Chapter 1

Measurements and Modeling

A quantitative model of a physical system is expressed in the language of mathematics. A qualitative model often precedes a quantitative model. For many years clinicians used medical x-ray images without employing a precise quantitative model. X-rays were thought of as high frequency ‘light’ with three very useful properties:

1. If x-rays are incident on a human body, some fraction of the incident radiation is absorbed or scattered, though a sizable fraction is transmitted. The fraction absorbed or scattered is proportional to the total density of the material encountered. The overall decrease in the intensity of the x-ray beam is called attenuation.

2. A beam of x-ray light travels in a straight line.

3. X-rays darken photographic film. The opacity of the film is a monotone function of the incident energy.

Taken together, these properties mean that using x-rays one can “see through” a human body to obtain a shadow or projection of the internal anatomy on a sheet of film [Figure 1.1(a)].

This model was adequate given the available technology. In their time, x-ray images led to a revolution in the practice of medicine because they opened the door to non-invasive examination of internal anatomy. They are still useful for locating bone fractures, dental caries, and foreign objects, but their ability to visualize soft tissues and more detailed anatomic structure is limited. There are several reasons for this. The principal difficulty is that an x-ray image is a two-dimensional representation of a three-dimensional object. In Figure 1.1(b), the opacity of the film at a point on the film plane is inversely proportional to an average of the density of the object, measured along the line joining the point to the x-ray source. This renders it impossible to deduce the spatial ordering in the missing third dimension.
(a) A old-fashioned chest x-ray image. (Image provided courtesy of Dr. David S. Feigin, ENS Sherri Rudinsky, and Dr. James G. Smirniotopoulos of the Uniformed Services University of the Health Sciences, Dept. of Radiology, Bethesda, MD.)

(b) Depth information is lost in a projection.

**Figure 1.1.** The world of old fashioned x-rays imaging.

A second problem is connected to the “detector” used in traditional x-ray imaging. Photographic film is used to record the total energy in the x rays that are transmitted through the object. Unfortunately film is rather insensitive to x rays. To get a usable image, a light emitting phosphor is sandwiched with the film. This increases the sensitivity of the overall “detector,” but even so, large differences in the intensity of the incident x-ray beam produce small differences in the opacity of film. This means that the contrast between different soft tissues is poor. Beyond this there are other problems caused by the scattering of x rays and noise. Because of these limitations a qualitative theory was adequate for the interpretation of traditional x-ray images.

A desire to improve upon this situation led Alan Cormack, [24], and Godfrey Hounsfield, [64], to independently develop x-ray *tomography* or slice imaging. The first step in their work was to use a quantitative theory for the attenuation of x-rays. Such a theory already existed and is little more than a quantitative restatement of (1) and (2). It is not needed for old fashioned x-ray imaging because traditional x-ray images are read “by eye,” and no further processing is done after the film is developed. Both Cormack and Hounsfield realized that mathematics could be used to infer three-dimensional anatomic structure from a large collection of different two-dimensional projections. The possibility for making this idea work relied on two technological advances:

1. The availability of scintillation crystals to use as detectors

2. Powerful, digital computers to process the tens of thousands of measurements needed to form a usable image

A detector using a scintillation crystal is about a hundred times more sensitive than photographic film. Increasing the dynamic range in the basic measurements makes possible much
finer distinctions. As millions of arithmetic operations are needed for each image, fast computers are a necessity for reconstructing an image from the available measurements. It is an interesting historical note that the mathematics underlying x-ray tomography was done in 1917 by Johan Radon, [105]. It had been largely forgotten, and both Hounsfield and Cormack worked out solutions to the problem of reconstructing an image from its projections. Indeed, this problem had arisen and been solved in contexts as diverse as radio astronomy and statistics.

This book is a detailed exploration of the mathematics that underpins the reconstruction of images in x-ray tomography. While our emphasis is on understanding these mathematical foundations, we constantly return to the practicalities of x-ray tomography. Of particular interest is the relationship between the mathematical treatment of a problem and the realities of numerical computation and physical measurement. There are many different imaging modalities in common use today, such as x-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasound, optical imaging, and electrical impedance imaging. Because each relies on a different physical principle, each provides different information. In every case the mathematics needed to process and interpret the data has a large overlap with that used in x-ray CT. We concentrate on x-ray CT because of the simplicity of the physical principles underlying the measurement process. Detailed descriptions of the other modalities can be found in [91], [76], or [6].

Mathematics is the language in which any quantitative theory or model is eventually expressed. In this introductory chapter we consider a variety of examples of physical systems, measurement processes, and the mathematical models used to describe them. These models illustrate different aspects of more complicated models used in medical imaging. We define the notion of degrees of freedom and relate it to the dimension of a vector space. The chapter concludes by analyzing the problem of reconstructing a region in the plane from measurements of the shadows it casts.

### 1.1 Mathematical Modeling

The first step in giving a mathematical description of a system is to isolate that system from the universe in which it sits. While it is no doubt true that a butterfly flapping its wings in Siberia in midsummer will affect the amount of rainfall in the Amazon rain forest a decade hence, it is surely a tiny effect, impossible to accurately quantify. A practical model includes the system of interest and the major influences on it. Small effects are ignored, though they may come back, as measurement error and noise, to haunt the model. After the system is isolated, we need to find a collection of numerical parameters that describe its state. In this generality these parameters are called state variables. In the idealized world of an isolated system the exact measurement of the state variables uniquely determines the state of the system. It may happen that the parameters that give a convenient description of the system are not directly measurable. The mathematical model then describes relations among the state variables. Using these relations, the state of the system can often be determined from feasible measurements. A simple example should help clarify these
abstract-sounding concepts.

**Example 1.1.1.** Suppose the system is a ball on a rod. For simplicity we assume that the ball has radius zero. The state of the system is described by \((x, y)\), the coordinates of the ball. These are the state variables. If the rod is of length \(r\) and one end of it is fixed at the point \((0, 0)\), then the state variables satisfy the relation

\[ x^2 + y^2 = r^2. \]  

(1.1)

![Diagram of a rod casting a shadow on lineland](image)

**Figure 1.2.** A rod of length \(r\) casts a shadow on lineland.

Imagine now that one-dimensional creatures, living on the \(x\)-axis \(\{y = 0\}\), can observe a shadow of the ball, cast by very distant light sources so that the rays of light are perpendicular to the \(x\)-axis (Figure 1.2). The line creatures want to predict whether or not the ball is about to collide with their world. Locating the shadow determines the \(x\)-coordinate of the ball and using equation (1.1) gives

\[ y = \pm \sqrt{r^2 - x^2}. \]

To determine the sign of the \(y\)-coordinate requires additional information not available in the model. On the other hand, this information is adequate if one only wants to predict if the ball is about to collide with the \(x\)-axis. If the \(x\)-axis is illuminated by red light from above and blue light from below, then a ball approaching from below would cast a red shadow while a ball approaching from above would cast a blue shadow. With these additional data, the location of the ball is completely determined.
Ordered pairs of real numbers, \{(x, y)\}, are the state variables for the system in Example 1.1.1. Because of the constraint (1.1), not every pair defines a state of this system. Generally we define the state space to be values of state variables which correspond to actual states of the system. The state space in Example 1.1.1 is the circle of radius \(r\) centered at \((0, 0)\).

**Exercises**

**Exercise 1.1.1.** Suppose that in Example 1.1.1 light sources are located at \((0, \pm R)\). What is the relationship between the \(x\)-coordinate and the shadow?

**Exercise 1.1.2.** Suppose that in Example 1.1.1 the ball is tethered to \((0, 0)\) by a string of length \(r\). What relations do the state variables \((x, y)\) satisfy? Is there a measurement the line creatures can make to determine the location of the ball? What is the state space for this system?

**Exercise 1.1.3.** Suppose that the ball is untethered but is constrained to lie in the region \(\{(x, y) : 0 \leq y < R\}\). Assume that the points \((x_1, y_1), (x_2, y_2), (x_3, y_3)\) do not lie on a line and have \(y_j > R\). Show that the shadows cast on the line \(y = 0\) by light sources located at these three points determine the location of the ball. Find a formula for \((x, y)\) in terms of the shadow locations. Why are three sources needed?

### 1.1.1 Finitely Many Degrees of Freedom

See: A.1, B.3, B.4, B.5.

The collection of ordered \(n\)-tuples of real numbers

\[
\{(x_1, \ldots, x_n) : x_j \in \mathbb{R}, j = 1, \ldots, n\}
\]

is called Euclidean \(n\)-space and is denoted by \(\mathbb{R}^n\). We often use boldface letters \(x, y\) to denote points in \(\mathbb{R}^n\), which we sometimes call vectors. Recall that if \(x = (x_1, \ldots, x_n)\) and \(y = (y_1, \ldots, y_n)\), then their sum \(x + y\) is defined by

\[
x + y = (x_1 + y_1, \ldots, x_n + y_n),
\]

and if \(a \in \mathbb{R}\), then \(ax\) is defined by

\[
ax = (ax_1, \ldots, ax_n).
\]

These two operations make \(\mathbb{R}^n\) into a real vector space.

**Definition 1.1.1.** If the state of a system is described by a finite collection of real numbers, then the system has finitely many degrees of freedom.
Euclidean $n$-space is the simplest state space for a system with $n$ degrees of freedom. Most systems encountered in elementary physics and engineering have finitely many degrees of freedom. Suppose that the state of a system is specified by a point $x \in \mathbb{R}^n$. Then the mathematical model is expressed as relations that these variables satisfy. These often take the form of functional relations,

$$
\begin{align*}
    f_1(x_1, \ldots, x_n) &= 0 \\
    \vdots \\
    f_m(x_1, \ldots, x_n) &= 0.
\end{align*}
$$

(1.4)

In addition to conditions like those in (1.4) the parameters describing the state of a system might also satisfy inequalities of the form

$$
\begin{align*}
    g_1(x_1, \ldots, x_n) &\geq 0 \\
    \vdots \\
    g_l(x_1, \ldots, x_n) &\geq 0.
\end{align*}
$$

(1.5)

The state space for the system is then the subset of $\mathbb{R}^n$ consisting of solutions to (1.4) which also satisfy (1.5).

**Definition 1.1.2.** An equation or inequality that must be satisfied by a point belonging to the state space of a system is called a constraint.

Example 1.1.1 considers a system with one degree of freedom. The state space for this system is the subset of $\mathbb{R}^2$ consisting of points satisfying (1.1). If the state variables satisfy constraints, then this generally reduces the number of degrees of freedom.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is linear if it satisfies the conditions

$$
\begin{align*}
    f(x + y) &= f(x) + f(y) \text{ for all } x, y \in \mathbb{R}^n \quad \text{and} \\
    f(ax) &= af(x) \text{ for all } a \in \mathbb{R} \text{ and } x \in \mathbb{R}^n.
\end{align*}
$$

(1.6)

Recall that the dot or inner product is the map from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$
\langle x, y \rangle = \sum_{j=1}^{n} x_j y_j.
$$

(1.7)

Sometimes it is denoted by $x \cdot y$. The Euclidean length of $x \in \mathbb{R}^n$ is defined to be

$$
\|x\| = \sqrt{\langle x, x \rangle} = \left[ \sum_{j=1}^{n} x_j^2 \right]^{\frac{1}{2}}.
$$

(1.8)

From the definition it is easy to establish that

$$
\begin{align*}
    \langle x, y \rangle &= \langle y, x \rangle \text{ for all } x, y \in \mathbb{R}^n, \\
    \langle ax, y \rangle &= a \langle x, y \rangle \text{ for all } a \in \mathbb{R} \text{ and } x \in \mathbb{R}^n, \\
    \langle x_1 + x_2, y \rangle &= \langle x_1, y \rangle + \langle x_2, y \rangle \text{ for all } x_1, x_2, y \in \mathbb{R}^n, \\
    \|cx\| &= |c|\|x\| \text{ for all } c \in \mathbb{R} \text{ and } x \in \mathbb{R}^n.
\end{align*}
$$

(1.9)
For $y$ a point in $\mathbb{R}^n$, define the function $f_y(x) = \langle x, y \rangle$. The second and third relations in (1.9) show that $f_y$ is linear. Indeed, every linear function has a such a representation.

**Proposition 1.1.1.** If $f : \mathbb{R}^n \to \mathbb{R}$ is a linear function, then there is a unique vector $y_f$ such that $f(x) = \langle x, y_f \rangle$.

This fact is proved in Exercise 1.1.5. The inner product satisfies a basic inequality called the Cauchy-Schwarz inequality.

**Proposition 1.1.2 (Cauchy-Schwarz inequality).** If $x, y \in \mathbb{R}^n$, then

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$  \hspace{1cm} (1.10)

A proof of this result is outlined in Exercise 1.1.6. The Cauchy-Schwarz inequality shows that if neither $x$ nor $y$ is zero, then

$$-1 \leq \frac{\langle x, y \rangle}{\|x\| \|y\|} \leq 1;$$

this in turn allows us the define the angle between two vectors.

**Definition 1.1.3.** If $x, y \in \mathbb{R}^n$ are both nonvanishing, then the angle $\theta \in [0, \pi]$, between $x$ and $y$ is defined by

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \|y\|}.$$  \hspace{1cm} (1.11)

In particular, two vector are orthogonal if $\langle x, y \rangle = 0$.

The Cauchy-Schwarz inequality implies that the Euclidean length satisfies the triangle inequality.

**Proposition 1.1.3.** For $x, y \in \mathbb{R}^n$, the following inequality holds:

$$\|x + y\| \leq \|x\| + \|y\|.$$  \hspace{1cm} (1.12)

This is called the triangle inequality.

**Remark 1.1.1.** The Euclidean length is an example of a norm on $\mathbb{R}^n$. A real-valued function $N$ defined on $\mathbb{R}^n$ is a norm provided it satisfies the following conditions:

**Non-degeneracy:**

$N(x) = 0$ if and only if $x = 0$,

**Homogeneity:**

$N(ax) = |a|N(x)$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$,

**The triangle inequality:**

$N(x + y) \leq N(x) + N(y)$ for all $x, y \in \mathbb{R}^n$. 
Any norm provides a way to measure distances. The distance between \(x\) and \(y\) is defined to be

\[d_N(x, y) = N(x - y)\]

Suppose that the state of a system is specified by a point in \(\mathbb{R}^n\) subject to the constraints in (1.4). If all the functions \(\{f_1, \ldots, f_m\}\) are linear, then we say that this is a linear model. This is the simplest type of model and also the most common in applications. In this case the set of solutions to (1.4) is a subspace of \(\mathbb{R}^n\). We recall the definition.

**Definition 1.1.4.** A subset \(S \subset \mathbb{R}^n\) is a subspace if

1. the zero vector belongs to \(S\),
2. \(x_1, x_2 \in S\), then \(x_1 + x_2 \in S\),
3. if \(c \in \mathbb{R}\) and \(x \in S\), then \(c x \in S\).

For a linear model it is a simple matter to determine the number of degrees of freedom. Suppose the state space consists of vectors satisfying a single linear equation. In light of Proposition 1.1.1, it can be expressed in the form

\[\langle a_1, x \rangle = 0, \quad (1.13)\]

with \(a_1\) a nonzero vector. This is the equation of a hyperplane in \(\mathbb{R}^n\). The solutions to (1.13) are the vectors in \(\mathbb{R}^n\) orthogonal to \(a_1\). Recall the following definition:

**Definition 1.1.5.** The vectors \(\{v_1, \ldots, v_k\}\) are linearly independent if the only linear combination, \(c_1 v_1 + \cdots + c_k v_k\), that vanishes has all its coefficients, \(\{c_i\}\), equal to zero. Otherwise the vectors are linearly dependent.

The dimension of a subspace of \(\mathbb{R}^n\) can now be defined.

**Definition 1.1.6.** Let \(S \subset \mathbb{R}^n\) be a subspace. If there is a set of \(k\) linearly independent vectors contained in \(S\) but any set with \(k + 1\) or more vectors is linearly dependent, then the dimension of \(S\) equals \(k\). In this case we write \(\text{dim } S = k\).

There is a collection of \((n - 1)\) linearly independent \(n\)-vectors \(\{v_1, \ldots, v_{n-1}\}\) so that \(\langle a_1, x \rangle = 0\) if and only if

\[x = \sum_{i=1}^{n-1} c_i v_i.\]

The hyperplane has dimension \(n - 1\), and therefore a system described by a single linear equation has \(n - 1\) degrees of freedom. The general case is not much harder. Suppose that the state space is the solution set of the system of linear equations

\[
\begin{align*}
\langle a_1, x \rangle &= 0 \\
&\vdots \\
\langle a_m, x \rangle &= 0.
\end{align*}
\]
1.1. Mathematical Modeling

Suppose that \( k \leq m \) is the largest number of linearly independent vectors in the collection \( \{a_1, \ldots, a_m\} \). By renumbering, we can assume that \( \{a_1, \ldots, a_k\} \) are linearly independent, and for any \( l > k \) the vector \( a_l \) is a linear combination of these vectors. Hence if \( x \) satisfies

\[
\langle a_i, x \rangle = 0 \text{ for } 1 \leq i \leq k
\]

then it also satisfies \( \langle a_l, x \rangle = 0 \) for any \( l \) greater than \( k \). The argument in the previous paragraph can be applied recursively to conclude that there is a collection of \( n - k \) linearly independent vectors \( \{u_1, \ldots, u_{n-k}\} \) so that \( x \) solves (1.14) if and only if

\[
x = \sum_{i=1}^{n-k} c_i u_i.
\]

Thus the system has \( n - k \) degrees of freedom.

A nonlinear model can often be approximated by a linear model. If \( f \) is a differentiable function, then the gradient of \( f \) at \( x \) is defined to be

\[
\nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right).
\]

From the definition of the derivative it follows that

\[
f(x_0 + x_1) = f(x_0) + \langle x_1, \nabla f(x_0) \rangle + e(x_1), \quad (1.15)
\]

where the error \( e(x_1) \) satisfies

\[
\lim_{x_1 \to 0} \frac{|e(x_1)|}{\|x_1\|} = 0.
\]

In this case we write

\[
f(x_0 + x_1) \approx f(x_0) + \langle x_1, \nabla f(x_0) \rangle. \quad (1.16)
\]

Suppose that the functions in (1.4) are differentiable and \( f_j(x_0) = 0 \) for \( j = 1, \ldots, m \). Then

\[
f_j(x_0 + x_1) \approx \langle x_1, \nabla f_j(x_0) \rangle.
\]

For small values of \( x_1 \) the system of equations (1.4) can be approximated, near to \( x_0 \), by a system of linear equations,

\[
\begin{align*}
\langle x_1, \nabla f_1(x_0) \rangle &= 0 \\
\vdots \\
\langle x_1, \nabla f_m(x_0) \rangle &= 0,
\end{align*} \quad (1.17)
\]

This provides a linear model that approximates the non-linear model. The accuracy of this approximation depends, in a subtle way, on the collection of vectors \( \{\nabla f_j(x)\} \), for \( x \) near to \( x_0 \). The simplest situation is when these vectors are linearly independent at \( x_0 \). In this case the solutions to

\[
f_j(x_0 + x_1) = 0, \quad j = 1, \ldots, m,
\]
are well approximated, for small $x_1$, by the solutions of (1.17). This is a consequence of the implicit function theorem; see [119].

Often the state variables for a system are divided into two sets, the input variables, $(w_1, \ldots, w_k)$, and output variables, $(z_1, \ldots, z_m)$, with constraints rewritten in the form

\[
\begin{align*}
F_1(w_1, \ldots, w_k) &= z_1 \\
& \vdots \\
F_m(w_1, \ldots, w_k) &= z_m.
\end{align*}
\]  

(1.18)

The output variables are thought of as being measured; the remaining variables must then be determined by solving this system of equations. For a linear model this amounts to solving a system of linear equations. We now consider some examples of physical systems and their mathematical models.

**Example 1.1.2.** We would like to find the height of a mountain without climbing it. To that end, the distance $x$ between the point $P$ and the base of the mountain, as well as the angle $\theta$, are measured (Figure 1.3). If $x$ and $\theta$ are measured exactly, then the height, $h$, of the mountain is given by

\[ h(x, \theta) = x \tan \theta. \]  

(1.19)

Measurements are never exact; using the model and elementary calculus, we can relate the error in the measurement $\theta$ to the error in the computed value of $h$. Suppose that $x$ is measured exactly but there is an uncertainty $\Delta \theta$ in the value of $\theta$. Equation (1.16) gives the linear approximation

\[ h(x, \theta + \Delta \theta) - h(x, \theta) \approx \frac{\partial h}{\partial \theta}(x, \theta) \Delta \theta. \]

As $\partial_\theta h = x \sec^2 \theta$, the height, $h_m$, predicted from the measurement of the angle is given by

\[ h_m = x \tan(\theta + \Delta \theta) \approx x(\tan \theta + \sec^2 \theta \Delta \theta). \]

The approximate value of the absolute error is

\[ h_m - h \approx x \frac{\Delta \theta}{\cos^2 \theta}. \]

The absolute error is a number with the same units as $h$; in general, it is not an interesting quantity. If, for example, the true measurement were 10,000 m, then an error of size 1 m would not be too significant. If the true measurement were 2 m, then this error would be significant. To avoid this obvious pitfall, we normally consider the relative error. In this problem the relative error is

\[ \frac{h_m - h}{h} = \frac{\Delta \theta}{\cos^2 \theta \tan \theta} = \frac{\Delta \theta}{\sin \theta \cos \theta}. \]

Generally the relative error is the absolute error divided by the correct value. It is a dimensionless quantity that gives a quantitative assessment of the accuracy of the measurement.
1.1. Mathematical Modeling

![Figure 1.3](image1.png)

**Figure 1.3.** Using trigonometry to find the height of a mountain.

If the angle \( \theta \) is measured from a point too near to or too far from the mountain (i.e., \( \theta \) is very close to 0 or \( \pi/2 \)), then small measurement errors result in a substantial loss of accuracy. A useful feature of a precise mathematical model is the possibility of estimating how errors in measurement affect the accuracy of the parameters we wish to determine. In Exercise 1.1.13 we consider how to estimate the error entailed in using a linear approximation.

**Example 1.1.3.** In a real situation we cannot measure the distance to the base of the mountain. Suppose that we measure the angles, \( \theta_1 \) and \( \theta_2 \), from two different points, \( P_1 \) and \( P_2 \), as well as the distance \( x_2 - x_1 \) between the two points, as shown in Figure 1.4.

![Figure 1.4](image2.png)

**Figure 1.4.** A more realistic measurement.

From the previous example we know that

\[
\begin{align*}
    h &= x_1 \tan \theta_1, \\
    h &= x_2 \tan \theta_2.
\end{align*}
\]

(1.20)

Using these equations and elementary trigonometry, we deduce that

\[
    x_1 = \frac{x_2 - x_1}{\tan \theta_1 - \tan \theta_2}.
\]

(1.21)
which implies that
\[
h = x_1 \tan \theta_1 \\
= (x_2 - x_1) \frac{\sin \theta_1 \sin \theta_2}{\sin(\theta_1 - \theta_2)}.
\] (1.22)

Thus \( h \) can be determined from \( \theta_1, \theta_2 \) and \( x_2 - x_1 \). With \( d = x_2 - x_1 \), equation (1.22) expresses \( h \) as a function of \( (d, \theta_1, \theta_2) \). At the beginning of this example, \((x_1, \theta_1, x_2, \theta_2, h)\) were the state variables describing our system; by the end we used \((d, \theta_1, \theta_2, h)\). The first three are directly measurable, and the last is an explicit function of the others. The models in this and the previous example, as expressed by equations (1.22) and (1.19), respectively, are nonlinear models.

In this example there are many different ways that the model may fail to capture important features of the physical situation. We now consider a few potential problems.

1. If the shape of a mountain looks like that in Figure 1.5 and we measure the distance and angle at the point \( P \), we are certainly not finding the real height of the mountain. Some a priori information is always incorporated in a mathematical model.

![Figure 1.5](image)

\textbf{Figure 1.5.} Not exactly what we predicted!

2. The curvature of the earth is ignored. A more sophisticated geometric model is needed to correct for such errors. This becomes a significant problem as soon as the distances, \( x, x_1, x_2 \), are large compared to the distance to the horizon (about 25 km for a 2-meter-tall person). The approximations used in the model must be adapted to the actual physical conditions of the measurements.

3. The geometry of the underlying measurements could be quite different from the simple Euclidean geometry used in the model. To measure the angles \( \theta_1, \theta_2 \), we would normally use a transit to sight the peak of the mountain. If the mountain is far away, then the light traveling from the mountain to the transit passes through air of varying density. The light is refracted by the air and therefore the ray path is not a straight line, as assumed in the model. To include this effect would vastly complicate the model. This is an important consideration in the similar problem of creating a map of the sky from earth based observations of stars and planets.

Analogous problems arise in medical imaging. If the wavelength of the energy used to probe the human anatomy is very small compared to the size of the structures that are
present, then it is reasonable to assume that the waves are not refracted. For example, x-rays can be assumed to travel along straight lines. For energies with wavelengths comparable to the size of structures present in the human anatomy, this assumption is simply wrong. The waves are then bent and diffracted by the medium, and the difficulty of modeling the ray paths is considerable. This is an important issue in ultrasound imaging that remains largely unresolved.

**Example 1.1.4.** Refraction provides another example of a simple physical system. Suppose that we have two fluids in a tank, as shown in Figure 1.6, and would like to determine the height of the interface between them. Suppose that the refractive indices of the fluids are known. Let $n_1$ be the refractive index of the upper fluid and $n_2$ the refractive index of the lower one. Snell’s law states that

$$\frac{\sin(\theta_1)}{\sin(\theta_2)} = \frac{n_2}{n_1}.$$ 

Let $h$ denote the total height of the fluid; then

$$h_1 + h_2 = h.$$ 

The measurement we make is the total displacement, $l$, of the light ray as it passes through the fluids. It satisfies the relationship

$$h_1 \tan(\theta_1) + h_2 \tan(\theta_2) = l.$$ 

The heights $h_1$ and $h_2$ are easily determined from these three formulæ. The assumption that we know $n_1$ implies, by Snell’s law, that we can determine $\theta_1$ from a measurement of

![Figure 1.6. Using refraction to determine the height of an interface.](image-url)
the angle of the light ray above the fluid. If \( n_2 \) is also known, then using these observations we can determine \( \theta_2 \) as well:

\[
\sin(\theta_2) = \frac{n_1}{n_2} \sin(\theta_1).
\]

The pair \((h_1, h_2)\) satisfies the \(2 \times 2\) linear system

\[
\begin{pmatrix}
1 & 1 \\
\tan(\theta_1) & \tan(\theta_2)
\end{pmatrix}
\begin{pmatrix}
h_1 \\
h_2
\end{pmatrix}
= \begin{pmatrix}
h \\\nt
\end{pmatrix}.
\] (1.23)

In Example 2.1.1 we consider a slightly more realistic situation where the refractive index of the lower fluid in not known. By using more measurements, \( n_2 \) can also be determined.

**Exercises**

**Exercise 1.1.4.** Prove the formulæ in (1.9).

**Exercise 1.1.5.** Let \( e_j \in \mathbb{R}^n \), \( j = 1, \ldots, n \) denote the vector with a 1 in the \( j \)th place and otherwise zero,

\[
e_1 = (1, 0, 0, \ldots, 0), \quad e_2 = (0, 1, 0, \ldots, 0), \ldots, \quad e_n = (0, \ldots, 0, 1).
\]

1. Show that if \( x = (x_1, \ldots, x_n) \), then

\[
x = \sum_{j=1}^n x_j e_j.
\]

2. Use the previous part to prove the existence statement in Proposition 1.1.1; that is, show that there is a vector \( y_f \) so that \( f(x) = \langle x, y_f \rangle \). Give a formula for \( y_f \).

3. Show that the uniqueness part of the proposition is equivalent to the statement “If \( y \in \mathbb{R}^n \) satisfies

\[
\langle x, y \rangle = 0 \text{ for all } x \in \mathbb{R}^n,
\]

then \( y = 0 \).” Prove this statement.

**Exercise 1.1.6.** In this exercise we use calculus to prove the Cauchy-Schwarz inequality. Let \( x, y \in \mathbb{R}^n \) be nonzero vectors. Define the function

\[
F(t) = \langle x + ty, x + ty \rangle.
\]

Use calculus to find the value of \( t \), where \( F \) assumes its minimum value. By using the fact that \( F(t) \geq 0 \) for all \( t \), deduce the Cauchy-Schwarz inequality.

**Exercise 1.1.7.** Show that (1.12) is a consequence of the Cauchy-Schwarz inequality. *Hint:* Consider \( \|x + y\|^2 \).
Exercise 1.1.8. Define a real-valued function on $\mathbb{R}^n$ by setting

$$N(x) = \max\{|x_1|, \ldots, |x_n|\}.$$ 

Show that $N$ defines a norm.

Exercise 1.1.9. Let $N$ be a norm on $\mathbb{R}^n$ and define $d(x, y) = N(x - y)$. Show that for any triple of points $x_1, x_2, x_3$, the following estimate holds:

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3).$$

Explain why this is also called the triangle inequality.

Exercise 1.1.10. Let $S \subset \mathbb{R}^n$ be a subspace of dimension $k$. Show that there exists a collection of vectors $\{v_1, \ldots, v_k\} \subset S$ such that every vector $x \in S$ has a unique representation of the form

$$x = c_1v_1 + \cdots + c_kv_k.$$ 

Exercise 1.1.11. Let $a$ be a nonzero $n$-vector. Show that there is a collection of $n-1$ linearly independent $n$-vectors, $\{v_1, \ldots, v_{n-1}\}$, so that $x$ solves $(a, x) = 0$ if and only if

$$x = \sum_{j=1}^{n-1} c_jv_j$$

for some real constants $\{c_1, \ldots, c_{n-1}\}$.

Exercise 1.1.12. Let $\{a_1, \ldots, a_k\}$ be linearly independent $n$-vectors. Show that there is a collection of $n-k$ linearly independent $n$-vectors, $\{v_1, \ldots, v_{n-k}\}$, so that $x$ solves

$$(a_j, x) = 0 \text{ for } j = 1, \ldots, k$$

if and only if

$$x = \sum_{j=1}^{n-k} c_jv_j$$

for some real constants $\{c_1, \ldots, c_{n-k}\}$. Hint: Use the previous exercise and an induction argument.

Exercise 1.1.13. If a function $f$ has two derivatives, then Taylor’s theorem gives a formula for the error $e(y) = f(x + y) - [f(x) + f'(x)y]$. There exists a $z$ between 0 and $y$ such that

$$e(z) = \frac{f''(z)y^2}{2};$$

see (B.13). Use this formula to bound the error made in replacing $h(x, \theta + \Delta \theta)$ with $h(x, \theta) + \partial_\theta h(x, \theta) \Delta \theta$. Hint: Find the value of $z$ between 0 and $\Delta \theta$ that maximizes the error term.

Exercise 1.1.14. In Example 1.1.3 compute the gradient of $h$ to determine how the absolute and relative errors depend on $\theta_1, \theta_2$, and $d$. 

---

Copyright ©2007 by the Society for Industrial and Applied Mathematics. 
This electronic version is for personal use and may not be duplicated or distributed. 

From "Introduction to the Mathematics of Medical Imaging, Second Edition" by Charles L. Epstein. 
This book is available for purchase at www.siam.org/catalog.
1.1.2 Infinitely Many Degrees of Freedom

In the previous section we examined some simple physical systems with finitely many degrees of freedom. In these examples, the problem of determining the state of the system from feasible measurements reduces to solving systems of finitely many equations in finitely many unknowns. In imaging applications the state of a system is usually described by a function or functions of continuous variables. These systems have infinitely many degrees of freedom. In this section we consider several examples.

![Particle trajectory diagram](Diagram.png)

**Figure 1.7.** Particle scattering can be used to explore the boundary of an unknown region.

**Example 1.1.5.** Suppose that we would like to determine the shape of a planar region, $D$, that cannot be seen. The object is lying inside a disk and we can fire particles at the object. Assume that the particles bounce off according to a simple scattering process. Each particle strikes the object once and is then scattered along a straight line off to infinity (Figure 1.7). The outline of the object can be determined by knowing the correspondence between incoming lines, $l_{in}$, and outgoing lines, $l_{out}$. Each intersection point $l_{in} \cap l_{out}$ lies on the boundary of the object. Measuring $\{l_{out}^j\}$ for finitely many incoming directions $\{l_{in}^j\}$ determines finitely many points $\{l_{in}^j \cap l_{out}^j\}$ on the boundary of $D$. In order to use this finite collection of points to make any assertions about the rest of the boundary of $D$, more information is required. If we know that $D$ consists of a single piece or component, then these points would lie on a single closed curve, though it might be difficult to decide in what
1.1. Mathematical Modeling

order they should appear on the curve. On the other hand, these measurements provide a lot of information about convex regions.

**Definition 1.1.7.** A region $D$ in the plane is convex if it has the following property: For each pair of points $p$ and $q$ lying in $D$, the line segment $\overline{pq}$ is also contained in $D$. See Figure 1.8.

Convex regions have many special properties. If $p$ and $q$ are on the boundary of $D$, then the line segment $\overline{pq}$ lies inside of $D$. From this observation we can show that if $\{p_1, \ldots, p_N\}$ are points on the boundary of a convex region, then the smallest polygon with these points as vertices lies entirely within $D$ [Figure 1.9(a)]. Convexity can also be defined by a property of the boundary of $D$: For each point $p$ on the boundary of $D$ there is a line $l_p$ that passes through $p$ but is otherwise disjoint from $D$. This line is called a support line through $p$. If the boundary is smooth at $p$, then the tangent line to the boundary is the unique support line. A line divides the plane into two half-planes. Let $l_p$ be a support line to $D$ at $p$. Since the interior of $D$ does not meet $l_p$, it must lie entirely in one of the half-planes determined by this line [see Figure 1.9(b)]. If each support line meets the boundary of $D$ at exactly one point, then the region is strictly convex.
A triangle is the boundary of a convex region, with each edge of the triangle a support line. As infinitely many points of the boundary belong to each edge, the region bounded by a triangle in not strictly convex. On the other hand, through each vertex of the triangle, there are infinitely many support lines. These observations are illustrated in Figure 1.9(c).

Suppose that the object is convex and more is known about the scattering process: for example, that the angle of incidence is equal to the angle of reflection. From a finite number of incoming and outgoing pairs, \( \{ (l^i_{\text{in}}, l^i_{\text{out}}) : i = 1, \ldots, N \} \), we can now determine an approximation to \( D \) with an estimate for the error. The intersection points, \( p_i = l^i_{\text{in}} \cap l^i_{\text{out}} \), lie on the boundary of the convex region, \( D \). If we use these points as the vertices of a polygon \( P^\text{in}_N \), then, as remarked previously, \( P^\text{in}_N \) is completely contained within \( D \). On the other hand, as the angle of incidence equals the angle of reflection, we can also determine the tangent lines \( \{ l_{p_i} \} \) to the boundary of \( D \) at the points \( \{ p_i \} \). These lines are support lines for \( D \). Hence by intersecting the half-planes that contain \( D \), defined by these tangent lines, we obtain another convex polygon, \( P^\text{out}_N \), that contains \( D \). Thus with these \( N \)-measurements we obtain the both an inner and outer approximation to \( D \):

\[
P^\text{in}_N \subset D \subset P^\text{out}_N.
\]

An example is shown in Figure 1.10.

A convex region is determined by its boundary, and each point on the boundary is, in effect, a state variable. Therefore, the collection of convex regions is a system with infinitely many degrees of freedom. A nice description for the state space of smooth convex regions is developed in Section 1.2.2. As we have seen, a convex region can be approximated by polygons. Once the number of sides is fixed, then we are again considering a system with finitely many degrees of freedom. In all practical problems, a system with infinitely many degrees of freedom must eventually be approximated by a system with finitely many degrees of freedom.

Remark 1.1.1. For a non-convex body, the preceding method does not work as the correspondence between incoming and outgoing lines can be complicated: Some incoming lines...
1.2. A Simple Model Problem for Image Reconstruction

may undergo multiple reflections before escaping, and in fact some lines might become permanently trapped.

Figure 1.10. An inner and an outer approximation to a convex region.

Exercises

Exercise 1.1.15. Find state variables to describe the set of polygons with $n$-vertices in the plane. For the case of triangles, find the relations satisfied by your variables. Extra credit: Find a condition, in terms of your parameters, implying that the polygon is convex.

Exercise 1.1.16. Suppose that $D_1$ and $D_2$ are convex regions in the plane. Show that their intersection $D_1 \cap D_2$ is also a convex region.

Exercise 1.1.17.* Suppose that $D$ is a possibly non-convex region in the plane. Define a new region $D'$ as the intersection of all the half-planes that contain $D$. Show that $D = D'$ if and only if $D$ is convex.

Exercise 1.1.18. Find an example of a planar region such that at least one particle trajectory is trapped forever.

1.2 A Simple Model Problem for Image Reconstruction

The problem of image reconstruction in x-ray tomography is sometimes described as reconstructing an object from its “projections.” Of course, these are projections under the illumination of x-ray “light.” In this section we consider the analogous but simpler problem of determining the outline of a convex object from its shadows. As is also the case in
medical applications, we consider a two-dimensional problem. Let \( D \) be a convex region in the plane. Imagine that a light source is placed very far from \( D \). Since the light source is very far away, the rays of light are all traveling in essentially the same direction. We can think of them as a collection of parallel lines. We want to measure the shadow that \( D \) casts for each position of the light source. To describe the measurements imagine that a screen is placed on the “other side” of \( D \) perpendicular to the direction of the light rays (Figure 1.11). In a real apparatus sensors would be placed on the screen, allowing us to determine where the shadow begins and ends.

![Figure 1.11. The shadow of a convex region.](image)

The region, \( D \), blocks a certain collection of light rays and allows the rest to pass. Locating the shadow amounts to determining the “first” and “last” lines in this family of parallel lines to intersect \( D \). To describe the object completely, we need to rotate the source and detector through \( \pi \) radians, measuring, at each angle, where the shadow begins and ends.

The first and last lines to intersect a region just meet it along its boundary. These lines are therefore tangent to the boundary of \( D \). The problem of reconstructing a region from its shadows is mathematically the same as the problem of reconstructing a region from a knowledge of the tangent lines to its boundary. As a first step in this direction we need a good way to organize our measurements. To that end we give a description for the space of lines in the plane.

### 1.2.1 The Space of Lines in the Plane

A line in the plane is a set of points that satisfies an equation of the form

\[
ax + by = c,
\]
where \(a^2 + b^2 \neq 0\). We could use \((a, b, c)\) to parameterize the set of lines, but note that we get the same set of points if we replace this equation by

\[
\frac{a}{\sqrt{a^2 + b^2}} x + \frac{b}{\sqrt{a^2 + b^2}} y = \frac{c}{\sqrt{a^2 + b^2}}.
\]

The coefficients \(\frac{a}{\sqrt{a^2 + b^2}}, \frac{b}{\sqrt{a^2 + b^2}}\) define a point \(\omega\) on the unit circle, \(S^1 \subset \mathbb{R}^2\), and the constant \(\frac{c}{\sqrt{a^2 + b^2}}\) can be any number. The lines in the plane are parameterized by a pair consisting of a unit vector

\[
\omega = (\omega_1, \omega_2)
\]

and a real number \(t\). The line \(l_{t, \omega}\) is the set of points in \(\mathbb{R}^2\) satisfying the equation

\[
\langle (x, y), \omega \rangle = t.
\]

(1.24)

The vector \(\omega\) is perpendicular to this line (Figure 1.12).

It is often convenient to parameterize the points on the unit circle by a real number; to that end we set

\[
\omega(\theta) = (\cos(\theta), \sin(\theta)).
\]

(1.25)

Since \(\cos\) and \(\sin\) are \(2\pi\)-periodic, it clear that \(\omega(\theta) = \omega(\theta + 2\pi)\) are the same point on the unit circle. Using this parameterization for points on the circle, the line \(l_{t, \omega} \overset{d}{=} l_{t, \omega(\theta)}\) is the set of solutions to the equation

\[
\langle (x, y), (\cos(\theta), \sin(\theta)) \rangle = t.
\]

Both notations for lines and points on the circle are used in the sequel.

While the parameterization provided by \((t, \omega)\) is much more efficient than that provided by \((a, b, c)\), note that the set of points satisfying (1.24) is unchanged if \((t, \omega)\) is replaced by \((-t, -\omega)\). Thus, as sets,

\[
l_{t, \omega} = l_{-t, -\omega}.
\]

(1.26)

It is not difficult to show that if \(l_{t_1, \omega_1} = l_{t_2, \omega_2}\), then either \(t_1 = t_2\) and \(\omega_1 = \omega_2\) or \(t_1 = -t_2\) and \(\omega_1 = -\omega_2\).

The pair \((t, \omega)\) actually specifies an oriented line. That is, we can use these data to define the positive direction along the line. The vector

\[
\hat{\omega} = (-\omega_2, \omega_1)
\]

is perpendicular to \(\omega\) and is therefore parallel to \(l_{t, \omega}\). In fact, \(\hat{\omega}\) and \(-\hat{\omega}\) are both unit vectors that are parallel to \(l_{t, \omega}\). The vector \(\hat{\omega}\) is selected by using the condition that the \(2 \times 2\) matrix,

\[
\begin{pmatrix}
\omega_1 & -\omega_2 \\
\omega_2 & \omega_1
\end{pmatrix},
\]

has determinant \(+1\). The vector \(\hat{\omega}\) defines the positive direction or orientation of the line \(l_{t, \omega}\). This explains how the pair \((t, \omega)\) determines an oriented line. We summarize these computations in a proposition.
Proposition 1.2.1. The pairs \((t, \omega) \in \mathbb{R} \times S^1\) are in one-to-one correspondence with the set of oriented lines in the plane.

Figure 1.12. Parameterization of oriented lines in the plane.

The vector \(\omega\) is the direction orthogonal to the line and the number \(t\) is called the affine parameter of the line; \(|t|\) is the distance from the line to the origin of the coordinate system.

The pair \((t, \omega)\) defines two half-planes

\[
H^+_t,\omega = \{ x \in \mathbb{R}^2 \mid \langle x, \omega \rangle > t \} \quad \text{and} \quad H^-_{t,\omega} = \{ x \in \mathbb{R}^2 \mid \langle x, \omega \rangle < t \};
\]

the line \(l_{t,\omega}\) is the common boundary of these half-planes. Facing along the line \(l_{t,\omega}\) in the direction specified by \(\hat{\omega}\), the half-plane \(H^-_{t,\omega}\) lies to the left.

**Exercises**

**Exercise 1.2.1.** Show that \(l_{t,\omega}\) is given parametrically as the set of points

\[
l_{t,\omega} = \{ t\omega + s\hat{\omega} : s \in (-\infty, \infty) \}.
\]

**Exercise 1.2.2.** Show that if \(\omega = (\cos(\theta), \sin(\theta))\), then \(\hat{\omega} = (-\sin(\theta), \cos(\theta))\), and as a function of \(\theta\):

\[
\hat{\omega}(\theta) = \partial_\theta \omega(\theta).
\]

**Exercise 1.2.3.** Suppose that \((t, \omega)\) and \((t_1, \omega_1)\) are different points in \(\mathbb{R} \times S^1\) such that \(l_{t,\omega} = l_{t_1,\omega_1}\). Show that \((t_1, \omega_1) = (-t_2, -\omega_2)\).
**Exercise 1.2.4.** Show that
\[ |t| = \min \{ \sqrt{x^2 + y^2} : (x, y) \in l_t, \omega \}. \]

**Exercise 1.2.5.** Show that if \( \omega \) is fixed, then the lines in the family \( \{ l_t, \omega : t \in \mathbb{R} \} \) are parallel.

**Exercise 1.2.6.** Show that every line in the family \( \{ l_t, \omega : t \in \mathbb{R} \} \) is orthogonal to every line in the family \( \{ l_t, \omega : t \in \mathbb{R} \} \).

**Exercise 1.2.7.** Each choice of direction \( \omega \) defines a coordinate system on \( \mathbb{R}^2 \),
\[ (x, y) = t \omega + s \hat{\omega}. \]
Find the inverse, expressing \( (t, s) \) as functions of \( (x, y) \). Show that the area element in the plane satisfies
\[ dx \, dy = dt \, ds. \]

### 1.2.2 Reconstructing an Object from Its Shadows

Now we can quantitatively describe the shadow. Because there are two lines in each family of parallel lines that are tangent to the boundary of \( D \), we need a way to select one of them. To do this we choose an orientation for the boundary of \( D \); this operation is familiar from Green’s theorem in the plane. The positive direction on the boundary is selected so that, when facing in this direction the region lies to the left; the counterclockwise direction is, by convention, the positive direction (Figure 1.13).

Fix a source position \( \omega(\theta) \). In the family of parallel lines \( \{ l_t, \omega(\theta) : t \in \mathbb{R} \} \) there are two values of \( t \), \( t_0 < t_1 \), such that the lines \( l_{t_0}, \omega(\theta) \) and \( l_{t_1}, \omega(\theta) \) are tangent to the boundary of \( D \) (Figure 1.13). Examining the diagram, it is clear that the orientation of the boundary at the point of tangency and that of the oriented line agree for \( l_{t_1}, \omega \), and are opposite for \( l_{t_0}, \omega \).

Define \( h_D \), the shadow function of \( D \), by setting
\[ h_D(\theta) = t_1 \text{ and } h_D(\theta + \pi) = -t_0. \]
(1.28)

The shadow function is completely determined by values of \( \theta \) belonging to an interval of length \( \pi \). Because \( \omega(\theta) = \omega(\theta + 2\pi) \), the shadow function can be regarded as a \( 2\pi \)-periodic function defined on the whole real line. The mathematical formulation of reconstruction problem is as follows: Can the boundary of the region \( D \) be determined from \( h_D \)?

As \( \omega(\theta) = (\cos(\theta), \sin(\theta)) \), the line \( l_{h_D(\theta), \omega(\theta)} \) is given parametrically by
\[ \{ h_D(\theta)(\cos(\theta), \sin(\theta)) + s(-\sin(\theta), \cos(\theta)) \mid s \in (-\infty, \infty) \}. \]

To determine the boundary of \( D \), it would suffice to determine the point of tangency of \( l_{h_D(\theta), \omega(\theta)} \) with the boundary of \( D \); in other words, we would like to find the function \( s(\theta) \) so that for each \( \theta \),
\[ (x(\theta), y(\theta)) = h_D(\theta)(\cos(\theta), \sin(\theta)) + s(\theta)(-\sin(\theta), \cos(\theta)) \]
(1.29)
is a point on the boundary of $D$. For the remainder of this section we suppose that $s$ is differentiable.

The function $s$ is found by recalling that, at the point of tangency, the direction of the tangent line to $D$ is $\hat{\omega}(\theta)$. For a curve in the plane given parametrically by differentiable functions $(x(\theta), y(\theta))$, the direction of the tangent line is found by differentiating. At a parameter value $\theta_0$ the direction of the tangent line is the same as that of the vector $(x'(\theta_0), y'(\theta_0))$. Differentiating the expression given in (1.29) and using the fact that $\partial_\theta \omega = \hat{\omega}$, we find that

$$
(x'(\theta), y'(\theta)) = (h'_{D}(\theta) - s(\theta))\omega(\theta) + (h_{D}(\theta) + s'(\theta))\hat{\omega}(\theta).
$$

(1.30)

Since the tangent line at $(x(\theta), y(\theta))$ is parallel to $\hat{\omega}(\theta)$ it follows from (1.30) that

$$
h'_{D}(\theta) - s(\theta) = 0.
$$

(1.31)

This gives a parametric representation for the boundary of a convex region in terms of its shadow function: If the shadow function is $h_{D}(\theta)$, then the boundary of $D$ is given parametrically by

$$
(x(\theta), y(\theta)) = h_{D}(\theta)\omega(\theta) + h'_{D}(\theta)\hat{\omega}(\theta).
$$

(1.32)

Note that we have assumed that $D$ is strictly convex and the $h_{D}(\theta)$ is a differentiable function. This is not always true; for example, if the region $D$ is a polygon, then neither assumption holds.

Let $D$ denote a convex region and $h_{D}$ its shadow function. We can think of $D \mapsto h_{D}$ as a mapping from convex regions in the plane to $2\pi$-periodic functions. It is reasonable
1.2. A Simple Model Problem for Image Reconstruction

to enquire if every $2\pi$-periodic function is the shadow function of a convex region. The answer to this question is no. For strictly convex regions with smooth boundaries, we are able to characterize the range of this mapping. If $h$ is twice differentiable, then the tangent vector to the curve defined by

\[(x(\theta), y(\theta)) = h(\theta)\omega(\theta) + h'(\theta)\hat{\omega}(\theta)\]  

is given by

\[(x'(\theta), y'(\theta)) = (h''(\theta) + h(\theta))\hat{\omega}(\theta).\]

In our construction of the shadow function, we observed that the tangent vector to the curve at $(x(\theta), y(\theta))$ and the vector $\hat{\omega}(\theta)$ point in the same direction. From our formula for the tangent vector, we see that this implies that

\[h''(\theta) + h(\theta) > 0\text{ for all } \theta \in [0, 2\pi]. \]  

(1.34)

This gives a necessary condition for a twice differentiable function $h$ to be the shadow function for a strictly convex region with a smooth boundary. Mathematically we are determining the range of the map that takes a convex body $D \subset \mathbb{R}^2$ to its shadow function $h_D$, under the assumption that $h_D$ is twice differentiable. This is a convenient mathematical assumption, though in an applied context it is likely to be overly restrictive. The state space of the “system” which consists of strictly convex regions with smooth boundaries is parameterized by the set of smooth, $2\pi$-periodic functions satisfying the inequality (1.34). This is an example of a system where the constraint defining the state space is an inequality rather than an equality.

**Exercises**

**Exercise 1.2.8.** Justify the definition of $h_D(\theta + \pi)$ in (1.28) by showing that the orientation of the boundary at the point of tangency with $l_{t_0, \omega(\theta)}$ agrees with that of $l_{-t_0, \omega(\theta + \pi)}$.

**Exercise 1.2.9.** Suppose that $D_n$ is a regular $n$-gon. Determine the shadow function $h_{D_n}(\theta)$.

**Exercise 1.2.10.** Suppose that $D$ is a bounded, convex planar region. Show that the shadow function $h_D$ is a continuous function of $\theta$.

**Exercise 1.2.11.** Suppose that $h$ is a $2\pi$-periodic, twice differentiable function that satisfies (1.34). Show that the curve given by (1.33) is the boundary of a strictly convex region.

**Exercise 1.2.12.** How is the assumption that $D$ is strictly convex used in the derivation of (1.31)?

**Exercise 1.2.13.** If $h$ is a differentiable function, then equation (1.33) defines a curve. By plotting examples, determine what happens if the condition (1.34) is not satisfied.

**Exercise 1.2.14.** Suppose that $h$ is a function satisfying (1.34). Show that the area of $D_h$ is given by the

\[
\text{Area}(D_h) = \frac{1}{2} \int_0^{2\pi} \left[ (h(\theta))^2 - (h'(\theta))^2 \right] d\theta.
\]
Explain why this implies that a function satisfying (1.34) also satisfies the estimate
\[ \int_0^{2\pi} (h'(\theta))^2 \, d\theta < \int_0^{2\pi} (h(\theta))^2 \, d\theta. \]

**Exercise 1.2.15.** Let \( h \) be a smooth \( 2\pi \)-periodic function that satisfies (1.34). Prove that the curvature of the boundary of the region with this shadow function, at the point \( h(\theta)\omega(\theta) + h'(\theta)\hat{\omega}(\theta) \), is given by
\[ \kappa(\theta) = \frac{1}{h(\theta) + h''(\theta)}. \] (1.35)

**Exercise 1.2.16.** Suppose that \( h \) is a function satisfying (1.34). Show that another parametric representation for the boundary of the region with this shadow function is
\[ \theta \mapsto \left( -\int_0^\theta (h(s) + h''(s))\sin(s) \, ds, \int_0^\theta (h(s) + h''(s))\cos(s) \, ds \right). \]

**Exercise 1.2.17.** In this exercise we determine which positive functions \( \kappa \) defined on \( S^1 \) are the curvatures of closed strictly convex curves. Prove the following result: A positive function \( \kappa \) on \( S^1 \) is the curvature of a closed, strictly convex curve (parameterized by its tangent direction) if and only if
\[ \int_0^\infty \frac{\sin(s) \, ds}{\kappa(s)} = 0 = \int_0^\infty \frac{\cos(s) \, ds}{\kappa(s)}. \]

**Exercise 1.2.18.** Let \( D \) be a convex region with shadow function \( h_D \). For a vector \( v \in \mathbb{R}^2 \), define the translated region
\[ D^v = \{ x + v : x \in D \}. \]
Find the relation between \( h_D \) and \( h_{D^v} \). Explain why this answer is inevitable in light of the formula (1.35) for the curvature.

**Exercise 1.2.19.** Let \( D \) be a convex region with shadow function \( h_D \). For a rotation \( A = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \), define the rotated region
\[ D^A = \{ Ax : x \in D \}. \]
Find the relation between \( h_D \) and \( h_{D^A} \).

**Exercise 1.2.20.** If \( h_1 \) and \( h_2 \) are \( 2\pi \)-periodic functions satisfying (1.34) then they are the shadow functions of convex regions \( D_1 \) and \( D_2 \). The sum \( h_1 + h_2 \) also satisfies (1.34) and so is the shadow function of a convex region, \( D_3 \). Describe geometrically how \( D_3 \) is determined by \( D_1 \) and \( D_2 \).

**Exercise 1.2.21.** Suppose that \( D \) is non-convex planar region. The shadow function \( h_D \) is defined as before. What information about \( D \) is encoded in \( h_D \)?
Approximate Reconstructions

In a realistic situation the shadow function is measured at a finite set of angles
\[ \{\theta_1, \ldots, \theta_m\} \].

How can the data, \( \{h_D(\theta_1), \ldots, h_D(\theta_m)\} \), be used to construct an approximation to the region \( D \)? We consider two different strategies; each relies on the special geometric properties of convex regions. Recall that a convex region always lies in one of the half-planes determined by the support line at any point of its boundary. Since the boundary of \( D \) and \( h(\theta), \omega(\theta) \) have the same orientation at the point of contact, it follows that \( D \) lies in each of the half-planes
\[ H_{h(\theta_j), \omega(\theta_j)}, \quad j = 1, \ldots, m; \]
see (1.27). As \( D \) lies in each of these half-planes, it also lies in their intersection. This defines a convex polygon
\[ P_m = \bigcap_{j=1}^{m} H_{h(\theta_j), \omega(\theta_j)} \]
that contains \( D \). This polygon provides one sort of approximation for \( D \) from the measurement of a finite set of shadows. It is a stable approximation to \( D \) because small errors in the measurements of either the angles \( \theta_j \) or the corresponding affine parameters \( h(\theta_j) \) lead to small changes in the approximating polygon.

The difficulty with using the exact reconstruction formula (1.32) is that \( h \) is only known at finitely many values, \( \{\theta_j\} \). From this information it is not possible to compute the exact values of the derivatives, \( h'(\theta_j) \). We could use a finite difference approximation for the derivative to determine a finite set of points that approximate points on the boundary of \( D \):
\[ (x_j, y_j) = h(\theta_j)\omega(\theta_j) + \frac{h(\theta_j) - h(\theta_{j+1})}{\theta_j - \theta_{j+1}}\omega(\theta_j). \]
If the measurements were perfect, the boundary of \( D \) smooth and the numbers \( \{|\theta_j - \theta_{j+1}|\} \) small, then the finite difference approximations to \( h'(\theta_j) \) would be accurate and these points would lie close to points on the boundary of \( D \). Joining these points in the given order gives a polygon, \( P' \), that approximates \( D \). If \( h'(\theta_j) \) could be computed exactly, then \( P' \) would be contained in \( D \). With approximate values this cannot be asserted with certainty, though \( P' \) should be largely contained within \( D \).

This gives a different way to reconstruct an approximation to \( D \) from a finite set of measurements. This method is not as robust as the first technique because it requires the measured data to be differentiated. In order for the finite difference \( \frac{h(\theta_j) - h(\theta_{j+1})}{\theta_j - \theta_{j+1}} \) to be a good approximation to \( h'(\theta_j) \), it is generally necessary for \( |\theta_j - \theta_{j+1}| \) to be small. Moreover,
the errors in the measurements of $h(\theta_j)$ and $h(\theta_{j+1})$ must also be small compared to $|\theta_j - \theta_{j+1}|$. This difficulty arises in solution of the reconstruction problem in x-ray CT; the exact reconstruction formula calls for the measured data to be differentiated.

In general, measured data are corrupted by noise, and noise is usually non-differentiable. This means that the measurements cannot be used directly to approximate the derivatives of a putative underlying smooth function. This calls for finding a way to improve the accuracy of the measurements. If the errors in individual measurements are random then repeating the same measurement many times and averaging the results should give a good approximation to the true value. This is the approach taken in magnetic resonance imaging. Another possibility is to make a large number of measurements at closely spaced angles $\{(h_j, j \Delta \theta) : j = 1, \ldots, N\}$, which are then averaged to give less noisy approximations on a coarser grid. There are many ways to do the averaging. One way is to find a differentiable function, $H$, belonging to a family of functions of dimension $M < N$ that minimizes the square error

$$e(H) = \sum_{j=1}^{N} (h_j - H(j \Delta \theta))^2.$$ 

For example, $H$ could be taken to be a polynomial of degree $M - 1$, or a continuously differentiable, piecewise cubic function. The reconstruction formula can be applied to $H$ to obtain a different approximation to $D$. The use of averaging reduces the effects of noise but fine structure in the boundary is also blurred by any such procedure.

**Exercises**

**Exercise 1.2.22.** Suppose that the angles $\{\theta_j\}$ can be measured exactly but there is an uncertainty of size $\epsilon$ in the measurement of the affine parameters, $h(\theta_j)$. Find a polygon $P_{m,\epsilon}$ that gives the best possible approximation to $D$ and certainly contains $D$.

**Exercise 1.2.23.** Suppose that we know that $|h''(\theta)| < M$, and the measurement errors are bounded by $\epsilon > 0$. For what angle spacing $\Delta \theta$ is the error, using a finite difference approximation for $h'$, due to the uncertainty in the measurements equal to that caused by the nonlinearity of $h$ itself?

### 1.2.4 Can an Object Be Reconstructed from Its Width?

To measure the location of the shadow requires an expensive detector that can accurately locate a transition from light to dark. It would be much cheaper to build a device, similar to the exposure meter in a camera, to measure the length of the shadow region without determining its precise location. It is therefore an interesting question whether or not the boundary of a region can be reconstructed from measurements of the *widths* of its shadows. Let $w_D(\theta)$ denote the width of the shadow in direction $\theta$. A moment’s consideration shows that

$$w_D(\theta) = h_D(\theta) + h_D(\theta + \pi).$$ (1.36)
1.2. A Simple Model Problem for Image Reconstruction

Using this formula and Exercise 1.2.11, it is easy to show that \( w_D \) does not determine \( D \). From Exercise 1.2.11 we know that if \( h_D \) has two derivatives such that \( h''_D + h_D > 0 \), then \( h_D \) is the shadow function of a strictly convex region. Let \( e \) be an \textit{odd} smooth function [i.e., \( e(\theta) + e(\theta + \pi) \equiv 0 \)] such that

\[
h''_D + h_D + e'' + e > 0.
\]

If \( e \neq 0 \), then \( h_D + e \) is the shadow function for \( D' \), a different strictly convex region. Observe that \( D' \) has the same \textit{width} of shadow for each direction as \( D \); that is,

\[
w_D(\theta) = (h_D(\theta) + e(\theta)) + (h_D(\theta + \pi) + e(\theta + \pi)) = w_{D'}(\theta).
\]

To complete this discussion, note that any function expressible as a series of the form

\[
e(\theta) = \sum_{j=0}^{\infty} [a_j \sin(2j + 1)\theta + b_j \cos(2j + 1)\theta]
\]

is an odd function. This is an infinite-dimensional space of functions. This implies that if \( w_D \) is the width of the shadow function for a convex region \( D \), then there is an infinite-dimensional set of regions with the same width of the shadow function. Consequently, the simpler measurement is inadequate to reconstruct the boundary of a convex region. Figure 1.14 shows the unit disk and another region that has constant shadow width equal to 2.

![Figure 1.14. Two regions of constant width 2.](image)

**Exercises**

**Exercise 1.2.24.** Justify the formula (1.36) for the shadow width.

**Exercise 1.2.25.** Show that the width function satisfies \( w''_D + w_D > 0 \).

**Exercise 1.2.26.** Is it true that every twice differentiable, \( \pi \)-periodic function, \( w \) satisfying \( w'' + w > 0 \) is the width function of a convex domain?
Exercise 1.2.27. We considered whether or not a convex body is determined by the width of its shadows in order use a less expensive detector. The cheaper detector can only measure the width of the covered region. Can you find a way to use a detector that only measures the length of an illuminated region to locate the edge of the shadow? Hint: Cover only half of the detector with photosensitive material.

1.3 Conclusion

By considering examples, we have seen how physical systems can be described using mathematical models. The problem of determining the state of the system from measurements is replaced by that of solving equations or systems of equations. It is important to keep in mind that mathematical models are just models, indeed often toy models. A good model must satisfy two opposing requirements: The model should accurately depict the system under study while at the same time being simple enough to be usable. In addition, it must also have accurate, finite-dimensional approximations.

In mathematics, problems of determining the state of a physical system from feasible measurements are gathered under the rubric of inverse problems. The division of problems into inverse problems and direct problems is often a matter of history. Usually a physical theory that models how the state of the system determines feasible measurements preceded a description of the inverse process: how to use measurements to determine the state of the system. While many of the problems that arise in medical imaging are considered to be inverse problems, we do not give a systematic development of this subject. The curious reader is referred to the article by Joe Keller, [80], which contains analyses of many classical inverse problems or the book Introduction to Inverse Problems in Imaging, [10].

The models used in medical imaging usually involve infinitely many degrees of freedom. The state of the system is described by a function of continuous variables. Ultimately only a finite number of measurements can be made and only a finite amount of time is available to process them. Our analysis of the reconstruction process in x-ray CT passes through several stages. We begin with a description of the complete, perfect data situation. The measurement is described by a function on the space of lines. By finding an explicit inversion formula, we show that the state of the system can be determined from these measurements. The main tool in this analysis is the Fourier transform. We next consider the consequences of having only discrete samples of these measurements. This leads us to sampling theory and Fourier series. In the next chapter we quickly review linear algebra and the theory of linear equations, recasting this material in the language of measurement. The chapter ends with a brief introduction to the issues that arise in the extension of linear algebra to infinite-dimensional spaces.