx = x - g(x)/g'(x). \tag{31.12}$$

If $g$ is a vector-valued function, $g : R^J \to R^J$, then $g(x)$ has the form $g(x) = (g_1(x), ..., g_J(x))^T$, where $g_j : R^J \to R$ are the component functions of $g(x)$. The NR algorithm is then as follows:

**Algorithm 31.2 (Newton-Raphson)** *Let $x^0 \in R^J$ be arbitrary. Having calculated $x^k$, let*

$$x_{k+1} = x_k - [\mathcal{J}(g)(x_k)]^{-1} g(x_k). \tag{31.13}$$

Here $\mathcal{J}(g)(x)$ is the Jacobian matrix of first partial derivatives of the component functions of $g$; that is, its entries are $\frac{\partial g_m}{\partial x_j}(x)$. The operator $T$ is now

$$Tx = x - [\mathcal{J}(g)(x)]^{-1} g(x). \tag{31.14}$$

Convergence of the NR algorithm is not guaranteed and depends on the starting point being sufficiently close to a solution. When it does converge, however, it does so fairly rapidly. In both the scalar and vector cases, the limit is a fixed point of $T$, and therefore a root of $g(x)$.

## 31.9 Newton-Raphson and Chaos

It is interesting to consider how the behavior of the NR iteration depends on the starting point.

### 31.9.1 A Simple Case

The complex-valued function $f(z) = z^2 - 1$ of the complex variable $z$ has two roots, $z = 1$ and $z = -1$. The NR method for finding a root now has the iterative step

$$z_{k+1} = Tz_k = \frac{z_k}{2} + \frac{1}{2z_k}. \tag{31.15}$$

If $z_0$ is selected closer to $z = 1$ than to $z = -1$ then the iterative sequence converges to $z = 1$; similarly, if $z_0$ is closer to $z = -1$, the limit is $z = -1$. If $z_0$ is on the vertical axis of points with real part equal to zero, then the sequence does not converge, and is not even defined for $z_0 = 0$. This axis separates the two *basins of attraction* of the algorithm.

### 31.9.2 A Not-So-Simple Case

Now consider the function $f(z) = z^3 - 1$, which has the three roots $z = 1$, $z = \omega = e^{2\pi i/3}$, and $z = \omega^2 = e^{4\pi i/3}$. The NR method for finding a root now has the iterative step

$$z_{k+1} = Tz_k = \frac{2z_k}{3} + \frac{1}{3z_k^2}. \tag{31.16}$$

Where are the *basins of attraction* now? Is the complex plane divided up as three people would divide a pizza, into three wedge-shaped slices, each containing one of the roots? Far from it, as Figure 31.8 shows. In this figure the color of a point indicates the root to which the iteration will converge, if it is started at that point. In fact, it can be shown that, if the sequence starting at $z_0 = a$ converges to $z = 1$ and the sequence starting at $z_0 = b$ converges to $\omega$, then there is a starting point $z_0 = c$, closer to $a$ than $b$ is, whose sequence converges to $\omega^2$. For more details, see Schroeder's delightful book [41].

## 31.10 The Cantor Game

The *Cantor Game* is played as follows. Select a starting point $x_0$ in the interior of the unit interval $[0, 1]$. Let $a$ be the end point closer to $x_0$, with $a = 0$ if there is a tie. Let $x_1$ be the point whose distance from $a$ is three times the distance from $a$ to $x_0$. Now repeat the process, with $x_1$ replacing $x_0$. In order to win the game, we must select a starting point $x_0$ in such a way that all the subsequent points $x_n$ remain within the interior of the interval $[0, 1]$. Where are the winning starting points?

## 31.11 The Sir Pinski Game

In [41] Schroeder discusses several iterative sequences that lead to fractals or chaotic behavior. The *Sir Pinski Game*, a two-dimensional variant of the Cantor Game, has the following rules. Let $P_0$ be a point chosen arbitrarily within the interior of the equilateral triangle with vertices $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$. Let $V$ be the vertex closest to $P_0$ and $P_1$ chosen so that $P_0$ is the midpoint of the line segment $VP_1$. Repeat the process, with $P_1$ in place of $P_0$. The game is lost when $P_n$ falls outside the original triangle. The objective of the game is to select $P_0$ that will allow the player to win the game. Where are these winning points?

The *inverse Sir Pinski Game* is similar. Select any point $P_0$ in the plane of the equilateral triangle, let $V$ be the most distance vertex, and $P_1$ the midpoint of the line segment $P_0V$. Replace $P_0$ with $P_1$ and repeat the

procedure. The resulting sequence of points is convergent. Which points are limit points of sequences obtained in this way?

## 31.12 The Chaos Game

Schroeder also mentions Barnsley's *Chaos Game*. Select $P_0$ inside the equilateral triangle. Roll a fair die and let $V = (1, 0, 0)$ if 1 or 2 is rolled, $V = (0, 1, 0)$ if 3 or 4 is rolled, and $V = (0, 0, 1)$ if 5 or 6 is rolled. Let $P_1$ again be the midpoint of $VP_0$. Replace $P_0$ with $P_1$ and repeat the procedure. Which points are limits of such sequences of points?

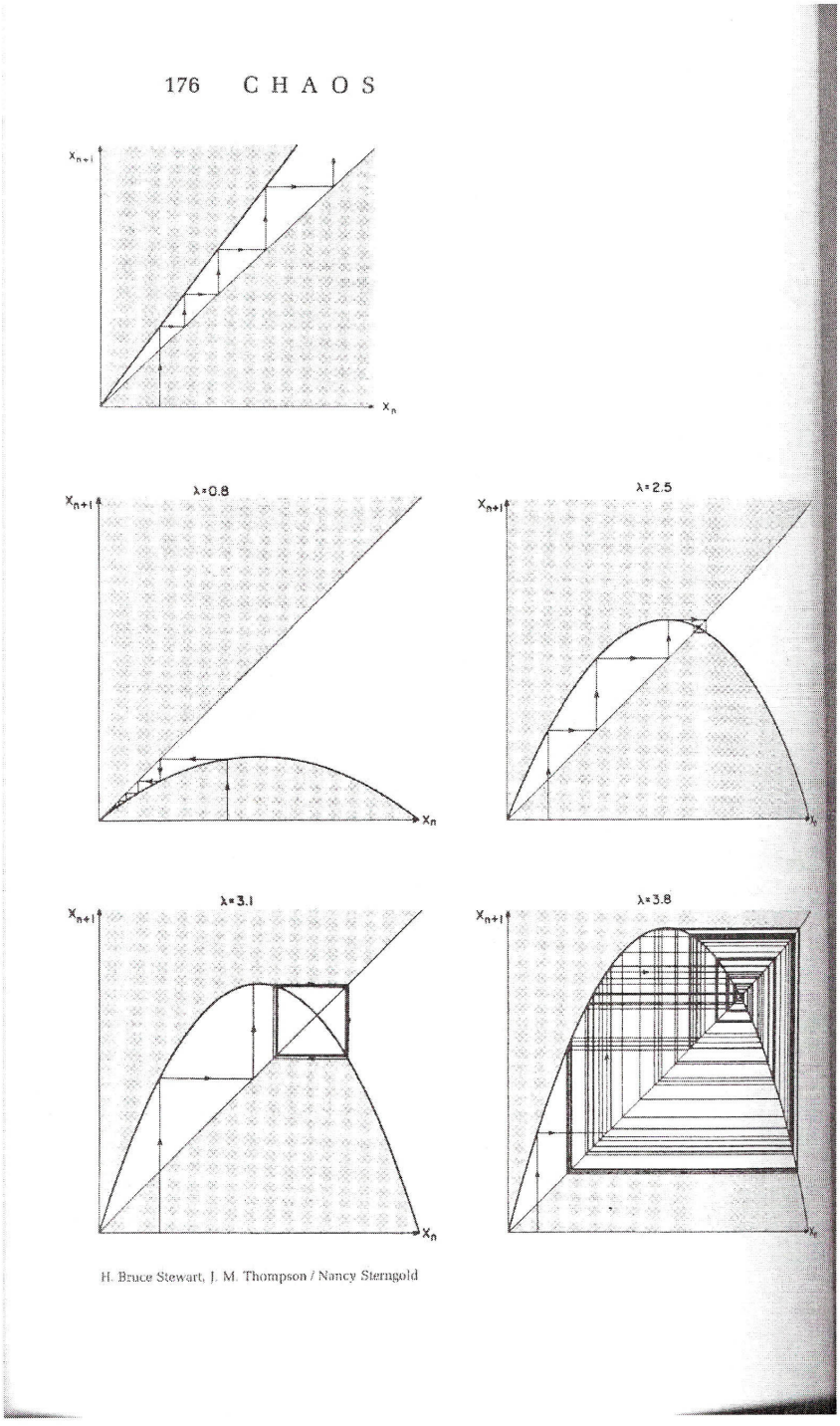



Figure 31.2: Iterations for various values of $r$.

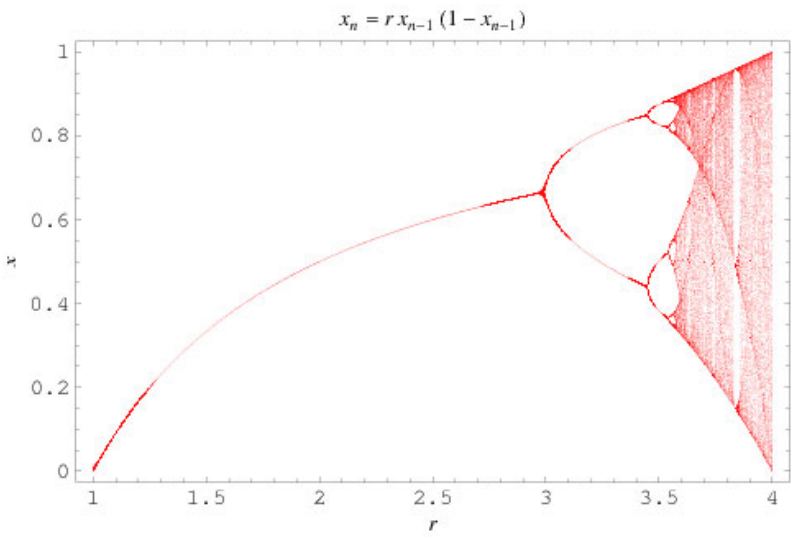



Figure 31.3: Limit behavior for various $r$.

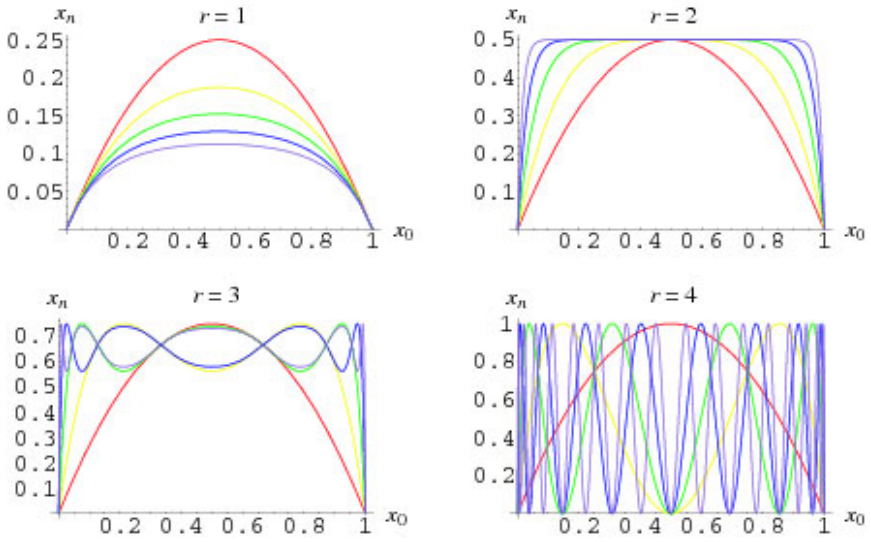



Figure 31.4: Iteration maps.

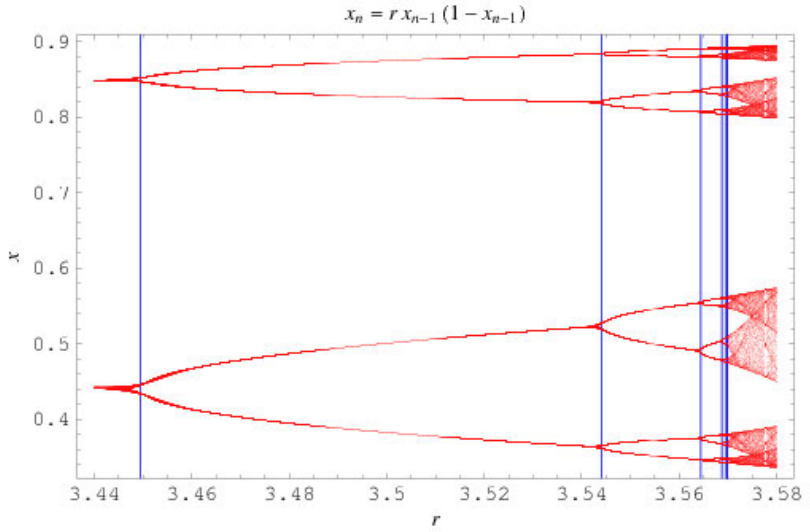
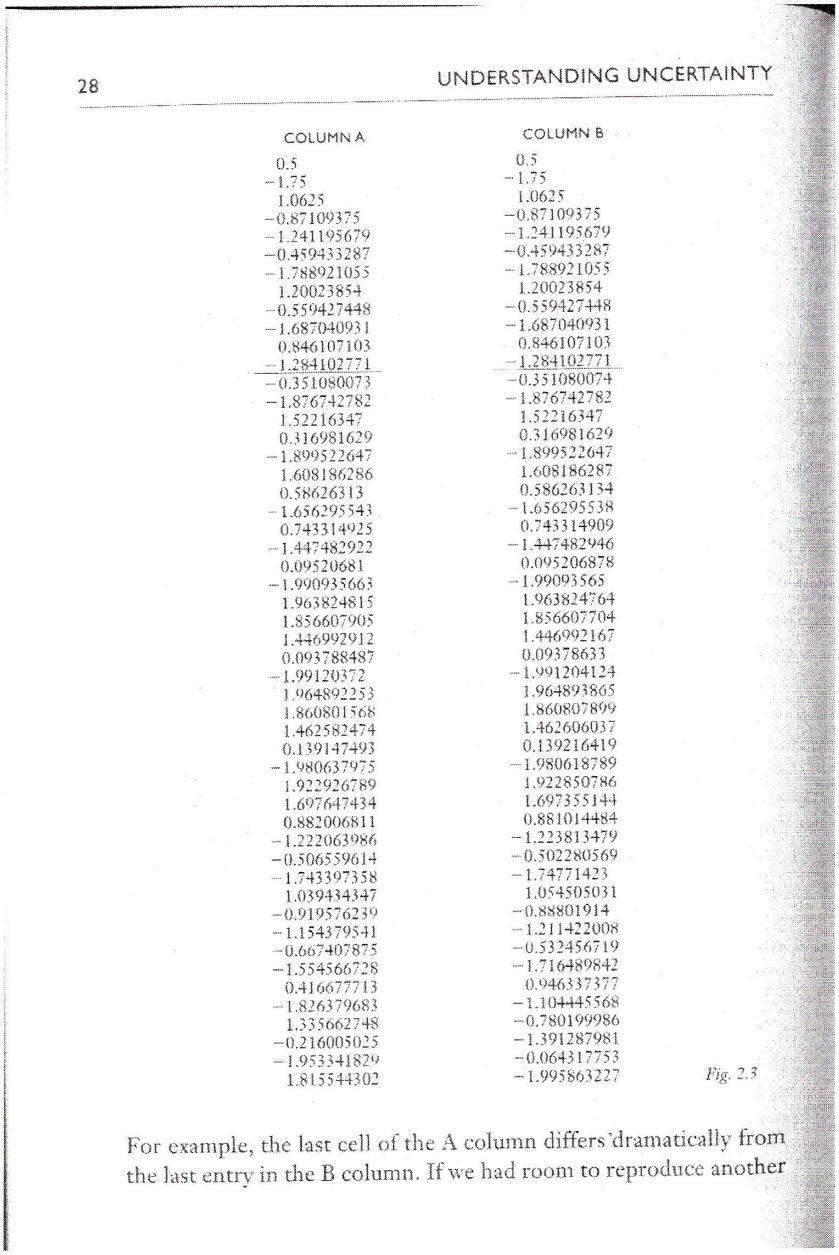



Figure 31.5: Close-up of Figure 31.3.



28 UNDERSTANDING UNCERTAINTY

| COLUMN A | COLUMN B |
|---|---|
| 0.5 | 0.5 |
| −1.75 | −1.75 |
| 1.0625 | 1.0625 |
| −0.87109375 | −0.87109375 |
| −1.241195679 | −1.241195679 |
| −0.459433287 | −0.459433287 |
| −1.788921055 | −1.788921055 |
| 1.20023854 | 1.20023854 |
| −0.559427448 | −0.559427448 |
| −1.687040931 | −1.687040931 |
| 0.846107103 | 0.846107103 |
| −1.284102771 | −1.284102771 |
| −0.351080073 | −0.351080074 |
| −1.876742782 | −1.876742782 |
| 1.52216347 | 1.52216347 |
| 0.316981629 | 0.316981629 |
| −1.899522647 | −1.899522647 |
| 1.608186286 | 1.608186287 |
| 0.58626313 | 0.586263134 |
| −1.656295543 | −1.656295538 |
| 0.743314925 | 0.743314909 |
| −1.447482922 | −1.447482946 |
| 0.09520681 | 0.095206878 |
| −1.990935663 | −1.99093565 |
| 1.963824815 | 1.963824764 |
| 1.856607905 | 1.856607704 |
| 1.446992912 | 1.446992167 |
| 0.093788487 | 0.09378633 |
| −1.99120372 | −1.991204124 |
| 1.964892253 | 1.964893865 |
| 1.860801568 | 1.860807899 |
| 1.462582474 | 1.462606037 |
| 0.139147493 | 0.139216419 |
| −1.980637975 | −1.980618789 |
| 1.922926789 | 1.922850786 |
| 1.697647434 | 1.697355144 |
| 0.882006811 | 0.881014484 |
| −1.222063986 | −1.223813479 |
| −0.506559614 | −0.502280569 |
| −1.743397358 | −1.74771423 |
| 1.039434347 | 1.054505031 |
| −0.919576239 | −0.88801914 |
| −1.154379541 | −1.211422008 |
| −0.667407875 | −0.532456719 |
| −1.554566728 | −1.716489842 |
| 0.416677713 | 0.946337377 |
| −1.826379683 | −1.104445568 |
| 1.335662748 | −0.780199986 |
| −0.216005025 | −1.391287981 |
| −1.953341829 | −0.064317753 |
| 1.815544302 | −1.995863227 |

*Fig. 2.3*

For example, the last cell of the A column differs dramatically from the last entry in the B column. If we had room to reproduce another

Figure 31.6: Sensitivity to initial conditions.

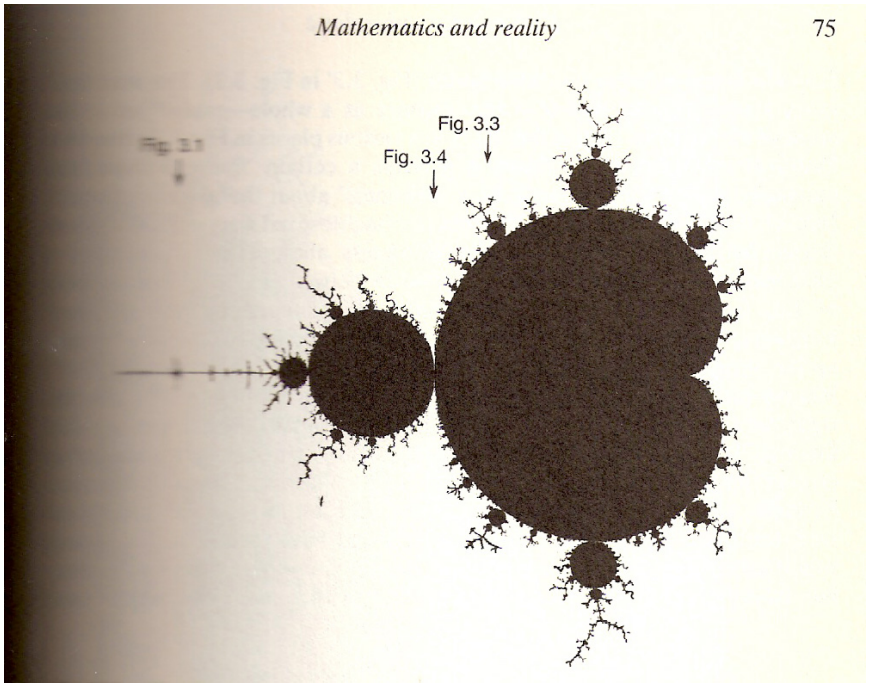



Figure 31.7: The Mandelbrot Set

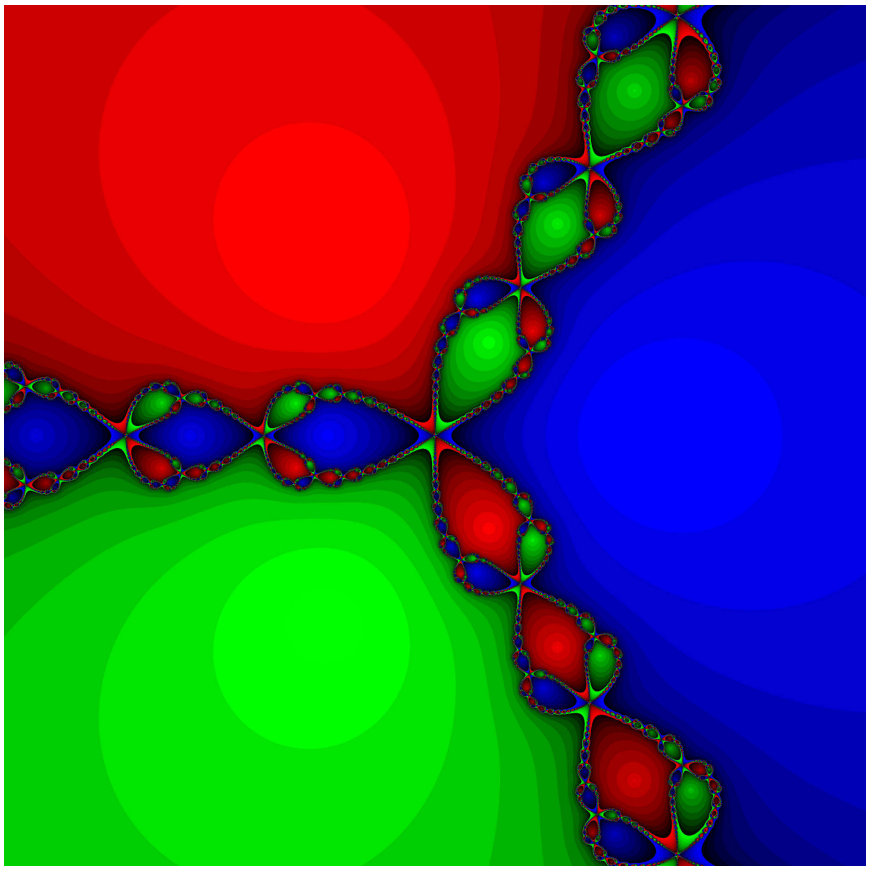


Figure 31.8: Basins of attraction for Equation (31.16).

# Chapter 32

# Wavelets

## 32.1 Analysis and Synthesis

In our discussion of special functions, we saw that the Bessel functions, the Legendre polynomials and the Hermite polynomials provide building blocks, or *basis elements* for certain classes of functions. Each family exhibits a form of orthogonality that makes it easy to calculate the coefficients in an expansion. Wavelets provide other important families of orthogonal basis functions.

An important theme that runs through most of mathematics, from the geometry of the early Greeks to modern signal processing, is *analysis and synthesis*, or, less formally, *breaking up and putting back together*. The Greeks estimated the area of a circle by breaking it up into sectors that approximated triangles. The Riemann approach to integration involves breaking up the area under a curve into pieces that approximate rectangles or other simple shapes. Viewed differently, the Riemann approach is first to approximate the function to be integrated by a step function and then to integrate the step function.

Along with geometry, Euclid includes a good deal of number theory, in which we find analysis and synthesis. His theorem that every positive integer is divisible by a prime is analysis; division does the breaking up and the simple pieces are the primes. The fundamental theorem of arithmetic, which asserts that every positive integer can be written in an essentially unique way as the product of powers of primes, is synthesis, with the putting back together done by multiplication.

## 32.2 Polynomial Approximation

The individual power functions, $x^n$, are not particularly interesting by themselves, but when finitely many of them are scaled and added to form a polynomial, interesting functions can result, as the famous approximation theorem of Weierstrass confirms [32]:

**Theorem 32.1** *If $f : [a, b] \to \mathbb{R}$ is continuous and $\epsilon > 0$ is given, we can find a polynomial $P$ such that $|f(x) - P(x)| \leq \epsilon$ for every $x$ in $[a, b]$.*

The idea of building complicated functions from powers is carried a step further with the use of infinite series, such as Taylor series. The sine function, for example, can be represented for all real $x$ by the infinite power series

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + ....$$

The most interesting thing to note about this is that the sine function has properties that none of the individual power functions possess; for example, it is bounded and periodic. So we see that an infinite sum of simple functions can be qualitatively different from the components in the sum. If we take the sum of only finitely many terms in the Taylor series for the sine function we get a polynomial, which cannot provide a good approximation of the sine function for all $x$; that is, the finite sum does not approximate the sine function uniformly over the real line. The approximation is better for $x$ near zero and poorer as we move away from zero. However, for any selected $x$ and for any $\epsilon > 0$, there is a positive integer $N$, depending on the $x$ and on the $\epsilon$, with the sum of the first $n$ terms of the series within $\epsilon$ of $\sin x$ for $n \geq N$; that is, the series converges pointwise to $\sin x$ for each real $x$. In Fourier analysis the trigonometric functions themselves are viewed as the simple functions, and we try to build more complicated functions as (possibly infinite) sums of trig functions. In wavelet analysis we have more freedom to design the simple functions to fit the problem at hand.

## 32.3 A Radar Problem

To help motivate wavelets, we turn to a signal-processing problem arising in *radar*. The connection between radar signal processing and wavelets is discussed in some detail in Kaiser's book [30].

### 32.3.1 Stationary Target

In radar a real-valued function $\psi(t)$ representing a time-varying voltage is converted by an antenna in transmission mode into a propagating electromagnetic wave. When this wave encounters a reflecting target an echo is

produced. The antenna, now in receiving mode, picks up the echo $f(t)$, which is related to the original signal by

$$f(t) = A\psi(t - d(t)),$$

where $d(t)$ is the time required for the original signal to make the round trip from the antenna to the target and return back at time $t$. The amplitude $A$ incorporates the reflectivity of the target as well as attenuation suffered by the signal. As we shall see shortly, the delay $d(t)$ depends on the distance from the antenna to the target and, if the target is moving, on its radial velocity. The main signal-processing problem here is to determine target range and radial velocity from knowledge of $f(t)$ and $\psi(t)$.

If the target is stationary, at a distance $r_0$ from the antenna, then $d(t) = 2r_0/c$, where $c$ is the speed of light. In this case the original signal and the received echo are related simply by

$$f(t) = A\psi(t - b),$$

for $b = 2r_0/c$.

### 32.3.2 Moving Target

When the target is moving so that its distance to the antenna, $r(t)$, is time-dependent, the relationship between $f$ and $\psi$ is more complicated.

**Exercise 32.1** *Suppose the target is at a distance $r_0 > 0$ from the antenna at time $t = 0$, and has radial velocity $v$, with $v > 0$ indicating away from the antenna. Show that the delay function $d(t)$ is now*

$$d(t) = 2\frac{r_0 + vt}{c + v}$$

*and $f(t)$ is related to $\psi(t)$ according to*

$$f(t) = A\psi(\frac{t - b}{a}), \tag{32.1}$$

*for*

$$a = \frac{c + v}{c - v}$$

*and*

$$b = \frac{2r_0}{c - v}.$$

*Show also that if we select $A = (\frac{c-v}{c+v})^{1/2}$ then energy is preserved; that is, $||f|| = ||\psi||$.*

**Exercise 32.2** *Let $\Psi(\omega)$ be the Fourier transform of the signal $\psi(t)$. Show that the Fourier transform of the echo $f(t)$ in Equation (32.1) is then*

$$F(\omega) = Aae^{ib\omega}\Psi(a\omega). \tag{32.2}$$

The basic problem is to determine $a$ and $b$, and therefore the range and radial velocity of the target, from knowledge of $f(t)$ and $\psi(t)$. An obvious approach is to do a matched filter.

### 32.3.3 The Wideband Cross-Ambiguity Function

Note that the received echo $f(t)$ is related to the original signal by the operations of rescaling and shifting. We therefore match the received echo with all the shifted and rescaled versions of the original signal. For each $a > 0$ and real $b$, let

$$\psi_{a,b}(t) = \psi(\frac{t-b}{a}).$$

The *wideband cross-ambiguity function* (WCAF) is

$$(W_\psi f)(b,a) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)dt. \tag{32.3}$$

In the ideal case the values of $a$ and $b$ for which the WCAF takes on its largest absolute value should be the true values of $a$ and $b$.

## 32.4 Wavelets

### 32.4.1 Background

The fantastic increase in computer power over the last few decades has made possible, even routine, the use of digital procedures for solving problems that were believed earlier to be intractable, such as the modeling of large-scale systems. At the same time, it has created new applications unimagined previously, such as medical imaging. In some cases the mathematical formulation of the problem is known and progress has come with the introduction of efficient computational algorithms, as with the Fast Fourier Transform. In other cases, the mathematics is developed, or perhaps rediscovered, as needed by the people involved in the applications. Only later it is realized that the theory already existed, as with the development of computerized tomography without Radon's earlier work on reconstruction of functions from their line integrals.

It can happen that applications give a theoretical field of mathematics a rebirth; such seems to be the case with *wavelets* [28]. Sometime in the 1980s researchers working on various problems in electrical engineering, quantum mechanics, image processing, and other areas became aware that what the

others were doing was related to their own work. As connections became established, similarities with the earlier mathematical theory of approximation in functional analysis were noticed. Meetings began to take place, and a common language began to emerge around this reborn area, now called wavelets. One of the most significant meetings took place in June of 1990, at the University of Massachusetts Lowell. The keynote speaker was Ingrid Daubechies; the lectures she gave that week were subsequently published in the book [11].

There are a number of good books on wavelets, such as [30], [4], and [45]. A recent issue of the IEEE Signal Processing Magazine has an interesting article on using wavelet analysis of paintings for artist identification [29].

Fourier analysis and synthesis concerns the decomposition, filtering, compressing, and reconstruction of signals using complex exponential functions as the building blocks; wavelet theory provides a framework in which other building blocks, better suited to the problem at hand, can be used. As always, efficient algorithms provide the bridge between theory and practice.

### 32.4.2 A Simple Example

Imagine that $f(t)$ is defined for all real $t$ and we have sampled $f(t)$ every half-second. We focus on the time interval $[0, 2)$. Suppose that $f(0) = 1$, $f(0.5) = -3$, $f(1) = 2$ and $f(1.5) = 4$. We approximate $f(t)$ within the interval $[0, 2)$ by replacing $f(t)$ with the step function that is 1 on $[0, 0.5)$, $-3$ on $[0.5, 1)$, 2 on $[1, 1.5)$, and 4 on $[1.5, 2)$; for notational convenience, we represent this step function by $(1, -3, 2, 4)$. We can decompose $(1, -3, 2, 4)$ into a sum of step functions

$$(1, -3, 2, 4) = 1(1, 1, 1, 1) - 2(1, 1, -1, -1) + 2(1, -1, 0, 0) - 1(0, 0, 1, -1).$$

The first basis element, $(1, 1, 1, 1)$, does not vary over a two-second interval. The second one, $(1, 1, -1, -1)$, is orthogonal to the first, and does not vary over a one-second interval. The other two, both orthogonal to the previous two and to each other, vary over half-second intervals. We can think of these basis functions as corresponding to different frequency components and time locations; that is, they are giving us a time-frequency decomposition.

Suppose we let $\phi_0(t)$ be the function that is 1 on the interval $[0, 1)$ and 0 elsewhere, and $\psi_0(t)$ the function that is 1 on the interval $[0, 0.5)$ and $-1$ on the interval $[0.5, 1)$. Then we say that

$$\phi_0(t) = (1, 1, 0, 0),$$

and

$$\psi_0(t) = (1, -1, 0, 0).$$

Then we write

$$\phi_{-1}(t) = (1, 1, 1, 1) = \phi_0(0.5t),$$

$$\psi_0(t-1) = (0, 0, 1, -1),$$

and

$$\psi_{-1}(t) = (1, 1, -1, -1) = \psi_0(0.5t).$$

So we have the decomposition of $(1, -3, 2, 4)$ as

$$(1, -3, 2, 4) = 1\phi_{-1}(t) - 2\psi_{-1}(t) + 2\psi_0(t) - 1\psi_0(t-1).$$

It what follows we shall be interested in extending these ideas, to find other functions $\phi_0(t)$ and $\psi_0(t)$ that lead to bases consisting of functions of the form

$$\psi_{j,k}(t) = \psi_0(2^j t - k).$$

These will be our *wavelet bases*.

### 32.4.3 The Integral Wavelet Transform

For real numbers $b$ and $a \neq 0$, the *integral wavelet transform* (IWT) of the signal $f(t)$ relative to the *basic wavelet* (or *mother* wavelet) $\psi(t)$ is

$$(W_\psi f)(b, a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t)\psi(\frac{t-b}{a})dt.$$

This function is also the wideband cross-ambiguity function in radar. The function $\psi(t)$ is also called a window function and, like Gaussian functions, it will be relatively localized in time. An example is the *Haar wavelet* $\psi_{Haar}(t)$ that has the value $+1$ for $0 \leq t < \frac{1}{2}$, $-1$ for $\frac{1}{2} \leq t < 1$ and zero otherwise.

As the scaling parameter $a$ grows larger the wavelet $\psi(t)$ grows wider, so choosing a small value of the scaling parameter permits us to focus on a neighborhood of the time $t = b$. The IWT then registers the contribution to $f(t)$ made by components with features on the scale determined by $a$, in the neightborhood of $t = b$. Calculations involving the uncertainty principle reveal that the IWT provides a flexible time-frequency window that narrows when we observe high frequency components and widens for lower frequencies.

Given the integral wavelet transform $(W_\psi f)(b, a)$, it is natural to ask how we might recover the signal $f(t)$. The following inversion formula answers that question: at points $t$ where $f(t)$ is continuous we have

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (W_\psi f)(b, a)\psi(\frac{t-b}{a})\frac{da}{a^2}db,$$

with

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega$$

for $\Psi(\omega)$ the Fourier transform of $\psi(t)$.

### 32.4.4 Wavelet Series Expansions

The Fourier series expansion of a function $f(t)$ on a finite interval is a representation of $f(t)$ as a sum of orthogonal complex exponentials. Localized alterations in $f(t)$ affect every one of the components of this sum. Wavelets, on the other hand, can be used to represent $f(t)$ so that localized alterations in $f(t)$ affect only a few of the components of the wavelet expansion. The simplest example of a wavelet expansion is with respect to the Haar wavelets.

**Exercise 32.3** *Let $w(t) = \psi_{Haar}(t)$. Show that the functions $w_{jk}(t) = w(2^j t - k)$ are mutually orthogonal on the interval $[0, 1]$, where $j = 0, 1, ...$ and $k = 0, 1, ..., 2^j - 1$.*

These functions $w_{jk}(t)$ are the *Haar wavelets*. Every continuous function $f(t)$ defined on $[0, 1]$ can be written as

$$f(t) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} c_{jk} w_{jk}(t)$$

for some choice of $c_0$ and $c_{jk}$. Notice that the *support of the function* $w_{jk}(t)$, the interval on which it is nonzero, gets smaller as $j$ increases. Therefore, the components corresponding to higher values of $j$ in the Haar expansion of $f(t)$ come from features that are localized in the variable $t$; such features are transients that live for only a short time. Such transient components affect all of the Fourier coefficients but only those Haar wavelet coefficients corresponding to terms supported in the region of the disturbance. This ability to isolate localized features is the main reason for the popularity of wavelet expansions.

### 32.4.5 More General Wavelets

The orthogonal functions used in the Haar wavelet expansion are themselves discontinuous, which presents a bit of a problem when we represent continuous functions. Wavelets that are themselves continuous, or better still, differentiable, should do a better job representing smooth functions.

We can obtain other wavelet series expansions by selecting a basic wavelet $\psi(t)$ and defining $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$, for integers $j$ and $k$. We then say that the function $\psi(t)$ is an *orthogonal wavelet* if the family $\{\psi_{jk}\}$ is an orthonormal basis for the space of square-integrable functions on the real line, the Hilbert space $L^2(R)$. This implies that for every such $f(t)$ there are coefficients $c_{jk}$ so that

$$f(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{jk}\psi_{jk}(t),$$

with convergence in the mean-square sense. The coefficients $c_{jk}$ are found using the IWT:

$$c_{jk} = (W_\psi f)(\frac{k}{2^j}, \frac{1}{2^j}).$$

As with Fourier series, wavelet series expansion permits the filtering of certain components, as well as signal compression. In the case of Fourier series, we might attribute high frequency components to noise and achieve a smoothing by setting to zero the coefficients associated with these high frequencies. In the case of wavelet series expansions, we might attribute to noise localized small-scale disturbances and remove them by setting to zero the coefficients corresponding to the appropriate $j$ and $k$. For both Fourier and wavelet series expansions we can achieve compression by ignoring those components whose coefficients are below some chosen level.

# Bibliography


1. Baggott, J. (1992) *The Meaning of Quantum Theory*, Oxford University Press.

2. Bliss, G.A. (1925) *Calculus of Variations* Carus Mathematical Monographs, American Mathematical Society.

3. Bochner, S. (1966) *The Role of Mathematics in the Rise of Science*. Princeton University Press.

4. Boggess, A. and Narcowich, F. (2001) *A First Course in Wavelets, with Fourier Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

5. Bracewell, R.C. (1979) "Image reconstruction in radio astronomy." in [27], pp. 81–104.

6. Brockman, M. (2009) *What's Next? Dispatches on the Future of Science*, Vintage Books, New York.

7. Bryan, K., and Leise, T. (2010) "Impedance imaging, inverse problems, and Harry Potter's cloak ." *SIAM Review*, **52 (2)**, pp. 359–377.

8. Burger, E., and Starbird, M. (2006) *Coincidences, Chaos, and All That Math Jazz* New York: W.W. Norton, Publ.

9. Butterfield, H. (1957) *The Origins of Modern Science: 1300–1800*, Free Press Paperback (MacMillan Co.).

10. Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

11. Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

12. Devaney, R. (1989) *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley.

13. Diamond, J. (1997) *Guns, Germs, and Steel*, Norton, Publ.

14. Fara, P. (2009) *Science: A Four Thousand Year History*, Oxford University Press.

15. Feynman, R., Leighton, R., and Sands, M. (1963) *The Feynman Lectures on Physics, Vol. 1*. Boston: Addison-Wesley.

16. Flanigan, F. (1983) *Complex Variables: Harmonic and Analytic Functions*, Dover Publ.

17. Fleming, W. (1965) *Functions of Several Variables*, Addison-Wesley.

18. Gleick, J. (1987) *Chaos: The Making of a New Science*. Penguin Books.

19. Gonzalez-Velasco, E. (1996) *Fourier Analysis and Boundary Value Problems*. Academic Press.

20. Gonzalez-Velasco, E. (2008) *personal communication*.

21. Graham-Eagle, J. (2008) unpublished notes in applied mathematics.

22. Greenblatt, S. (2011) *The Swerve: How the World Became Modern*. New York: W.W. Norton.

23. Greene, B. (2011) *The Hidden Reality: Parallel Universes and the Deep Laws of the Cosmos*. New York: Vintage Books.

24. Groetsch, C. (1999) *Inverse Problems: Activities for Undergraduates*, The Mathematical Association of America.

25. Heath, T. (1981) *Aristarchus of Samos: The Ancient Copernicus*. Dover Books.

26. Heisenberg, W. (1958) *Physics and Philosophy*, Harper Torchbooks.

27. Herman, G.T. (ed.) (1979) "Image Reconstruction from Projections" , *Topics in Applied Physics, Vol. 32*, Springer-Verlag, Berlin.

28. Hubbard, B. (1998) *The World According to Wavelets*. Natick, MA: A K Peters, Inc.

29. Johnson, C., Hendriks, E., Berezhnoy, I., Brevdo, E., Hughes, S., Daubechies, I., Li, J., Postma, E., and Wang, J. (2008) "Image Processing for Artist Identification" *IEEE Signal Processing Magazine*, **25(4)**, pp. 37–48.

30. Kaiser, G. (1994) *A Friendly Guide to Wavelets*. Boston: Birkhäuser.

31. Koestler, A. (1959) *The Sleepwalkers: A History of Man's Changing Vision of the Universe*, Penguin Books.

32. Körner, T. (1988) *Fourier Analysis*. Cambridge, UK: Cambridge University Press.

33. Li, T., and Yorke, J.A. (1975) "Period Three Implies Chaos" *American Mathematics Monthly*, **82**, pp. 985–992.

34. Lindberg, D. (1992) *The Beginnings of Western Science*, University of Chicago Press.

35. Lindley, D. (2007) *Uncertainty: Einstein, Heisenberg, Bohr, and the Struggle for the Soul of Science*, Doubleday.

36. Muller, R. (2008) *Physics for Future Presidents: the Science Behind the Headlines*, Norton.

37. Papoulis, A. (1977) *Signal Analysis*. New York: McGraw-Hill.

38. Penrose, R. (1989) *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

39. Rigden, J. (2005) *Einstein 1905: The Standard of Greatness*. Harvard University Press.

40. Schey, H.M. (1973) *Div, Curl, Grad, and All That*, W.W. Norton.

41. Schroeder, M. (1991) *Fractals, Chaos, Power Laws*, W.H. Freeman, New York.

42. Simmons, G. (1972) *Differential Equations, with Applications and Historical Notes*. New York: McGraw-Hill.

43. Smolin, L. (2006) *The Trouble with Physics*, Houghton Mifflin.

44. Twomey, S. (1996) *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurement*. New York: Dover Publ.

45. Walnut, D. (2002) *An Introduction to Wavelets*. Boston: Birkhäuser.

46. Witten, E. (2002) "Physical law and the quest for mathematical understanding." *Bulletin of the American Mathematical Society*, **40(1)**, pp. 21–29.

47. Wylie, C.R. (1966) *Advanced Engineering Mathematics*. New York: McGraw-Hill.

48. Young, R. (1992) *Excursions in Calculus: An Interplay of the Continuous and Discrete*, Dolciani Mathematical Expositions Number 13, The Mathematical Association of America.

49. Zajonc, A. (1993) *Catching the Light: the Entwined History of Light and Mind*. Oxford, UK: Oxford University Press.

# Index