1

Introduction

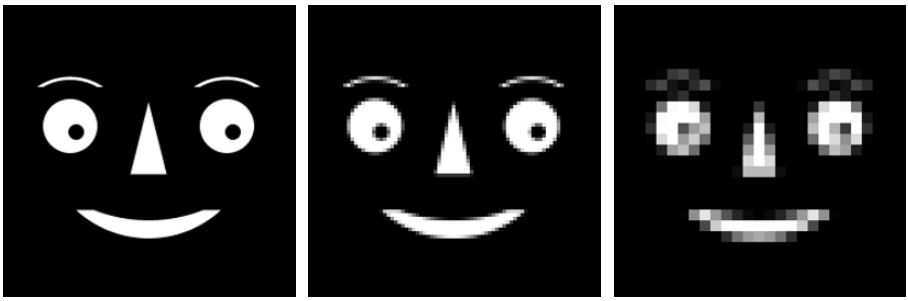
In this chapter we discuss some of the many different medical, engineering, and scientific areas in which the procedures described in this book are applicable. We also introduce some statistical concepts that are useful for understanding much of the later material.

1.1 Image Reconstruction from Projections

The problem of image reconstruction from projections has repeatedly arisen over the last 50 years in a large number of scientific, medical, and technical fields. The range of applicability is staggering. At one end of the scale, data from electron microscopes are used to reconstruct molecular structures; while at the other end, data from radio telescopes are used to reconstruct maps of radio emission from celestial objects. These seemingly different applications, and many others to be mentioned here, have the same mathematical and computational foundations. It is the purpose of this book to discuss these foundations.

Of all the applications, probably the greatest effect on the world at large has been in the area of diagnostic medicine: *computerized tomography* (CT) has revolutionized radiology. Images of cross sections of the human body are produced from data obtained by measuring the attenuation of x-rays along a large number of lines through the cross section. Most of this book uses CT as the framework within which the problems and solutions are presented. We therefore say very little about it in this section, but survey some of the numerous other applications.

We start with a simple artificial problem to demonstrate the underlying ideas. While the solution to this problem is of no known practical usefulness, the problem is very similar to a practical problem in astrophysics (to be mentioned in the following), and shares its basic structure with other applications of image reconstruction from projections.



(a) (b) (c)

Fig. 1.1: Three different digitizations of the same picture: (a) is a 243×243 digitization, (b) is an 81×81 digitization, and (c) is a 27×27 digitization.

Suppose that we have a rectangular area containing some sources of light. A simple example is a television screen displaying a still picture. Suppose that we also have a "detector" that can measure the *total* intensity of light in the picture. That, of course, would not help us to record the details in the picture. One way of getting at the details is to make a "collimator," by cutting a small square hole into a sheet of nontransparent material. If we put the collimator in front of the picture, the detector measures only the light emanating from the small region behind the square hole. By moving the hole in discrete steps across the picture and measuring the intensity each time, we can build up an image of the picture. The image is made up of small square regions whose brightness is proportional to the average intensity in the original picture in the corresponding region. We can move the collimator so that the small square regions are abutting and cover the whole picture. In such a case, the resulting image (referred to as a *digitization* of the picture later in this book) resembles the picture, provided only that the collimator's hole is small enough. This is illustrated in Fig. 1.1.

Suppose now that we lack the capability of cutting a small square hole into our opaque sheet. It may then appear that we can no longer produce an image of our picture. However, image reconstruction from projections comes to our rescue. We now illustrate the processes of "projection taking" and "reconstruction" on our simple problem.

The process of projection taking in this case consists of moving the opaque sheet across the picture in small discrete steps in a fixed direction. After each move, we use our detector to measure the total intensity of light in the uncovered part of the picture. Subtraction of the measured value of the total intensity at any time instance from the measured value of the total intensity at the next time instance provides us with the total intensity of light in each of a set of parallel abutting thin strips of known location (see Fig. 1.2). We can now repeat this process with the opaque sheet moving in a different direction. This way we get the total intensity of light in each of another set of

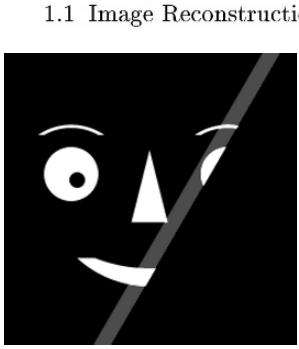
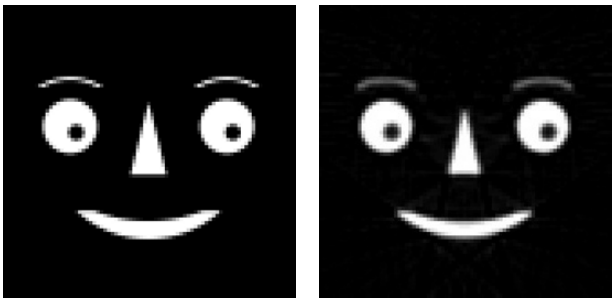


Fig. 1.2: The process of projection taking. The line integral of the brightness along the central line of a strip (shown half-illuminated) is estimated by dividing the total brightness in the strip by the width of the strip.

parallel abutting thin strips of known location. We estimate the *line integral* of the brightness along the central lines of these strips, by dividing the total brightness in the strip by the width of the strip. Doing this repeatedly (say 90 times, rotating the orientation of the opaque sheet by 2^o each time), we obtain many such sets of measurements. Each set of estimated line integrals is often called a *projection*, but in this book we preferentially use the word *view*, since "projection" is also used for a number of other things. The collection of all the estimated line integrals is referred to as the *projection data*.

The process of *reconstruction* produces an image of the picture from projection data of the picture. How this is done is the main topic of this book. In Fig. 1.3 we compare the 81×81 digitization of a picture with an 81×81 reconstruction from 90 views with 121 estimated line integrals in each.

This example illustrates the informal definition given in the following paragraph. While not all that comes under the heading "image reconstruction from projections" is covered by this informal definition (for example, in Chapter 13



(a) (b)

Fig. 1.3: The 81×81 digitization of a picture (a) compared to an 81×81 reconstructions from 90 views with 121 measurements in each view (b).

we discuss truly three-dimensional reconstruction), it is adequate for describing our attitude toward image reconstruction throughout most of this book.

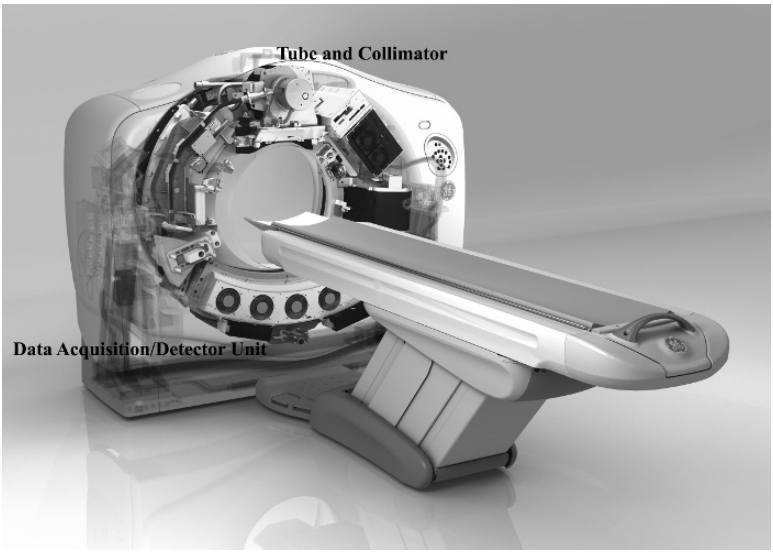
Image reconstruction from projections is the process of producing an image of a two-dimensional distribution (usually of some physical property) from estimates of its line integrals along a finite number of lines of known locations.

We now turn to some real-life applications. The order in which we take these is according to the size of the object to be reconstructed. The reader is warned that the following is by no means an exhaustive survey of all the application areas of image reconstruction from projections.

Actually, our simple artificial problem has a close analog in *astrophysics*. There are instruments for measuring the brightness distribution of radio sources in the sky that are of too low resolution to provide astrophysicists with the information they seek. However, if the moon moves across the portion of the sky that is of interest, it acts in an analogous fashion to the opaque sheet of our artificial example. The directions of the paths of the moon across the sky vary, providing us with a number of views, which in this field are referred to as profiles obtained from *lunar occultation* observations. From such observations the two-dimensional brightness distribution of radio sources can be reconstructed. Among the other applications of image reconstruction we mention its use for discovering the x-ray structure of supernova remnants and the electron-density distribution in the solar corona. In the latter case, display techniques allow us to make movies of the dynamic changes in the solar corona as would be observed from above the north pole of the sun's rotation, a view that cannot possibly be observed from earth! Coming down to earth, we note that there are numerous applications, such as applying tomographic methods to geodesy and to volcanology. However, we concentrate on the application in which image reconstruction from projections is probably applied more frequently than in any other, namely diagnostic medicine.

X-ray *transmission computerized tomography* (CT) is discussed in some detail in the succeeding chapters. Here we just indicate its nature using a few illustrations. Figure 1.4(a) shows a photograph of an x-ray CT scanner and Fig. 1.4(b) shows an engineering drawing of an apparatus for data collection in x-ray CT. The tube contains a single x-ray source, the detector unit contains an array of x-ray detectors. Suppose for the moment that the x-ray Tube and Collimator on the one side and the Data Acquisition/Detector Unit on the other side are stationary, and the patient on the table is moved between them at a steady rate. By shooting a fan beam of x-rays through the patient at frequent regular intervals and detecting them on the other side, we can build up a two-dimensional x-ray projection of the patient that is very similar in appearance to the image that is traditionally captured on an x-ray film. Such a projection is shown in Fig. 1.5(a). The brightness at a point is indicative of the total attenuation of the x-rays from the source to the detector. This mode of operation is *not* CT, it is just an alternative way of taking x-ray images. In the CT mode, the patient is kept stationary, but the tube and the detector unit rotate (together) around the patient. The fan beam of x-rays from the

(a)



(b)

Fig. 1.4: (a) A CT scanner of the LightSpeed Series of GE Healthcare. (Illustration provided by C. Yee of Jacobi Medical Center.) (b) Engineering rendering of a CT scanner released in 2008. (Photo provided by GE Healthcare.)

source to the detector determines a slice in the patient's body. The location of such a slice is shown by the horizontal line in Fig. 1.5(a). Data are collected for a number of fixed positions of the source and detector; these are referred

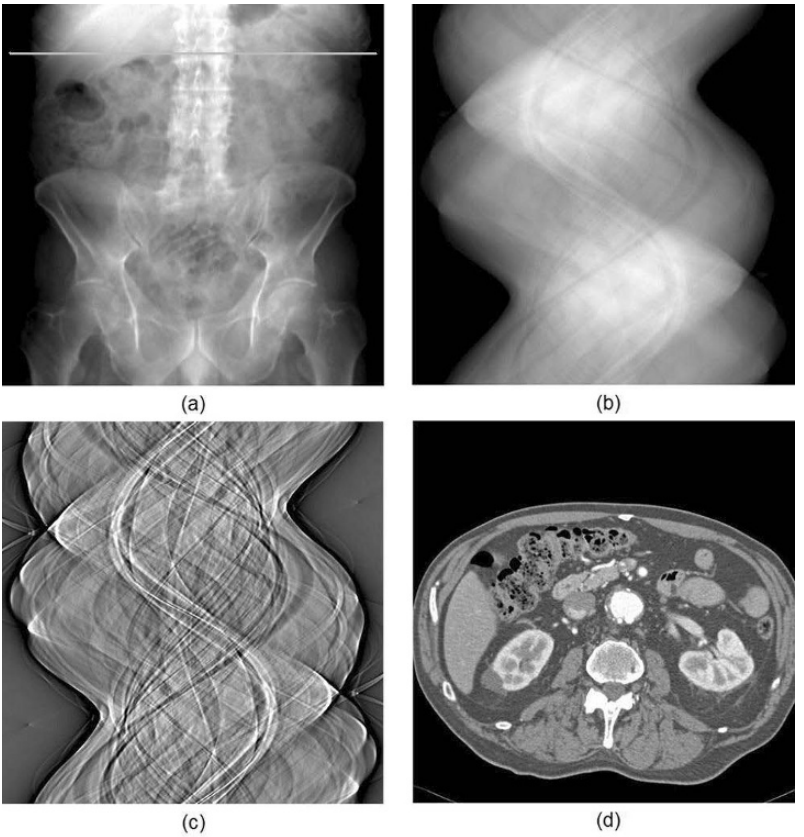


Fig. 1.5: (a) A digitally rendered (x-ray) radiograph with a horizontal line marking the location of the cross section for which the following images were obtained. (b) Sinogram of the projection data. (c) Sinogram of the convolved projection data. (d) A reconstruction from the projection data. (All images were obtained using a Siemens Sensation CT scanner by R. Fahrig and J. Starman at Stanford University.)

to as *views*. For each view, we have a reading by each of the detectors. All the detector readings for all the views can be represented as a *sinogram*, shown in Fig. 1.5(b). The intensities in the sinogram are proportional to the line integrals of the x-ray attenuation coefficient between the corresponding source and detector positions. From these line integrals, a two-dimensional image of the x-ray attenuation coefficient distribution in the slice of the body can be produced by the techniques of image reconstruction. Such an image is shown in Fig. 1.5(d). Inasmuch as different tissues have different x-ray attenuation coefficients, boundaries of organs can be delineated and healthy tissue can be distinguished from tumors. In this way CT produces cross-sectional slices of the human body without surgical intervention. (The picture in Fig. 1.5(c)

is a sinogram of the "convolved projection data," which is to be defined in (10.12).)

In addition to providing an excellent tool for diagnosis, the images produced by CT can be used for the planning of radiation therapy, in which beams of penetrating radiation are directed at malignancies in the body with the aim of destroying them. The goal is to deliver a sufficiently high dose to target volumes, such as tumor cells, but to avoid depositing a harmfully high dose to organs at risk. Modern devices use multileaf collimators that allow the treatment planner to control the intensity of radiation within relatively thin beams; this is referred to as *intensity modulated radiation therapy* (IMRT), see Fig. 1.6. The mathematical problem of IMRT can be considered to be

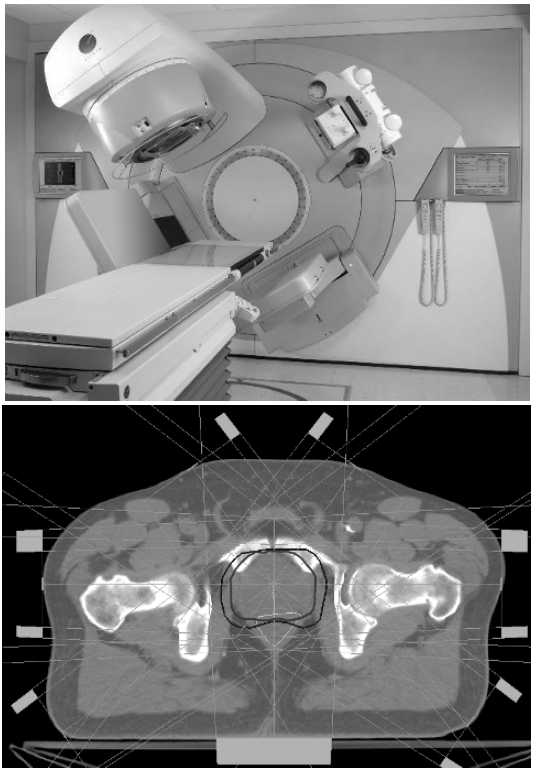


Fig. 1.6: Top: A treatment machine for intensity modulated radiation therapy (Elekta Synergy^{TM}) with multileaf collimator and a CT device for online image guided radiation therapy. (Illustration provided by Elekta, Inc.) Bottom: A CT slice of a cancer patient. Six beam angles are used and the treatment is planned so that the 75.6 Gy isodose line (in black) covers the target volume (in gray), but it bends so as to avoid the greater part of the rectum (outlined in white just below the center of the image). (Illustration provided by Y. Xiao of Thomas Jefferson University.)

"dual" to that of CT: in CT we try to recover the distribution of the x-ray attenuation in the body from measurements of total attenuation within thin beams of x-rays, in IMRT we are given information regarding the desired dose distribution in the body and we need to calculate the radiation intensity that needs to be sent into the body along thin beams in order to achieve such a dose distribution. Some of the mathematical techniques discussed later are applicable to both problems.

Another method of extreme usefulness in diagnostic medicine is *magnetic resonance imaging* (MRI). When an object is placed in a magnetic field gradient, the frequencies of magnetic resonance signals from its nuclei and unpaired electrons depend upon the value of the local applied magnetic field, as well as upon those molecular interactions usually studied by magnetic res-

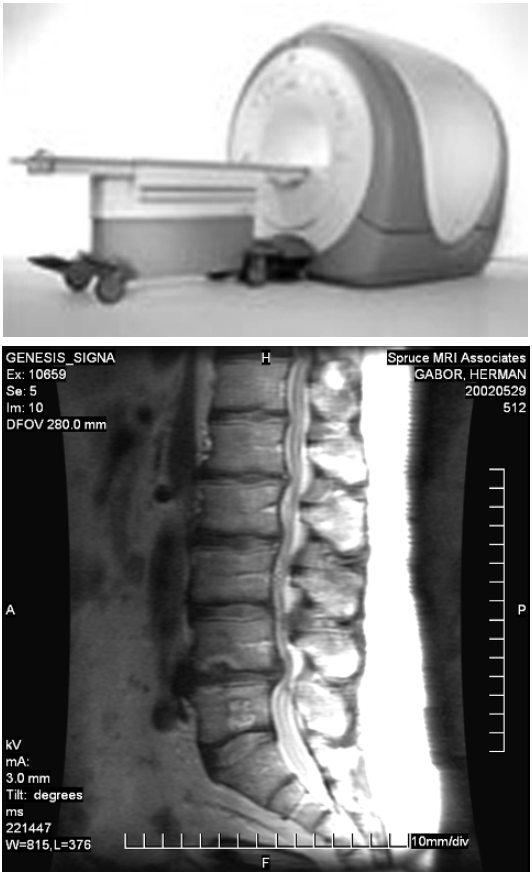


Fig. 1.7: Top: An MRI scanner (GE Healthcare Signa HD 1.5T). Bottom: A sagittal slice through the spine of the author obtained by such a scanner. Note the bulging disc pressing on the spinal nerve; this required surgical intervention.

onance methods. The integrated signal from the intersection of a surface of constant magnetic field with a three-dimensional object is one point on a one-dimensional projection of a three-dimensional signal. In a uniform linear field gradient, a plot of such signals against frequency is a one-dimensional projection, in a direction perpendicular to the gradient axis, of the total signal intensity. If the direction of the field gradient is varied other projections may be produced, and a two- or three-dimensional image may be reconstructed. However, such a "projection imaging" approach is not frequently used in practice: MRI usually relies on collecting data regarding the Fourier transform of the object to be imaged and then inverting this Fourier transform. (The associated mathematics is discussed below in Sections 8.4 and 9.1.) An MRI scanner and its output are illustrated in Fig. 1.7.

Emission computerized tomography has as its major emphasis the quantitative determination of the moment-to-moment changes in the chemistry and flow physiology of injected or inhaled compounds labeled with radioactive atoms. In this case the distribution to be reconstructed is the distribution of radioactivity in the body cross section, and the measurements are used to estimate the total activity along lines of known locations. Figure 1.8 illustrates

Fig. 1.8: A PET scanner that has 17,864 scintillation crystals to collect its data. Whole-body imaging is performed by acquiring seven to eight data sets (of approximately three-minutes duration each) with bed motion between acquisitions. (Illustration provided by J. Karp of the University of Pennsylvania.)

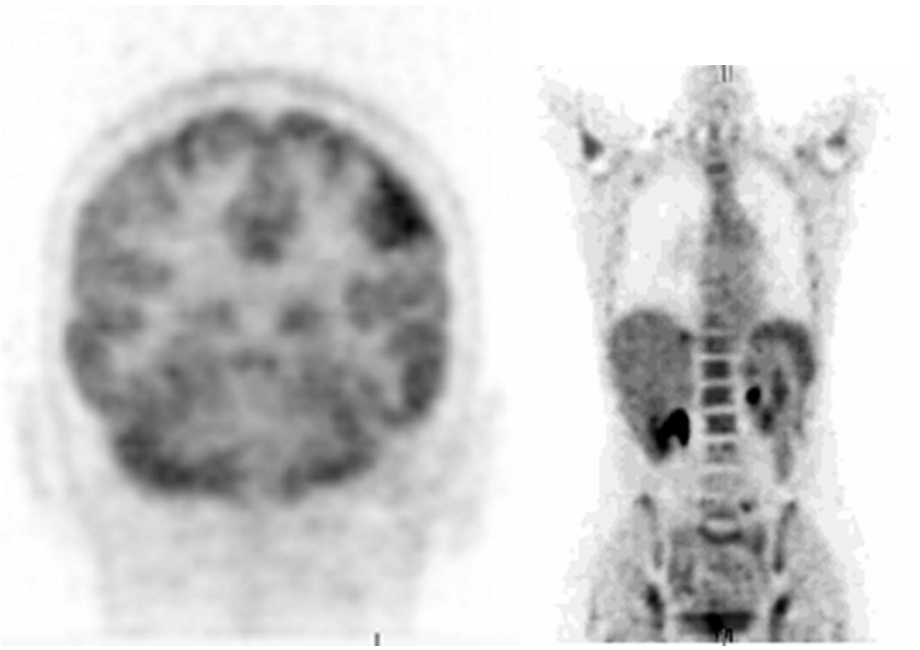


Fig. 1.9: Coronal sections of reconstructions of patients who have been injected with fluoro-deoxy-glucose (FDG) tagged with a radioactive isotope whose decay generates positrons that annihilate and produce pairs of gamma rays to be detected by the crystals to identify a line that contains the location of the annihilation. From the total activity along a large number of such lines the distribution of the annihilation frequency, and hence of the FDG, can be reconstructed. Left: Brain scan of an epilepsy patient; see the increased FDG uptake (dark) at about two o-clock, indicating the seizure focus site. Right: Whole-body scan of a patient with melanoma; see the increased FDG uptake in a small spot near the top, anterior part of the liver, indicating a lesion. (Images were obtained by the Philips Allegro scanner shown in Fig. 1.8 and were provided by J. Karp of the University of Pennsylvania.)

a device, a so-called *positron emission tomography* (PET) scanner, for doing this and Fig. 1.9 shows two clinical images produced by this device. As explained in the caption of that figure, the device allows us to image how the compound (in this case FDG) distributes itself in the body; if the compound uptake is increased in lesions, then the images can be used for locating such lesions.

Both x-ray CT and emission computerized tomography use potentially harmful ionizing radiation for their data collection, but this is not the case for MRI. Another modality of data collection with no demonstrated adverse effect on the patient is *ultrasound*.

In Fig. 1.10, we show a photograph of an apparatus used to collect ultrasonic data regarding the female breast. From these data one obtains three

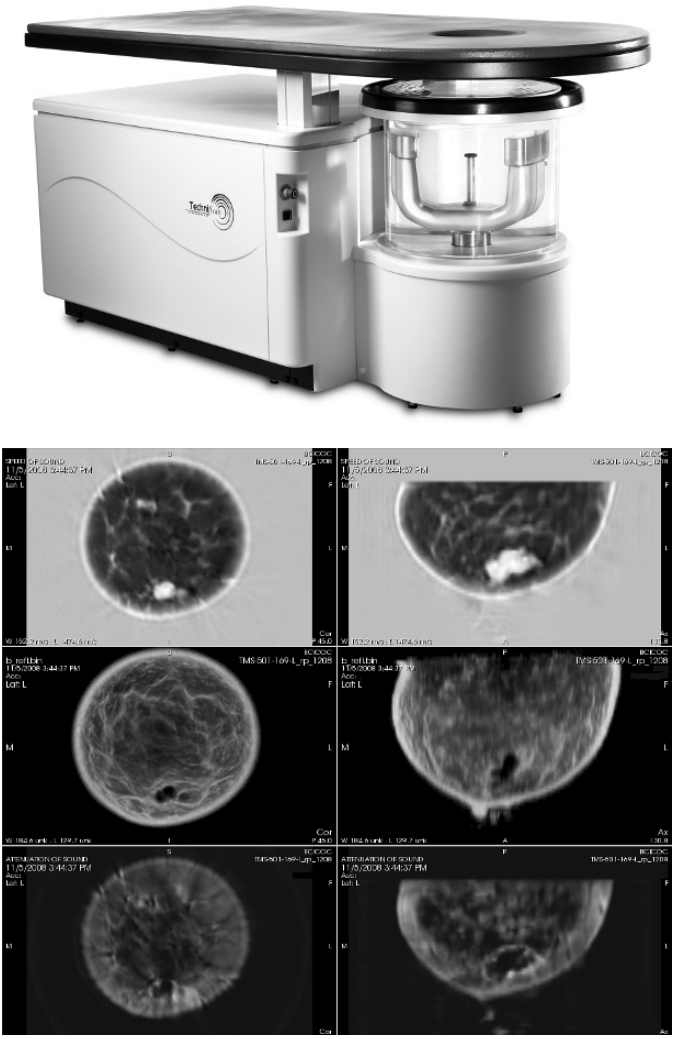


Fig. 1.10: Top: A Whole Breast Ultrasound (WBU^{TM}) scanner (TechniScan Medical Systems) used for collecting data to reconstruct breast images. Bottom: A screenshot of the speed of sound (top), reflection (middle) and the attenuation of sound (bottom) of a complex cyst, shown in coronal (left column) and axial (right) correlated slices. (Illustrations provided by TechniScan Medical Systems.)

separate but correlated reconstructions, providing different tissue characteristics. The three reconstructions looked at together give a great deal of information about the nature of the tumors (if any) present. The figure also shows

a screenshot on the scanner's viewer of the coronal slices (left column) and axial slices (right column) in the three 3D reconstructions.

Another method for tomographic imaging the body uses light. We give an illustration of such *optical tomography* in Fig. 1.11.

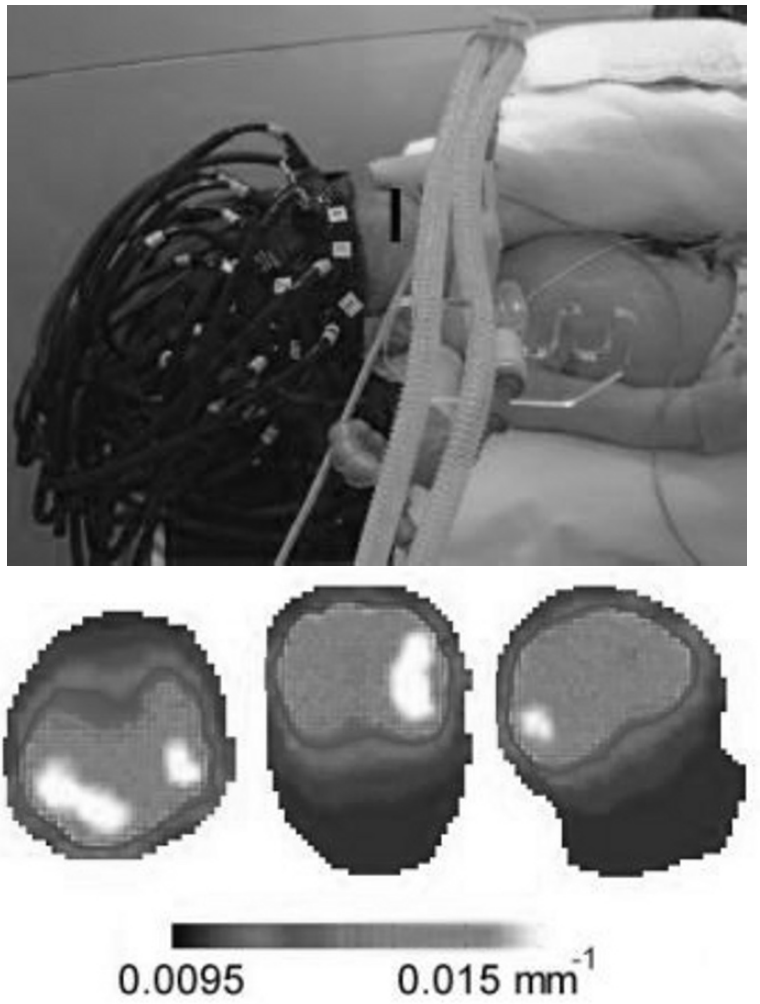


Fig. 1.11: Optical tomography. Top: A four-day old infant is being optically imaged using 29 source-detector pairs around her head. Bottom: Slices across the reconstructed 3D images of differences in light absorption at the the wavelength of 815 nm, caused by differences in blood distribution. Sequences of such images can be used to assess physiological functioning. (Illustration provided by S. Arridge of the University College London.)

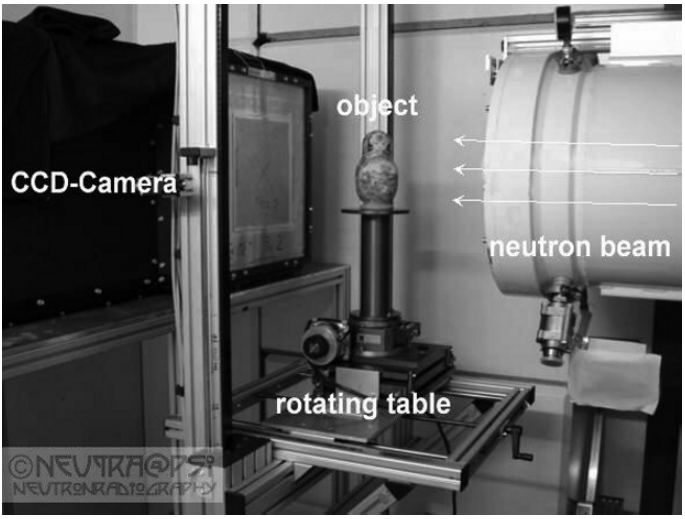
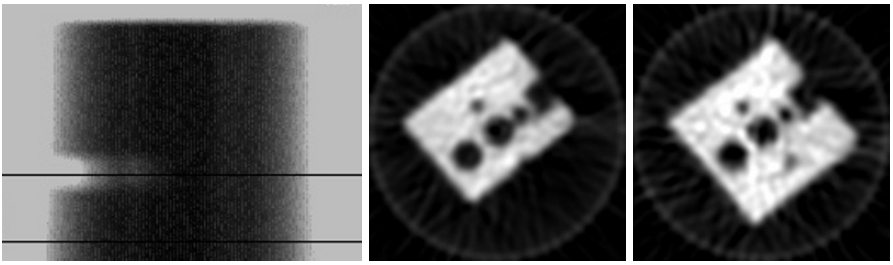


Fig. 1.12: Apparatus for neutron tomography data collection. (Illustration provided by the Neutron Imaging and Activation Group, Paul Scherrer Institute, Switzerland, http://neutra.web.psi.ch.)

Getting away from medicine, we note that image reconstruction from projections has been found useful in *nondestructive testing*. For example, a collection of transmission beam neutron radiographs can be used for the reconstruction (and hence inspection) of such objects as turbine blades and even whole engines. Such metallic objects would not be well penetrated by x-rays. Figure 1.12 shows a setup for collecting the necessary data. A typical neutron radiograph collected by such a device is shown in Fig. 1.13, together with two slices through the 3D reconstruction from multiple such radiographs.



(a) (b) (c)

Fig. 1.13: (a) A neutron radiograph with the location of two cross sections indicated. (b) and (c) Reconstructions of the indicated cross sections. (Illustration provided L. Ruskó of the University of Szeged, Hungary.)

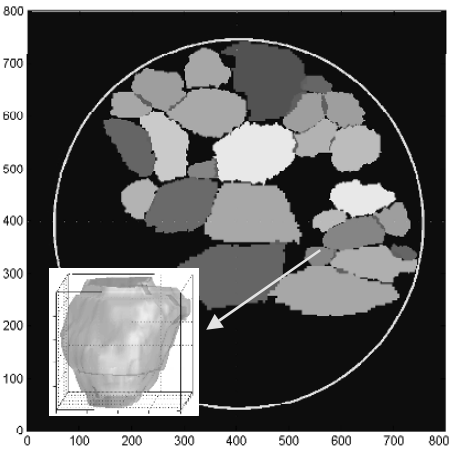


Fig. 1.14: Grains in a polycrystal. Multiple grains in a cross section are shown, a single one of them is displayed three-dimensionally. (Illustration provided by H.F. Poulsen of the Risø National Laboratory, Denmark.)

An emerging application of image reconstruction from projections is in *materials science*. In nature most materials such as rocks, ice, sand, and soil appear as aggregates comprised of a set of small crystals. Similarly, modern society is built on applications of metals, ceramics and other hard materials, which are also polycrystalline. An example of a *polycrystal* is shown in Fig. 1.14. The individual crystals are known as *grains*. Each grain is characterized by its position and shape as well as by the *orientation* of the 3D *crystalline lattice* (the discrete lattice of atom positions). The latter property is known as the *grain orientation*. The physical, chemical and mechanical properties of the material are to a large extent governed by the geometrical features of this 3D complex.

Three-dimensional x-ray diffraction (3DXRD) is one way to collect data for recovering the distribution and orientation of grains; see Fig. 1.15. The method is based on reconstruction using x-rays with a setup similar to that of CT. The vital difference is that in CT the absorption of the incident beam through the sample is probed, while in 3DXRD the diffracted beam is probed as it diverges from the sample on the exit side. The *diffraction pattern* on the detector typically is composed of a set of distinct *diffraction spots*. Acquiring images at a set of rotation angles, each grain gives rise to 5 to 30 spots, with positions and intensity distributions determined by the local orientation of the crystalline lattice. From such data it is possible to reconstruct not only the geometry of the grains, but also the variation of the orientations within the grains. Since orientations need three variables to specify them, an interesting aspect of such reconstructions is that what is reconstructed is an image of a distribution of three-dimensional vectors; see Fig. 1.16.

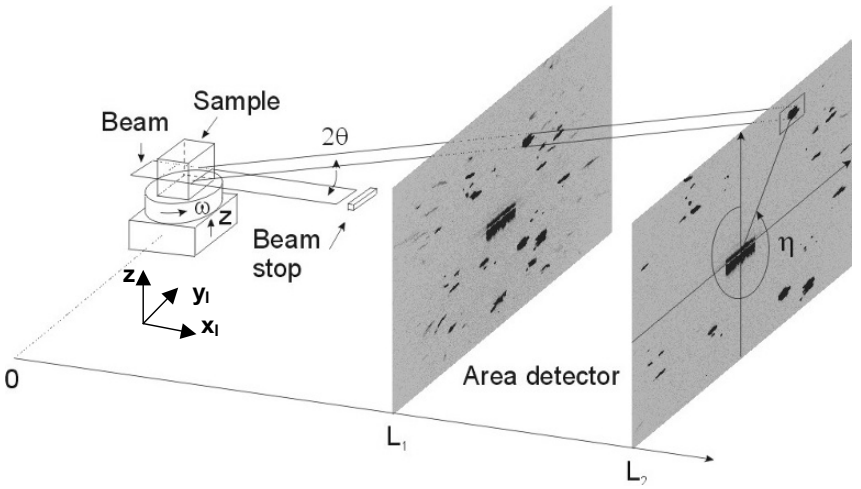


Fig. 1.15: The 3DXRD data collection geometry. Detectors are positioned perpendicular to the beam at various distances. (Illustration provided by H.F. Poulsen of the Risø National Laboratory, Denmark.)

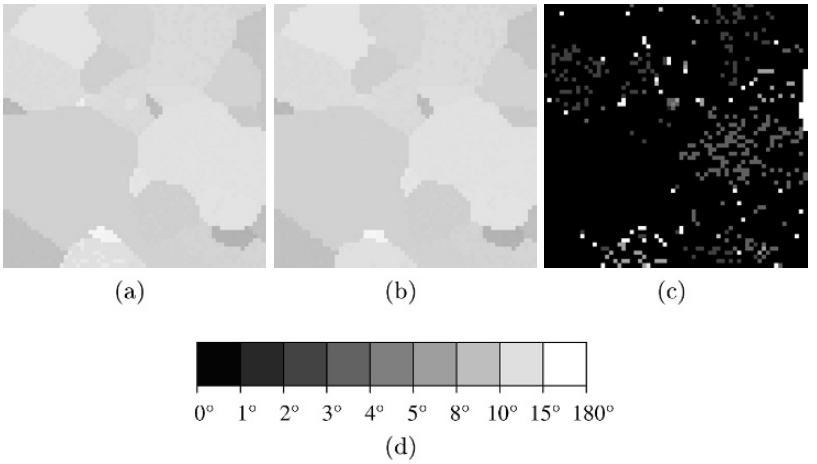


Fig. 1.16: (a) A test pattern of polycrystal orientations (obtained by electron microscopy). (b) A reconstruction from simulated noisy diffraction data. (c) Differences between the test and the reconstructed orientation distributions. (d) Gray scale indicating the angles in the difference map. Note that at nearly everywhere the difference angle is less than 15° and in the large majority of cases it is less than 1°.

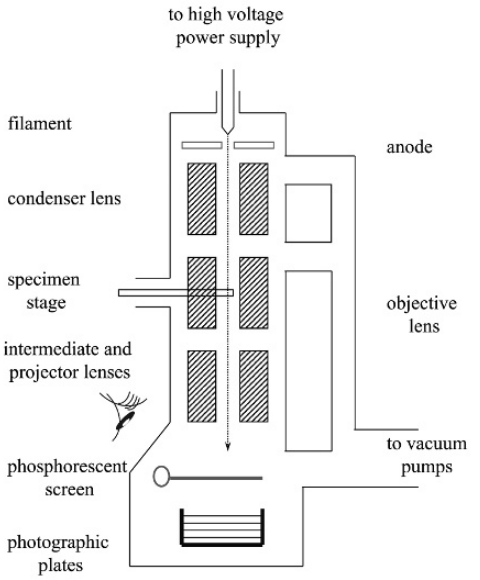


Fig. 1.17: Schematic drawing of a transmission electron microscope. (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain.)

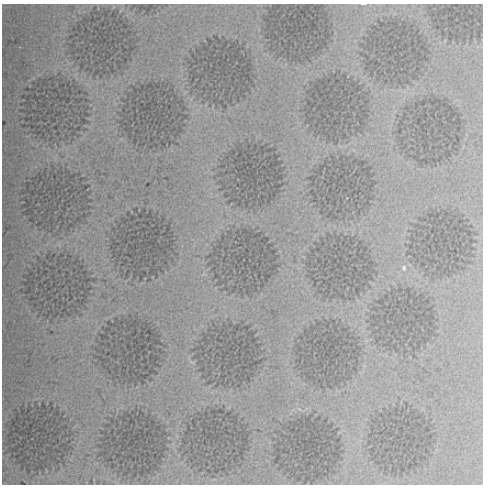


Fig. 1.18: Part of an electron micrograph containing projections of multiple copies of the human adenovirus type 5. (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain.)

Three-dimensional reconstruction of nano-scale objects (such as biological macromolecules) can be accomplished using data recorded with a transmission *electron microscope* (see Fig. 1.17) that produces *electron micrographs*, such as the one illustrated in Fig. 1.18, in which the grayness at each point is indicative of a line integral of a physical property of the object being imaged. From multiple electron micrographs one can recover the structure of the biological object that is being imaged; see Fig. 1.19.

This completes our survey of some of the applications of image reconstruction from projections. Except for some further discussion of x-ray CT, the rest of this book is devoted to the theory, rather than applications, of image reconstruction.

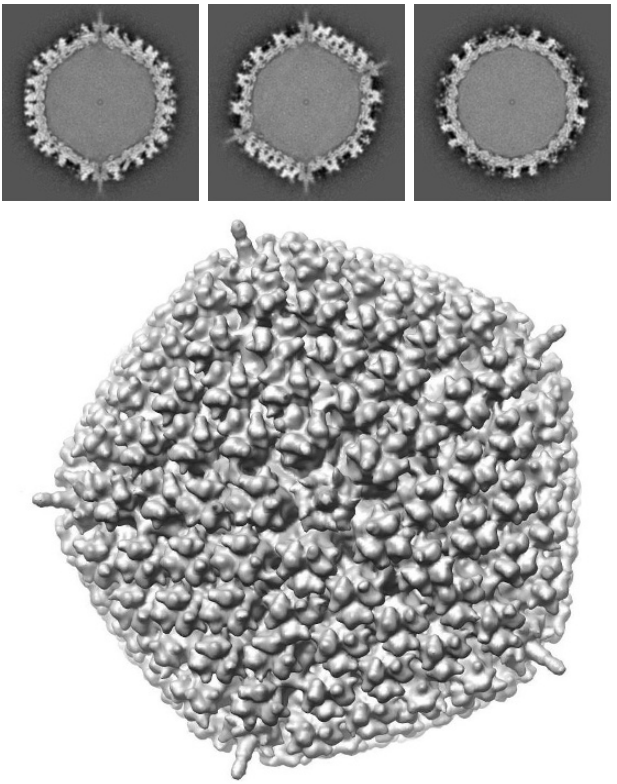


Fig. 1.19: Top: Reconstructed values, from electron microscopic data such as in Fig. 1.18, of the human adenovirus type 5 in three mutually orthogonal slices through the center of the reconstruction. Bottom: Computer graphic display of the surface of the virus based on the three-dimensional reconstruction. (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain.)

1.2 Probability and Random Variables

In order to discuss the processes involved in CT we need to know some of the basic concepts of probability theory. That is the purpose of this section. The reader may wish to skim it at first reading and return to it at times when the notions introduced here are actually used.

As an example, consider the situation depicted in Fig. 1.20. There is a slab of material and the line L goes through it. If an x-ray photon enters the slab along the line L through its top face, it will continue to travel along the line L until it is absorbed or scattered. Some photons will be neither absorbed nor scattered before exiting through the bottom face. We shall say such photons are *transmitted* through the slab. The point is that for any individual x-ray photon we cannot be certain whether or not it will be transmitted. All we can say is that, for any fixed energy $\bar{e}$, there is a fixed *probability* ρ that a photon at that energy that enters the slab is transmitted. We call ρ the *transmittance* at energy $\bar{e}$ of the slab along line L, and we define it as follows. The definition is typical of how the "probability" of something happening is defined.

To define ρ we carry out a "thought experiment." Such an experiment could be physically carried out if we had at our disposal an infinite amount of time and instruments with unlimited precision. We shoot photons at energy $\bar{e}$ one by one through the slab along the line L, and we test whether they are transmitted. Let $t_1(n)$ denote the number of photons transmitted out of the first n in this experiment. (The subscript 1 refers to the fact that this is the first such thought experiment, in the following we discuss a whole series of them.) Then ρ is defined as the limit of $t_1(n)/n$ as n tends to infinity:

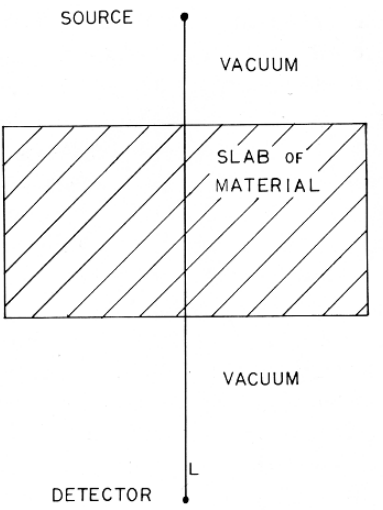


Fig. 1.20: Definition of transmittance.

$$\rho = \lim_{n\to\infty} (t_1(n)/n)\,. \tag{1.1}$$

That is, the transmittance ρ is a number such that given a positive real number ε, however small, there will always be an integer n_0, such that the difference between ρ and $t_1(n)/n$ is less than ε for all n greater than n_0.

Note that it is not a priori obvious that $t_1(n)/n$ has a limit as n tends to infinity. The claim that it does is based on physical experiments that approximate the thought experiment just described. Note also that it is assumed that the same value of ρ will be provided if the experiment is carried out again. More precisely, let the same thought experiment be carried out the second time, and let $t_2(n)$ denote the number of photons transmitted out of the first n in the second experiment. Then $\rho = \lim_{n\to\infty} (t_2(n)/n)$.

Even though the values of ρ defined by the limits of the two identical thought experiments are the same, it does not mean that we can assume that $t_1(n) = t_2(n)$, for any fixed n. Both $t_1(n)$ and $t_2(n)$ may assume any integer value between zero (no photons are transmitted) and n (all photons are transmitted). However, some of these values are more likely than others. It is reasonable to inquire as to what is the probability $p_{n,\rho}(m)$ that m photons out of n get transmitted.

To define $p_{n,\rho}(m)$, we carry out the previously described thought experiment repeatedly up to the point when n photons have entered the slab. Let $t_i(n)$ denote the number of transmitted photons in the ith thought experiment. Let $s(N)$ denote the number of times $t_i(n) = m$, for $1 \le i \le N$. Then

$$p_{n,\rho}(m) = \lim_{N\to\infty} (s(N)/N)\,. \tag{1.2}$$

Note that $p_{n,\rho}(m) = 0$ if m is negative or if m is greater than n. Since $p_{n,\rho}(m)$ is supposed to be the probability of m photons being transmitted out of n, this is reassuring. Also it is easy to see, by comparing the thought experiments used to define ρ and $p_{n,\rho}(m)$, that $p_{1,\rho}(1) = \rho$. It is somewhat more difficult to show that, in general for $0 \le m \le n$,

$$p_{n,\rho}(m) = \frac{n!}{m!(n-m)!}\rho^m(1-\rho)^{n-m}, \tag{1.3}$$

where, as usual, $m!$ denotes $m \times (m-1) \times (m-2) \times \cdots \times 2 \times 1$, with $0!$ defined to be one. Equation (1.3) is referred to as the *binomial probability law*. The values of $p_{30,0.7}(m)$, for $0 \le m \le 30$, are plotted in Fig. 1.21.

More generally, if S_X is a finite or a countably infinite set (such as the set all integers) of all possible outcomes of an experiment, then S_X together with the probability $p_X(x)$ of the outcome for each x in S_X is referred to as the *discrete random variable* X. Mathematically, we must have that $p_X(x) \ge 0$ for all x in S_X, and that the sum of the $p_X(x)$ over S_X is 1. In this book, we only allow experiments whose outcome is a number, or possibly a column vector of numbers. For example, for a fixed n, the set of all integers m together with the probability $p_{n,\rho}(m)$ is the *binomial random variable* with parameters n and ρ.

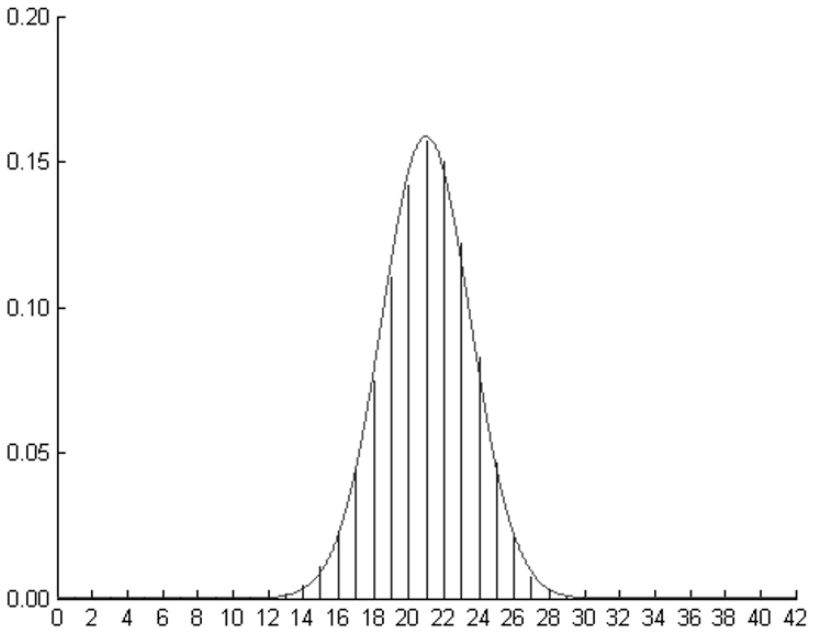


Fig. 1.21: Plot of the binomial probability function $p_{30,0.7}$, see (1.3), with values indicated by the vertical lines and of the Gaussian probability density function, see (1.10), with $\mu_X = 21$ and $V_X = 6.3$, shown as a continuous curve.

We call the outcome of a single experiment a *sample* of the random variable. For example, the value of $t_i(n)$, for a fixed i, in the thought experiment above is a sample of the binomial random variable with parameters n and ρ.

In summary, the number of photons that may be transmitted when n photons enter the slab is a discrete random variable. The actual number of photons transmitted in a single experiment with n photons entering the slab is a sample of the random variable. Later on we will see other examples of discrete random variables. A particularly important one for our purposes is associated with the number of photons emitted by an x-ray source in the direction of a detector during a unit period of time.

Two important properties of a discrete random variable X are its *mean* μ_X and its *variance* V_X, defined by

$$\mu_X = \sum_{x \in S_X} x p_X(x), \tag{1.4}$$

$$V_X = \sum_{x \in S_X} (x - \mu_X)^2 p_X(x). \tag{1.5}$$

Note that the averaged outcome of a very large number of experiments will approximate the mean. Also, the variance will be approximated by taking the average of the squares of the distances of the samples from the mean. Thus, the variance is a measure of the spread of the possible outcomes around the mean. For the binomial random variable with parameters n and ρ, the mean is $n\rho$ and the variance is $n\rho(1-\rho)$. The *standard deviation* σ_X of the random variable X is defined to be the nonnegative square root of its variance.

We introduce one more notion for the special case when all elements of S_X are real numbers. Let a denote either $-\infty$ or a real number and let b denote either a real number or ∞, such that $a < b$. Then we define

$$P_X(a,b] = \sum_{\substack{x \in S_X \\ a < x \leq b}} p_X(x). \tag{1.6}$$

Thus, $P_X(a,b]$ denotes the probability that a sample of the discrete random variable X is in the interval $(a,b]$.

We now discuss the notion of a *continuous random variable* X for the special case when the associated set S_X of possible outcomes is the set of all real numbers. In this case we use a *probability density function* p_X that maps S_X into the range $[0,1]$ (the closed interval of real numbers from 0 to 1) in such a way that the probability that a sample of X is in the interval $(a,b]$ is

$$P_X(a,b] = \int_a^b p_X(x)dx. \tag{1.7}$$

Note that this implies in particular that the integral of the $p_X(x)$ over S_X is 1. The *mean*, *variance* and *standard deviation* of such a continuous random variable X are defined by

$$\mu_X = \int_{-\infty}^{\infty} x p_X(x)dx, \tag{1.8}$$

$$V_X = \int_{-\infty}^{\infty} (x-\mu_X)^2 p_X(x)dx \tag{1.9}$$

and $\sigma_X = \sqrt{V_X}$.

The most important family of continuous random variables are the *Gaussian* (also called *normal*) *random variables* X that have the property that, for all real numbers x,

$$p_X(x) = \frac{1}{\sqrt{2\pi V_X}} \exp\left(-\frac{(x-\mu_X)^2}{2V_X}\right). \tag{1.10}$$

As usual, $\exp(x)$ denotes the value of the mathematical constant e raised to the power x. Such a probability density function can be seen in Fig. 1.21.

All Gaussian random variables "look the same" in the following well-defined sense. Let $(c, d]$ be an interval such that $-\infty \leq c < d \leq \infty$. Suppose that X and Y are Gaussian random variables. Then $P_X(\mu_X + c\sigma_X, \mu_X + d\sigma_X] = P_Y(\mu_Y + c\sigma_Y, \mu_Y + d\sigma_Y]$. In particular, they all look the same as the *standard Gaussian random variable* N, which is the Gaussian random variable with $\mu_N = 0$ and $V_N = 1$. This has the practically useful consequence that if we have a method (a table or a program) that allows us to calculate $P_N(c, d]$ for any interval $(c, d]$, then we can calculate $P_X(a, b]$ for any Gaussian random variable X and any interval $(a, b]$, since $P_X(a, b] = P_N\left(\frac{a-\mu_X}{\sigma_X}, \frac{b-\mu_X}{\sigma_X}\right]$. In particular, $P_N(-1, 1] > 0.67$, which implies that a sample of any random Gaussian variable will lie within one standard deviation of its mean in more than two cases out of three. We also know that $P_N(-2, 2] > 0.95$ and $P_N(-3, 3] > 0.995$. We can make use of these facts to estimate μ_X of a Gaussian random variable from a sample or samples. In more than 95 cases out of a 100, a sample will be within two standard deviations of μ_X. If we had a way of estimating the standard deviation (and we often do), then having observed a single sample we can say that we are 95% *confident* that the mean lies between two numbers, which are the sample plus/minus twice the standard deviation. Similarly, we can say that we are 99.5% confident that the mean lies between the sample plus/minus three standard deviations.

The reason for the importance of the Gaussian random variables is twofold. First, many random variables that occur in practice (in particular in things related to CT) can be closely approximated by some Gaussian random variable. We illustrate this in Fig. 1.21, where we plot the probabilities of the binomial random variable with $n = 30$ and $\rho = 0.7$ and the probability density function of the Gaussian random variable with the same mean (i.e., $n\rho = 21$) and the same variance (i.e., $n\rho(1-\rho) = 6.3$). Second, even if we start with a random variable X that is not at all similar to any Gaussian random variable, if we average a sufficient number (typically, thirty or more) of samples of X, then the random variable that corresponds to this average of samples will be very similar to a Gaussian random variable.

A mathematically precise statement of this is called the *central limit theorem*, it can be stated as follows. Let X be any random variable, discrete or continuous, with S_X consisting of real numbers and $V_X > 0$. For a fixed positive integer n, consider the following process for obtaining samples z of a random variable Z_n: pick n independent samples using X, sum them together, subtract $n\mu_X$, and divide by $\sqrt{nV_X}$. Then it will be the case that, for any interval $(a, b]$,

$$\lim_{n\to\infty} P_{Z_n}(a, b] = P_N(a, b]. \tag{1.11}$$

In words, the central limit theorem says that by taking the sum of a sufficiently large number of independent samples of any random variable X, then normalizing the sum by subtraction of $n\mu_X$ and by division by $\sqrt{n}\sigma_X$, we

get something that is indistinguishable from the standard Gaussian random variable!

A good example is provided if we start with the discrete random variable X for which $S_X = \{0, 1\}$ and p_X is set to $p_{1,0.7}$; i.e., $p_X(0) = 0.3$ and $p_X(1) = 0.7$. Let Y be the discrete random variable that is obtained by adding 30 random samples of X; then S_Y is the set of integers between 0 and 30 and p_Y is $p_{30,0.7}$ (recall the definition of the binomial random variables). The Z_{30} of (1.11) is obtained by taking a sample of Y, subtracting from it 21 and dividing the result by $\sqrt{6.3}$, which is just slightly larger than 2.5. If we now consider the interval (1,3], we see that $S_{Z_{30}}$ has five elements in this interval, the ones that correspond to 24, 25, 26, 27, and 28 in S_Y. So $P_{Z_{30}}(1, 3] = \sum_{m=24}^{28} p_{30,7}(m)$, which is approximately 0.159. Comparing this with $P_N(1, 3]$, which is approximately 0.157, we see that for the interval $(1, 3]$ the right-hand side of (1.11) is well approximated by its left-hand side already at the value $n = 30$. Except for the shift and the scaling that is used in the central limit theorem, the relationship between the two sides of (1.11) at $n = 30$ is demonstrated in Fig. 1.21. In fact, the discussion here indicates that, for every ρ strictly between 0 and 1, the binomial random variable determined by $p_{n,\rho}$ will be similar to a Gaussian random variable, provided that the n is chosen large enough.

In the definition of Z_n we used the expression *independent samples.* This means exactly what the language implies: when we pick one of the n samples we totally ignore the values that have been picked prior to that, we just pick a random sample from X. One might be tempted to describe the random variable whose samples are the sums of n independent samples of X by using the notation nX. But that notation is usually used for something quite different: assuming again that all elements of S_X are real numbers and that n is a positive integer, nX denotes the random variable for which $S_{nX} = \{nx \,|\, x \in S_X\}$, that is the set of all numbers nx for which x is in S_X, and $p_{nX}(nX) = p_X(X)$. It is easy to see that $\mu_{nX} = n\mu_X$ and $\sigma_{nX} = n\sigma_X$. On the other hand, it is a consequence of a soon-to-be-stated fact that, although the mean of the sum of n independent samples of X is also $n\mu_X$, the standard deviation of the sum is not $n\sigma_X$ but $\sqrt{n}\sigma_X$. Taking the average (rather than the sum) of n independent samples of X, we get a random variable whose mean is μ_X and whose standard deviation is $\sigma_X/\sqrt{n}$. Furthermore, according to the central limit theorem, this random variable becomes indistinguishable, as n increases, from the Gaussian random variable of mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. Thus, if it is our desire to estimate μ_X accurately, we can do this by averaging n independent samples of X for some large n, since (using a previously introduced terminology) we can be 99.5% confident that μ_X lies between the calculated average plus/minus $3\sigma_X/\sqrt{n}$, which can be made arbitrarily small by choosing n large enough.

One can think of the random variable nX introduced in the previous paragraph as the value of a *function on random variables* when that function is applied to X. Similarly, if X is a discrete random variable with el-

ements of S_X positive numbers, then $\ln(X)$ is a random variable such that $S_{\ln(X)}$ consists of the natural logarithms of elements of S_X and, for x in S_X, $p_{\ln(X)}\ln(x) = p_X(x)$.

Given any two continuous random variables X and Y over the real numbers, their sum $X+Y$ is a continuous random variable such that S_{X+Y} is also the set of real numbers and, for any real number z,

$$p_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)dx. \tag{1.12}$$

(A similar definition can be given for discrete random variables using a sum instead of the integral.) This corresponds to the process of sampling by picking a random sample x of X, then independently picking a random sample y of Y, and then producing the sample z by adding x and y. It is not difficult to prove that $\mu_{X+Y} = \mu_X + \mu_Y$ and $V_{X+Y} = V_X + V_Y$. We note that it is essential for these to be true that X and Y are sampled independently (statisticians would do this more formally, using the concept of *independent random variables*); as we have seen above, if we in fact chose Y to be the same random variable as X and always pick from S_Y what we have picked from S_X, then we would get the random variable $2X$, with $V_{2X} = 4V_X$. (We note by the way that the integral in (1.12) is the *convolution* of the functions p_X and p_Y; convolutions play an essential role in some methods for image reconstruction from projections and will be discussed further later on, especially in Chapter 8.) This concept of sum generalizes to any number of random variables: we sample $X_1 + \ldots + X_n$ by independently picking a sample of each of $X_1, \ldots, X_n$ and adding them together. It is a standard result in probability theory that

$$\mu_{X_1+\cdots+X_n} = \mu_{X_1} + \cdots + \mu_{X_n} \tag{1.13}$$

(this is true even if $X_1, \ldots, X_n$ are not independent) and

$$V_{X_1+\cdots+X_n} = V_{X_1} + \cdots + V_{X_n}. \tag{1.14}$$

In CT we frequently have to work with the ratios of two random variables. For example, in Section 2.5, we use A to denote the number of photons counted during an actual measurement and C to denote the number of photons counted during a calibration measurement, and then we take the ratio of these numbers. To put this into the context of the current discussion, let A and C be two random variables such that both S_A and S_C consist of the set of positive integers. Then A/C is the random variable for which $S_{A/C}$ is the set of positive rational numbers and, for any positive rational number q,

$$p_{A/C}(q) = \sum_{\substack{a,c \text{ positive integers} \\ (a/c)=q}} p_A(a)p_C(c). \tag{1.15}$$

Further concepts of probability theory will be introduced as and when they are needed.

Notes and References

An early book devoted mainly to the applications of image reconstruction from projections is [113]. That book contains survey articles on using image reconstruction to solve problems of finding the internal structure of the solar corona, the radio brightness of a portion of the sky, the distribution of radionuclides indicating the physiological functioning of the human body, and the dynamic behavior of the beating heart of a patient. Early applications of image reconstruction from projections to medicine were reported in [229], which includes articles on x-ray, proton, ultrasound, and emission computerized tomography. Another collection of articles that gives a rather mathematical approach to the field is [143] and a more recent book with a similar attitude is [211]. A relatively recent development is "discrete tomography" that also has interesting applications [125]; for a recent example see [278], which discusses this approach in geotomography. A recent book that concentrates on applications in various aspects of materials research, but covers quite a few topics in the process, is [18].

For a description of using lunar occultation observations for the reconstruction of two-dimensional brightness distribution of radio sources, see [257]. For a general tutorial of image reconstruction in radio astronomy, see [27]. For the reconstruction of the x-ray structure of a supernova remnant, see [204]. A survey on the reconstruction of the three-dimensional solar corona is given in [3], for a more recent article on the topic see [85]. For the use of tomography in geodesy see [149] and for its use in volcanology see [9, 172].

The first report in the open literature on an apparatus demonstrating the potential of reconstructive tomography in medicine was [213]. The procedure used for doing reconstruction in that paper is essentially the same as the backprojection method (to be discussed in Chapter 7). A more accurate method was proposed by A.M. Cormack in a paper [57] that also demonstrated the potential usefulness of reconstruction from x-ray projections in diagnostic medicine. The first commercially available x-ray CT scanner was designed by G.N. Hounsfield [145, 146]; it was used for scanning the head only. CT body scanning was introduced in [171]. The 1979 Nobel prize in physiology and medicine was awarded to G.N. Hounsfield and A.M. Cormack for their pioneering contributions to the development of computerized tomography. A book covering the technological state of the art for CT is [155].

Two books that cover the state of the art (up to 2004) in IMRT are [215, 268]. Two articles that discuss how the series expansion methods that were previously applied in image reconstruction from projections (and are presented in some detail below) can also be applied to IMRT are [42, 43]. A related development is intensity modulated proton therapy (IMPT); for a recent overview see [243]. It is also possible to do computerized tomography with protons, which has the potential of improving IMPT [240].

A brief early article (with a long bibliography) on medical magnetic resonance imaging was written by P.C. Lauterbur [170]. The 2003 Nobel prize

in physiology and medicine was awarded to P.C. Lauterbur and P. Mansfield for their discoveries concerning MRI. The original approach of Lauterbur was using reconstruction from projections; while nowadays most work is based on a different approach using Fourier transforms, there are some interesting developments that use projection imaging [152]. A book that discusses the technology of MRI, as well as some of its important applications to diagnostic medicine, is [13]; watch out for the soon-to-appear fourth edition! A review article on an important recent development in the field of MRI is [169].

A pioneering report on medical emission computerized tomography is [164] and an early survey of the topic is given by [35]. For more recent developments regarding PET, both basic science and clinical applications, see [17, 230, 263]. Clinical evaluation of the type of algorithms that can be used to produce reconstructions from the data collected by a PET scanner, such as the one shown in Fig. 1.8, to produce images, such as shown in Fig. 1.9, is reported in [53].

An early work on the topic of reconstructing the spatial distribution of acoustic absorption within tissue from acoustic projections is [101]. Two recent papers that are very relevant to Fig. 1.10 are [151, 269]. Note the 33-year spread between these publications! A thorough basic treatment of inverse problems for acoustic, electromagnetic and elastic wave scattering is [168]. Reconstruction from data collected by light (i.e., optical tomography) is a more recent development. A recent review of the topic is [91] and a very recent article is [162]. The methodology behind our Fig. 1.11 is discussed in [14]. The recent development of using reconstruction from data obtained by T-rays (alternatively called terahertz radiation) is discussed in [270].

The background material to our illustration of tomographic nondestructive testing (Figs. 1.12 and 1.13) can be obtained from [20, 173]. Other recent examples from the field of nondestructive testing are [96, 228]. For a general description of 3DXRD methodology, see [221]. For the specific algorithm used to obtain the reconstruction shown in Fig. 1.16, see [2, 233]. Recent developments combine diffraction data with attenuation data [150], use phase contrast tomography [207] or Friedel pairs [189], and image crystal growth [251].

The pioneering paper proposing the use of reconstruction from electron microscopic projections for the purpose of elucidation of biologically important molecular complexes was published by D.J. DeRosier and A. Klug [65]; the 1982 Nobel prize in chemistry was awarded to A. Klug for such work. Another example of an early paper on such techniques is [246]. The methods used in producing Fig. 1.19 are described in [238]; see also [236]. Two recent books that cover this field comprehensively are [83, 84]. A more recent paper that elucidates the tendency toward reconstruction from projections of larger cellular structures is [258]. Software tools for reconstruction from electron microscopic projections are described in the recent papers [239, 248].

Further details on probability and random variables can be found in [217], which provides additional material and the original references in this area. Two books written for users of statistics, rather than statisticians, are [1, 205].