1

Introduction and Preliminaries

Conceptual Outline

1.1 A deceptively simple model of the dynamics of a system is a deterministic iterative map applied to a single real variable. We characterize the dynamics by looking at its limiting behavior and the approach to this limiting behavior. Fixed points that attract or repel the dynamics, and cycles, are conventional limiting behaviors of a simple dynamic system. However, changing a parameter in a quadratic iterative map causes it to undergo a sequence of cycle doublings (bifurcations) until it reaches a regime of chaotic behavior which cannot be characterized in this way. This deterministic chaos reveals the potential importance of the influence of fine-scale details on large-scale behavior in the dynamics of systems.

1.2 A system that is subject to complex (external) influences has a dynamics that may be modeled statistically. The statistical treatment simplifies the complex unpredictable stochastic dynamics of a single system, to the simple predictable dynamics of an ensemble of systems subject to all possible influences. A random walk on a line is the prototype stochastic process. Over time, the random influence causes the ensemble of walkers to spread in space and form a Gaussian distribution. When there is a bias in the random walk, the walkers have a constant velocity superimposed on the spreading of the distribution.

1.3 While the microscopic dynamics of physical systems is rapid and complex, the macroscopic behavior of many materials is simple, even static. Before we can understand how complex systems have complex behaviors, we must understand why materials can be simple. The origin of simplicity is an averaging over the fast microscopic dynamics on the time scale of macroscopic observations (the ergodic theorem) and an averaging over microscopic spatial variations. The averaging can be performed theoretically using an ensemble representation of the physical system that assumes all microscopic states are realized. Using this as an assumption, a statistical treatment of microscopic states describes the macroscopic equilibrium behavior of systems. The final part of Section 1.3 introduces concepts that play a central role in the rest of the book. It discusses the differences between equilibrium and complex systems. Equilibrium systems are divisible and satisfy the ergodic theorem. Complex systems are composed out of interdependent parts and violate the ergodic theorem. They have many degrees of freedom whose time dependence is very slow on a microscopic scale.

1.4 To understand the separation of time scales between fast and slow degrees of freedom, a two-well system is a useful model. The description of a particle traveling in two wells can be simplified to the dynamics of a two-state (binary variable) system. The fast dynamics of the motion within a well is averaged by assuming that the system visits all states, represented as an ensemble. After taking the average, the dynamics of hopping between the wells is represented explicitly by the dynamics of a binary variable. The hopping rate depends exponentially on the ratio of the energy barrier and the temperature. When the temperature is low enough, the hopping is frozen. Even though the two wells are not in equilibrium with each other, equilibrium continues to hold within a well. The cooling of a two-state system serves as a simple model of a glass transition, where many microscopic degrees of freedom become frozen at the glass transition temperature.

1.5 Cellular automata are a general approach to modeling the dynamics of spatially distributed systems. Expanding the notion of an iterative map of a single variable, the variables that are updated are distributed on a lattice in space. The influence between variables is assumed to rely upon local interactions, and is homogeneous. Space and time are both discretized, and the variables are often simplified to include only a few possible states at each site. Various cellular automata can be designed to model key properties of physical and biological systems.

1.6 The equilibrium state of spatially distributed systems can be modeled by fields that are treated using statistical ensembles. The simplest is the Ising model, which captures the simple cooperative behavior found in magnets and many other systems. Cooperative behavior is a mechanism by which microscopic fast degrees of freedom can become slow collective degrees of freedom that violate the ergodic theorem and are visible macroscopically. Macroscopic phase transitions are the dynamics of the cooperative degrees of freedom. Cooperative behavior of many interacting elements is an important aspect of the behavior of complex systems. This should be contrasted to the two-state model (Section 1.4), where the slow dynamics occurs microscopically.

1.7 Computer simulations of models such as molecular dynamics or cellular automata provide important tools for the study of complex systems. Monte Carlo simulations enable the study of ensemble averages without necessarily describing the dynamics of a system. However, they can also be used to study random-walk dynamics. Minimization methods that use iterative progress to find a local minimum are often an important aspect of computer simulations. Simulated annealing is a method that can help find low energy states on complex energy surfaces.

1.8 We have treated systems using models without acknowledging explicitly that our objective is to describe them. All our efforts are designed to map a system onto a description of the system. For complex systems the description must be quite long, and the study of descriptions becomes essential. With this recognition, we turn to information theory. The information contained in a communication, typically a string of characters, may be defined quantitatively as the logarithm of the number of possible messages. When different messages have distinct probabilities P in an ensemble, then the information can be identified as $-\ln(P)$ and the average information is defined accordingly. Long messages can be modeled using the same concepts as a random walk, and we can use such models to estimate the information contained in human languages such as English.

1.9 In order to understand the relationship of information to systems, we must also understand what we can infer from information that is provided. The theory of logic is concerned with inference. It is directly linked to computation theory, which is concerned with the possible (deterministic) operations that can be performed on a string of characters. All operations on character strings can be constructed out of elementary logical (Boolean) operations on binary variables. Using Turing's model of computation, it is further shown that all computations can be performed by a universal Turing machine, as long as its input character string is suitably constructed. Computation theory is also related to our concern with the dynamics of physical systems because it explores the set of possible outcomes of discrete deterministic dynamic systems.

1.10 We return to issues of structure on microscopic and macroscopic scales by studying fractals that are self-similar geometric objects that embody the concept of progressively increasing structure on finer and finer length scales. A general approach to the scale dependence of system properties is described by scaling theory. The renormalization group methodology enables the study of scaling properties by relating a model of a system on one scale with a model of the system on another scale. Its use is illustrated by application to the Ising model (Section 1.6), and to the bifurcation route to chaos (Section 1.1). Renormalization helps us understand the basic concept of modeling systems, and formalizes the distinction between relevant and irrelevant microscopic parameters. Relevant parameters are the microscopic parameters that can affect the macroscopic behavior. The concept of universality is the notion that a whole class of microscopic models will give rise to the same macroscopic behavior, because many parameters are irrelevant. A conceptually related computational technique, the multigrid method, is based upon representing a problem on multiple scales.

The study of complex systems begins from a set of models that capture aspects of the dynamics of simple or complex systems. These models should be sufficiently general to encompass a wide range of possibilities but have sufficient structure to capture interesting features. An exciting bonus is that even the apparently simple models discussed in this chapter introduce features that are not typically treated in the conventional science of simple systems, but are appropriate introductions to the dynamics of complex systems.Our treatment of dynamics will often consider discrete rather than continuous time. Analytic treatments are often convenient to formulate in continuous variables and differential equations;however, computer simulations are often best formulated in discrete space-time variables with well-defined intervals. Moreover, the assumption of a smooth continuum at small scales is not usually a convenient starting point for the study of complex systems. We are also generally interested not only in one example of a system but rather in a class of systems that differ from each other but share a characteristic structure. The elements of such a class of systems are collectively known as an ensemble. As we introduce and study mathematical models, we should recognize that our primary objective is to represent properties of real systems. We must therefore develop an understanding of the nature of models and modeling, and how they can pertain to either simple or complex systems.

1.1 Iterative Maps (and Chaos)

An iterative map f is a function that evolves the state of a system s in discrete time

$$s(t) = f(s(t - \delta t)) \tag{1.1.1}$$

where $s(t)$ describes the state of the system at time t. For convenience we will generally measure time in units of δt which then has the value 1,and time takes integral values starting from the initial condition at $t = 0$.

Many of the complex systems we will consider in this text are of the form of Eq.(1.1.1),if we allow s to be a general variable of arbitrary dimension. The generality of iterative maps is discussed at the end of this section. We start by considering several examples of iterative maps where s is a single variable. We discuss briefly the binary variable case, $s = \pm 1$. Then we discuss in greater detail two types of maps with s a real variable, s , linear maps and quadratic maps. The quadratic iterative map is a simple model that can display complex dynamics. We assume that an iterative map may be started at any initial condition allowed by a specified domain of its system variable.

1.1.1 *Binary iterative maps*

There are only a few binary iterative maps.Question 1.1.1 is a complete enumeration of them.*

Question 1.1.1 Enumerate all possible iterative maps where the system is described by a single binary variable, $s = \pm 1$.

Solution 1.1.1 There are only four possibilities:

$$\begin{aligned} s(t) &= 1 \\ s(t) &= -1 \\ s(t) &= s(t-1) \\ s(t) &= -s(t-1) \end{aligned} \tag{1.1.2}$$

*Questions are an integral part of the text. They are designed to promote independent thought. The reader is encouraged to read the question, contemplate or work out an answer and then read the solution provided in the text. The continuation of the text assumes that solutions to questions have been read.

It is instructive to consider these possibilities in some detail. The main reason there are so few possibilities is that the form of the iterative map we are using depends,at most, on the value of the system in the previous time. The first two examples are constants and don't even depend on the value of the system at the previous time. The third map can only be distinguished from the first two by observation of its behavior when presented with two different initial conditions.

The last of the four maps is the only map that has any sustained dynamics. It cycles between two values in perpetuity. We can think about this as representing an oscillator. ▮

Question 1.1.2

a. In what way can the map $s(t) = -s(t - 1)$ represent a physical oscillator?

b. How can we think of the static map, $s(t) = s(t - 1)$, as an oscillator?

c. Can we do the same for the constant maps $s(t) = 1$ and $s(t) = -1$?

Solution 1.1.2 (*a*) By looking at the oscillator displacement with a strobe at half-cycle intervals,our measured values can be represented by this map. (*b*) By looking at an oscillator with a strobe at cycle intervals. (*c*) You might think we could, by picking a definite starting phase of the strobe with respect to the oscillator. However, the constant map ignores the first value, the oscillator does not. ▮

1.1.2 *Linear iterative maps: free motion, oscillation, decay and growth*

The simplest example of an iterative map with s real, s , is a constant map:

$$s(t) = s_0 \tag{1.1.3}$$

No matter what the initial value,this system always takes the particular value s_0. The constant map may seem trivial,however it will be useful to compare the constant map with the next class of maps.

A linear iterative map with unit coefficient is a model of free motion or propagation in space:

$$s(t) = s(t - 1) + v \tag{1.1.4}$$

at successive times the values of s are separated by v, which plays the role of the velocity.

Question 1.1.3 Consider the case of zero velocity

$$s(t) = s(t - 1) \tag{1.1.5}$$

How is this different from the constant map?

Solution 1.1.3 The two maps differ in their dependence on the initial value. ▮

Runaway growth or decay is a multiplicative iterative map:

$$s(t) = gs(t-1) \tag{1.1.6}$$

We can generate the values of this iterative map at all times by using the equivalent expression

$$s(t) = g^t s_0 = e^{\ln(g)t} s_0 \tag{1.1.7}$$

which is exponential growth or decay. The iterative map can be thought of as a sequence of snapshots of Eq.(1.1.7) at integral time. $g = 1$ reduces this map to the previous case.

Question 1.1.4 We have seen the case of free motion, and now jumped to the case of growth. What happened to accelerated motion? Usually we would consider accelerated motion as the next step after motion with a constant velocity. How can we write accelerated motion as an iterative map?

Solution 1.1.4 The description of accelerated motion requires two variables: position and velocity. The iterative map would look like:

$$\begin{aligned} x(t) &= x(t-1) + v(t-1) \\ v(t) &= v(t-1) + a \end{aligned} \tag{1.1.8}$$

This is a two-variable iterative map. To write this in the notation of Eq.(1.1.1) we would define s as a vector $s(t) = (x(t), v(t))$. ▮

Question 1.1.5 What happens in the rightmost exponential expression in Eq. (1.1.7) when g is negative?

Solution 1.1.5 The logarithm of a negative number results in a phase $i\pi$. The term $i\pi t$ in the exponent alternates sign every time step as one would expect from Eq. (1.1.6). ▮

At this point,it is convenient to introduce two graphical methods for describing an iterative map. The first is the usual way of plotting the value of s as a function of time. This is shown in the left panels of Fig. 1.1.1. The second type of plot,shown in the right panels, has a different purpose. This is a plot of the iterative relation $s(t)$ as a function of $s(t-1)$. On the same axis we also draw the line for the identity map $s(t) = s(t-1)$. These two plots enable us to graphically obtain the successive values of s as follows. Pick a starting value of s, which we can call $s(0)$. Mark this value on the abscissa. Mark the point on the graph of $s(t)$ that corresponds to the point whose abscissa is $s(0)$,i.e.,the point $(s(0), s(1))$.Draw a horizontal line to intersect the identity map. The intersection point is $(s(1), s(1))$. Draw a vertical line back to the iterative map. This is the point $(s(1), s(2))$. Successive values of $s(t)$ are obtained by iterating this graphical procedure. A few examples are plotted in the right panels of Fig. 1.1.1.

In order to discuss the iterative maps it is helpful to recognize several features of these maps.First,intersection points of the identity map and the iterative map are the fixed points of the iterative map:

$$s_0 = f(s_0) \tag{1.1.9}$$

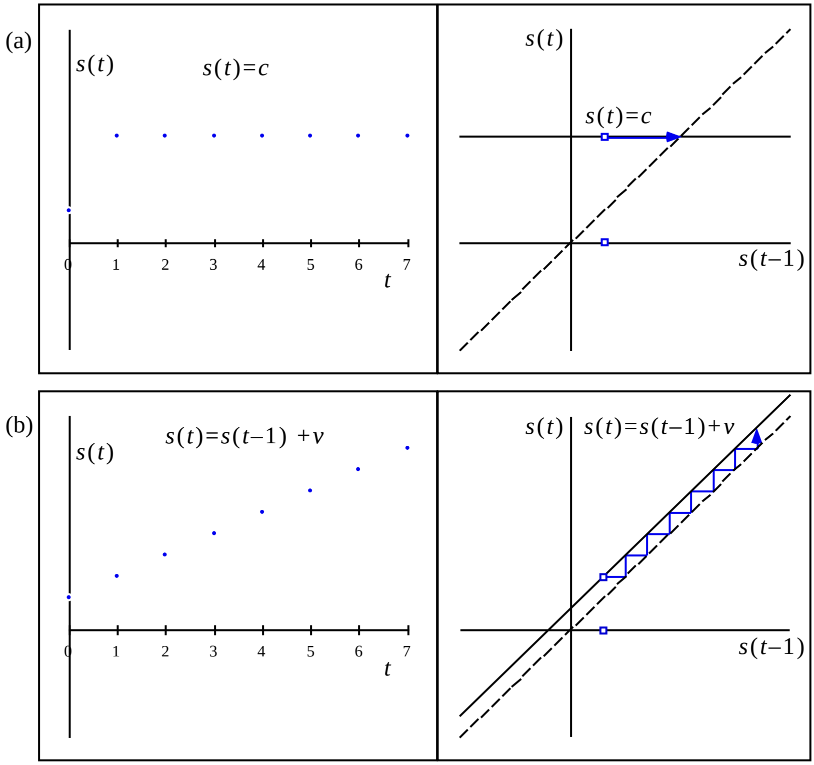
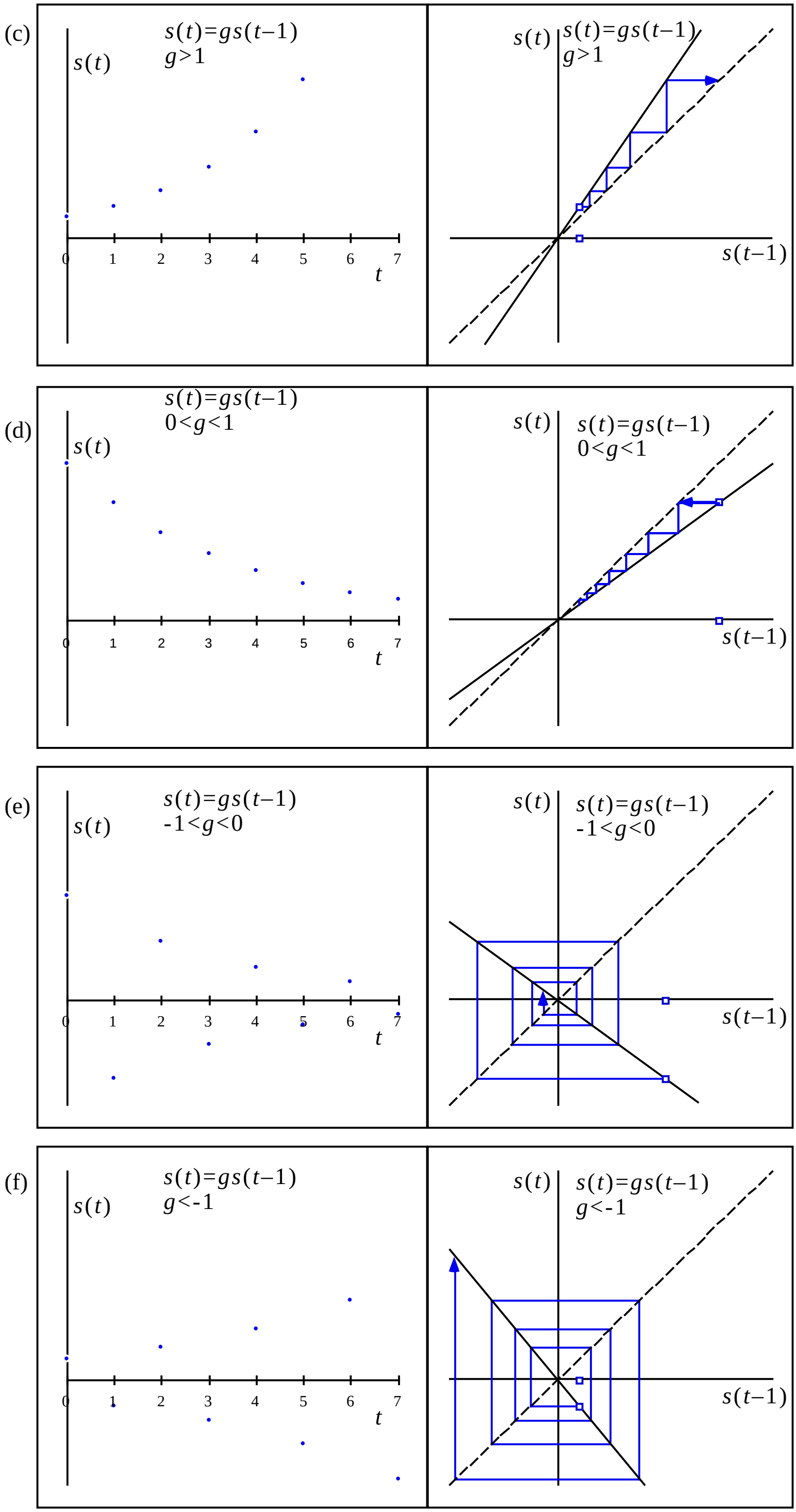


Figure 1.1.1 The left panels show the time-dependent value of the system variable $s(t)$ resulting from iterative maps. The first panel (a) shows the result of iterating the constant map; (b) shows the result of adding v to the previous value during each time interval; (c)–(f) show the result of multiplying by a constant g, where each figure shows the behavior for a different range of g values: (c) $g > 1$, (d) $0 < g < 1$, (e) $-1 < g < 0$, and (f) $g < -1$. The right panels are a different way of showing graphically the results of iterations and are constructed as follows. First plot the function $f(s)$ (solid line), where $s(t) = f(s(t-1))$. This can be thought of as plotting $s(t)$ vs. $s(t-1)$. Second, plot the identity map $s(t) = s(t-1)$ (dashed line). Mark the initial value $s(0)$ on the horizontal axis, and the point on the graph of $s(t)$ that corresponds to the point whose abscissa is $s(0)$, i.e. the point $(s(0), s(1))$. These are shown as squares. From the point $(s(0), s(1))$ draw a horizontal line to intersect the identity map. The intersection point is $(s(1), s(1))$. Draw a vertical line back to the iterative map. This is the point $(s(1), s(2))$. Successive values of $s(t)$ are obtained by iterating this graphical procedure. ▮

Fixed points,not surprisingly, play an important role in iterative maps. They help us describe the state and behavior of the system after many iterations. There are two kinds of fixed points—stable and unstable. Stable fixed points are characterized by “attracting” the result of iteration of points that are nearby. More precisely, there exists


(c)
s(t)=gs(t−1)
g>1
s(t)
0 1 2 3 4 5 6 7
t
s(t)
s(t)=gs(t−1)
g>1
s(t−1)
(d)
s(t)=gs(t−1)
0<g<1
s(t)
0 1 2 3 4 5 6 7
t
s(t)
s(t)=gs(t−1)
0<g<1
s(t−1)
(e)
s(t)=gs(t−1)
-1<g<0
s(t)
0 1 2 3 4 5 6 7
t
s(t)
s(t)=gs(t−1)
-1<g<0
s(t−1)
(f)
s(t)=gs(t−1)
g<-1
s(t)
0 1 2 3 4 5 6 7
t
s(t)
s(t)=gs(t−1)
g<-1
s(t−1)

a neighborhood of points of s_0 such that for any s in this neighborhood the sequence of points

$$\{s, f(s), f^2(s), f^3(s), \quad \} \tag{1.1.10}$$

converges to s_0. We are using the notation $f^2(s) = f(f(s))$ for the second iteration,and similar notation for higher iterations. This sequence is just the time series of the iterative map for the initial condition s. Unstable fixed points have the opposite behavior, in that iteration causes the system to leave the neighborhood of s_0. The two types of fixed points are also called attracting and repelling fixed points.

The family of multiplicative iterative maps in Eq.(1.1.6) all have a fixed point at $s_0 = 0$. Graphically from the figures, or analytically from Eq. (1.1.7), we see that the fixed point is stable for $|g| < 1$ and is unstable for $|g| > 1$. There is also distinct behavior of the system depending on whether g is positive or negative. For $g < 0$ the iterations alternate from one side to the other of the fixed point, whether it is attracted to or repelled from the fixed point. Specifically, if $s < s_0$ then $f(s) > s_0$ and vice versa, or $\mathrm{sign}(s - s_0) = -\mathrm{sign}(f(s) - s_0)$. For $g > 0$ the iteration does not alternate.

Question 1.1.6 Consider the iterative map.

$$s(t) = gs(t - 1) + v \tag{1.1.11}$$

convince yourself that v does not affect the nature of the fixed point, only shifts its position.

Question 1.1.7 Consider an arbitrary iterative map of the form Eq.(1.1.1), with a fixed point s_0 (Eq.(1.1.9)). If the iterative map can be expanded in a Taylor series around s_0 show that the first derivative

$$g = \left.\frac{df(s)}{ds}\right|_{s_0} \tag{1.1.12}$$

characterizes the fixed point as follows:

For $|g| < 1$, s_0 is an attracting fixed point.

For $|g| > 1$, s_0 is a repelling fixed point.

For $g < 0$, iterations alternate sides in a sufficiently small neighborhood of s_0.

For $g > 0$,iterations remain on one side in a sufficiently small neighborhood of s_0.

Extra credit: Prove the same theorem for a differentiable function (no Taylor expansion needed) using the mean value theorem.

Solution 1.1.7 If the iterative map can be expanded in a Taylor series we write that

$$f(s) = f(s_0) + g(s - s_0) + h(s - s_0)^2 + \tag{1.1.13}$$

where g is the first derivative at s_0, and h is one-half of the second derivative at s_0. Since s_0 is a fixed point $f(s_0) = s_0$ we can rewrite this as:

$$\frac{f(s) - s_0}{s - s_0} = g + h(s - s_0) + \tag{1.1.14}$$

If we did not have any higher-order terms beyond g, then by inspection each of the four conditions that we have to prove would follow from this expression without restrictions on s. For example, if $|g| > 1$, then taking the magnitude of both sides shows that $f(s) - s_0$ is larger than $s - s_0$ and the iterations take the point s away from s_0. If $g > 0$,then this expression says that $f(s)$ stays on the same side of s_0. The other conditions follow similarly.

To generalize this argument to include the higher-order terms of the expansion, we must guarantee that whichever domain g is in ($g > 1$, $0 < g < 1$, $-1 < g < 0$, or $g < -1$), the same is also true of the whole right side. For a Taylor expansion, by choosing a small enough neighborhood $|s - s_0| < \delta$, we can guarantee the higher-order terms are less than any number ε we choose. We choose ε to be half of the minimum of $|g - 1|$, $|g - 0|$ and $|g + 1|$. Then $g + \varepsilon$ is in the same domain as g. This provides the desired guarantee and the proof is complete.

We have proven that in the vicinity of a fixed point the iterative map may be completely characterized by its first-order expansion (with the exception of the special points $g = \pm 1,0$). ▮

Thus far we have not considered the special cases $g = \pm 1,0$. The special cases $g = 0$ and $g = 1$ have already been treated as simpler iterative maps. When $g = 0$, the fixed point at $s = 0$ is so attractive that it is the result of any iteration. When $g = 1$ all points are fixed points.

The new special case $g = -1$ has a different significance. In this case all points alternate between positive and negative values, repeating every other iteration. Such repetition is a generalization of the fixed point. Whereas in the fixed-point case we repeat every iteration, here we repeat after every two iterations. This is called a 2-cycle, and we can immediately consider the more general case of an n-cycle. In this terminology a fixed point is a 1-cycle.One way to describe an n-cycle is to say that iterating n times gives back the same result, or equivalently, that a new iterative map which is the nth fold composition of the original map $h = f^n$ has a fixed point. This description would include also fixed points of f and all points that are m-cycles, where m is a divisor of n. These are excluded from the definition of the n-cycles. While we have introduced cycles using a map where all points are 2-cycles,more general iterative maps have specific sets of points that are n-cycles. The set of points of an n-cycle is called an orbit. There are a variety of properties of fixed points and cycles that can be proven for an arbitrary map. One of these is discussed in Question 1.1.8.

Question 1.1.8 Prove that there is a fixed point between any two points of a 2-cycle if the iterating function f is continuous.

Solution 1.1.8 Let the 2-cycle be written as

$$\begin{aligned} s_2 &= f(s_1) \\ s_1 &= f(s_2) \end{aligned} \tag{1.1.15}$$

Consider the function $h(s) = f(s) - s$, $h(s_1)$ and $h(s_2)$ have opposite signs and therefore there must be an s_0 between s_1 and s_2 such that $h(s_0) = 0$—the fixed point. ▮

We can also generalize the definition of attracting and repelling fixed points to consider attracting and repelling n-cycles. Attraction and repulsion for the cycle is equivalent to the attraction and repulsion of the fixed point of f^n.

1.1.3 *Quadratic iterative maps: cycles and chaos*

The next iterative map we will consider describes the effect of nonlinearity (self-action):

$$s(t) = as(t-1)(1 - s(t-1)) \tag{1.1.16}$$

or equivalently

$$f(s) = as(1 - s) \tag{1.1.17}$$

This map has played a significant role in development of the theory of dynamical systems because even though it looks quite innocent,it has a dynamical behavior that is not described in the conventional science of simple systems. Instead, Eq. (1.1.16) is the basis of significant work on chaotic behavior, and the transition of behavior from simple to chaotic. We have chosen this form of quadratic map because it simplifies somewhat the discussion. Question 1.1.11 describes the relationship between this family of quadratic maps,parameterized by a, and what might otherwise appear to be a different family of quadratic maps.

We will focus on a values in the range $4 > a > 0$. For this range, any value of s in the interval s [0,1] stays within this interval. The minimum value $f(s) = 0$ occurs for $s = 0,1$ and the maximal value occurs for $s = 1/2$. For all values of a there is a fixed point at $s = 0$ and there can be at most two fixed points, since a quadratic can only intersect a line (Eq. (1.1.9)) in two points.

Taking the first derivative of the iterative map gives

$$\frac{df}{ds} = a(1-2s) \tag{1.1.18}$$

At $s = 0$ the derivative is a which, by Question 1.1.7,shows that $s = 0$ is a stable fixed point for $a < 1$ and an unstable fixed point for $a > 1$. The switching of the stability of the fixed point at $s = 0$ coincides with the introduction of a second fixed point in the interval [0,1] (when the slope at $s = 0$ is greater than one, $f(s) > s$ for small s, and since

$f(1) = 0$, we have that $f(s_1) = s_1$ for some s_1 in [0,1] by the same construction as in Question 1.1.8). We find s_1 by solving the equation

$$s_1 = as_1(1-s_1) \tag{1.1.19}$$

$$s_1 = (a-1)/a \tag{1.1.20}$$

Substituting this into Eq. (1.1.18) gives

$$\left.\frac{df}{ds}\right|_{s_1} = 2-a \tag{1.1.21}$$

This shows that for $1 < a < 3$, the new fixed point is stable by Question 1.1.7. Moreover, the derivative is positive for $1 < a < 2$, so s_1 is stable and convergence is from one side. The derivative is negative for $2 < a < 3$, so s_1 is stable and alternating.

Fig. 1.1.2(a)–(c) shows the three cases: $a = 0.5$, $a = 1.5$ and $a = 2.8$. For $a = 0.5$, starting from anywhere within [0,1] leads to convergence to $s = 0$. When $s(0) > 0.5$ the first iteration takes the system to $s(1) < 0.5$. The closer we start to $s(0) = 1$ the closer to $s = 0$ we get in the first jump. At $s(0) = 1$ the convergence to 0 occurs in the first jump. A similar behavior would be found for any value of $0 < a < 1$. For $a = 1.5$ the behavior is more complicated. Except for the points $s = 0,1$, the convergence is always to the fixed point $s_1 = (a - 1)/a$ between 0 and 1. For $a = 2.8$ the iterations converge to the same point; however, the convergence is alternating. Because there can be at most two fixed points for the quadratic map, one might think that this behavior would be all that would happen for $1 < a < 4$. One would be wrong. The first indication that this is not the case is the instability of the fixed point at s_1 starting from $a = 3$.

What happens for $a > 3$? Both of the fixed points that we have found, and the only ones that can exist for the quadratic map, are now unstable. We know that the iteration of the map has to go somewhere, and only within [0,1]. The only possibility, within our experience, is that there is an attracting n-cycle to which the fixed points are unstable. Let us then consider the map $f^2(s)$ whose fixed points are 2-cycles of the original map. $f^2(s)$ is shown in the right panels of Fig. 1.1.2 for increasing values of a. The fixed points of $f(s)$ are also fixed points of $f^2(s)$. However, we see that two additional fixed points exist for $a > 3$. We can also show analytically that two fixed points are introduced at exactly $a = 3$:

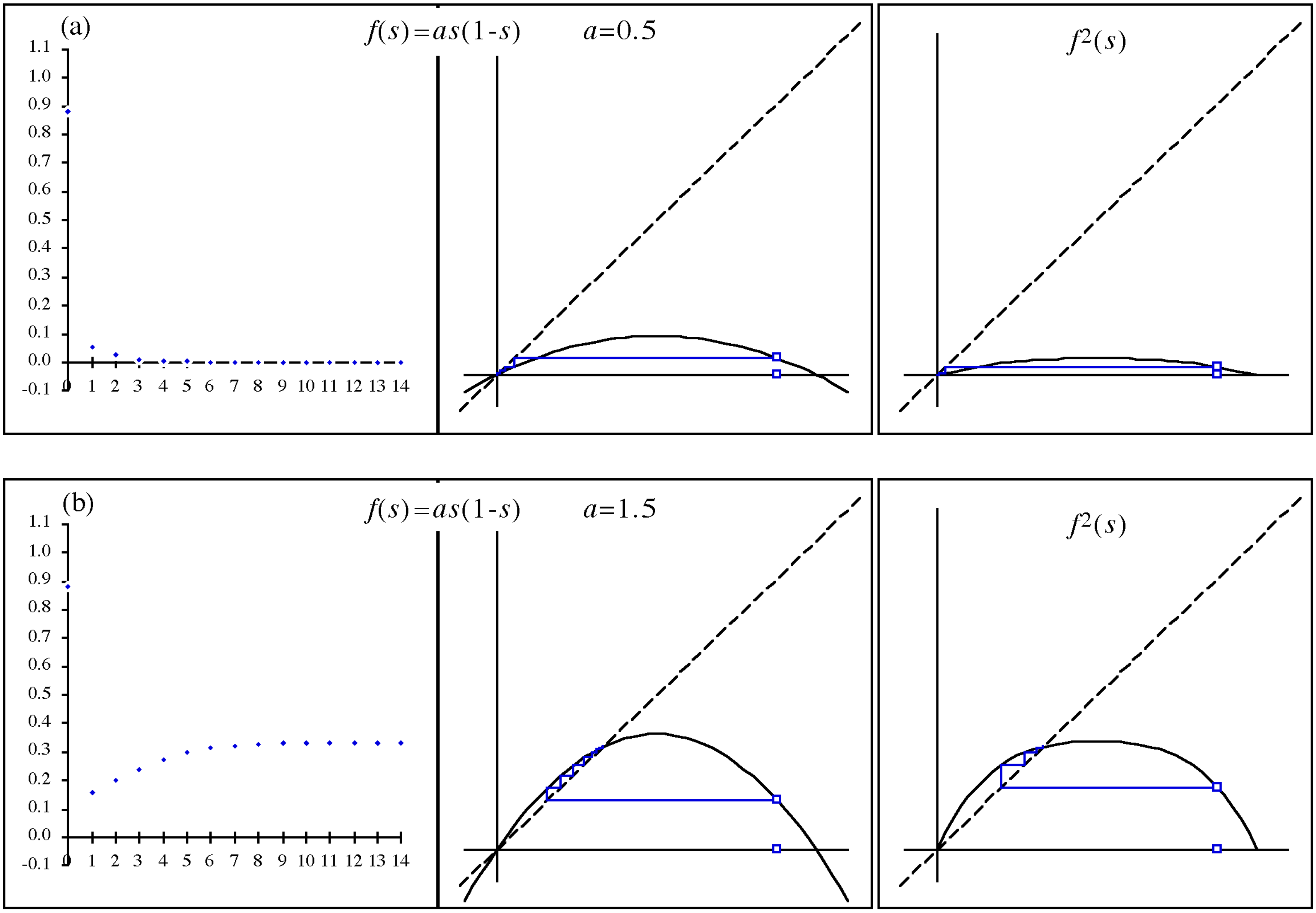
$$f^2(s) = a^2 s(1-s)(1-as(1-s)) \tag{1.1.22}$$

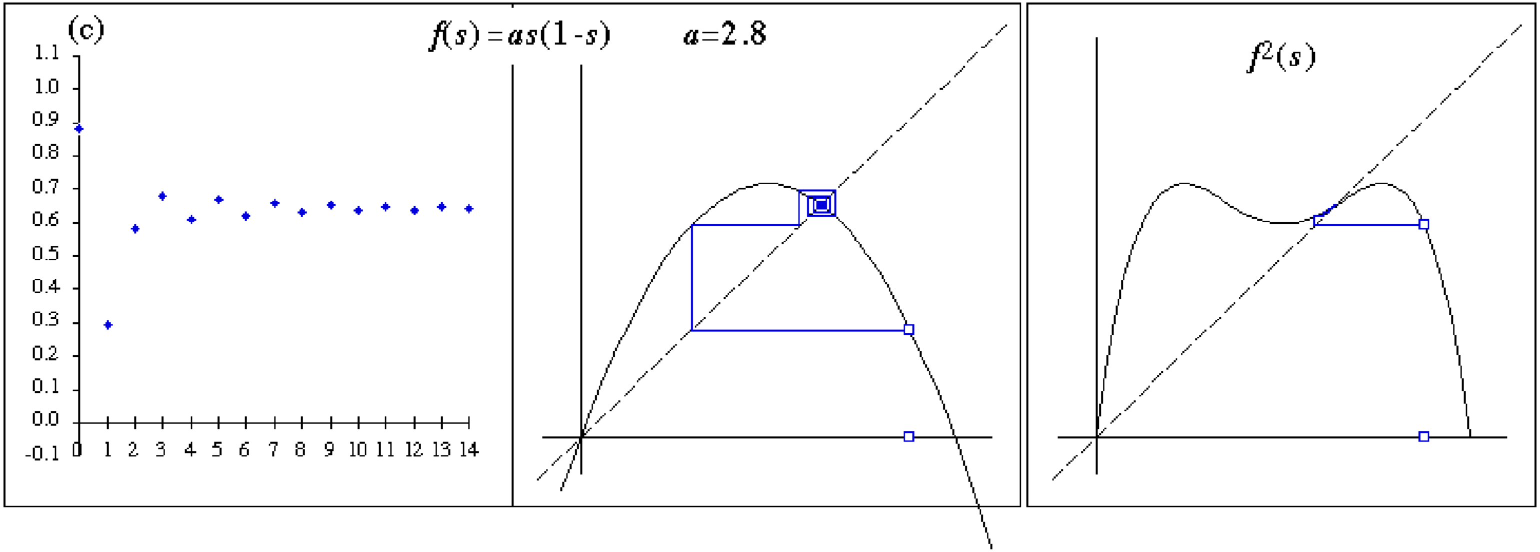
To find the fixed point we solve:

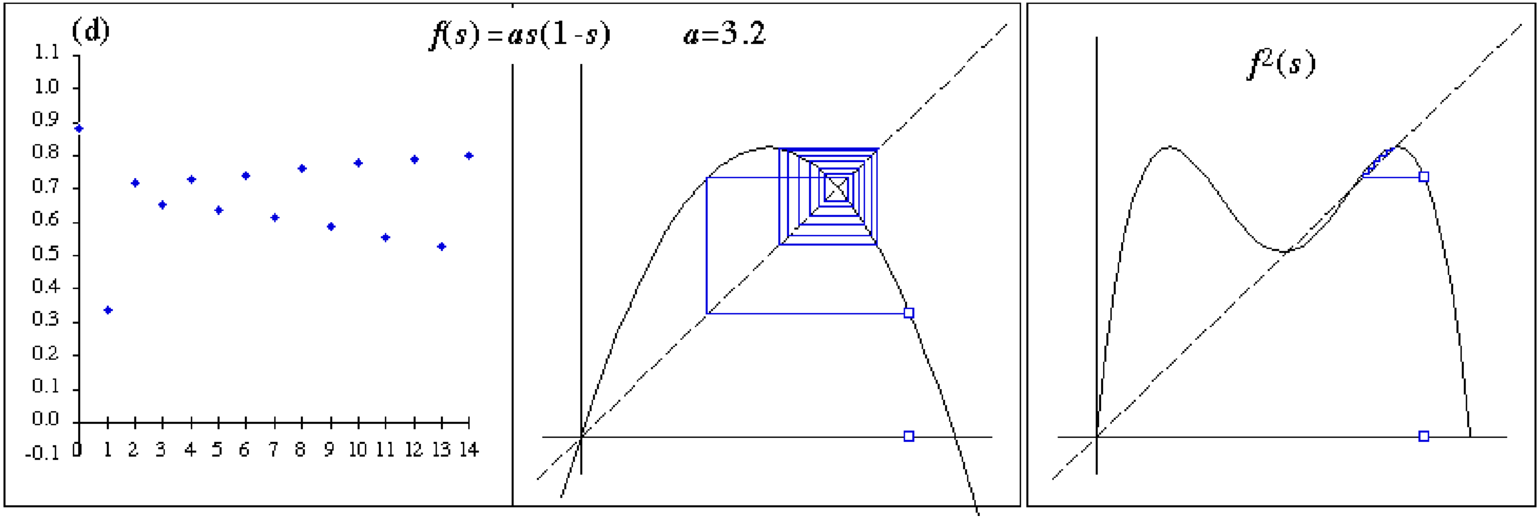
$$s = a^2 s(1-s)(1-as(1-s)) \tag{1.1.23}$$

We already know two solutions of this quartic equation—the fixed points of the map f. One of these at $s = 0$ is obvious. Dividing by s we have a cubic equation:

$$a^3s^3 - 2a^3s^2 + a^2(1+a)s + (1-a^2) = 0 \tag{1.1.24}$$


(a)
$f(s)=as(1-s)$
a=0.5
$f^2(s)$
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
(b)
$f(s)=as(1-s)$
a=1.5
$f^2(s)$
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14


(c)
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
$f(s) = as(1-s)$
a=2.8
$f^2(s)$


(d)
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
$f(s) = as(1-s)$
a=3.2
$f^2(s)$

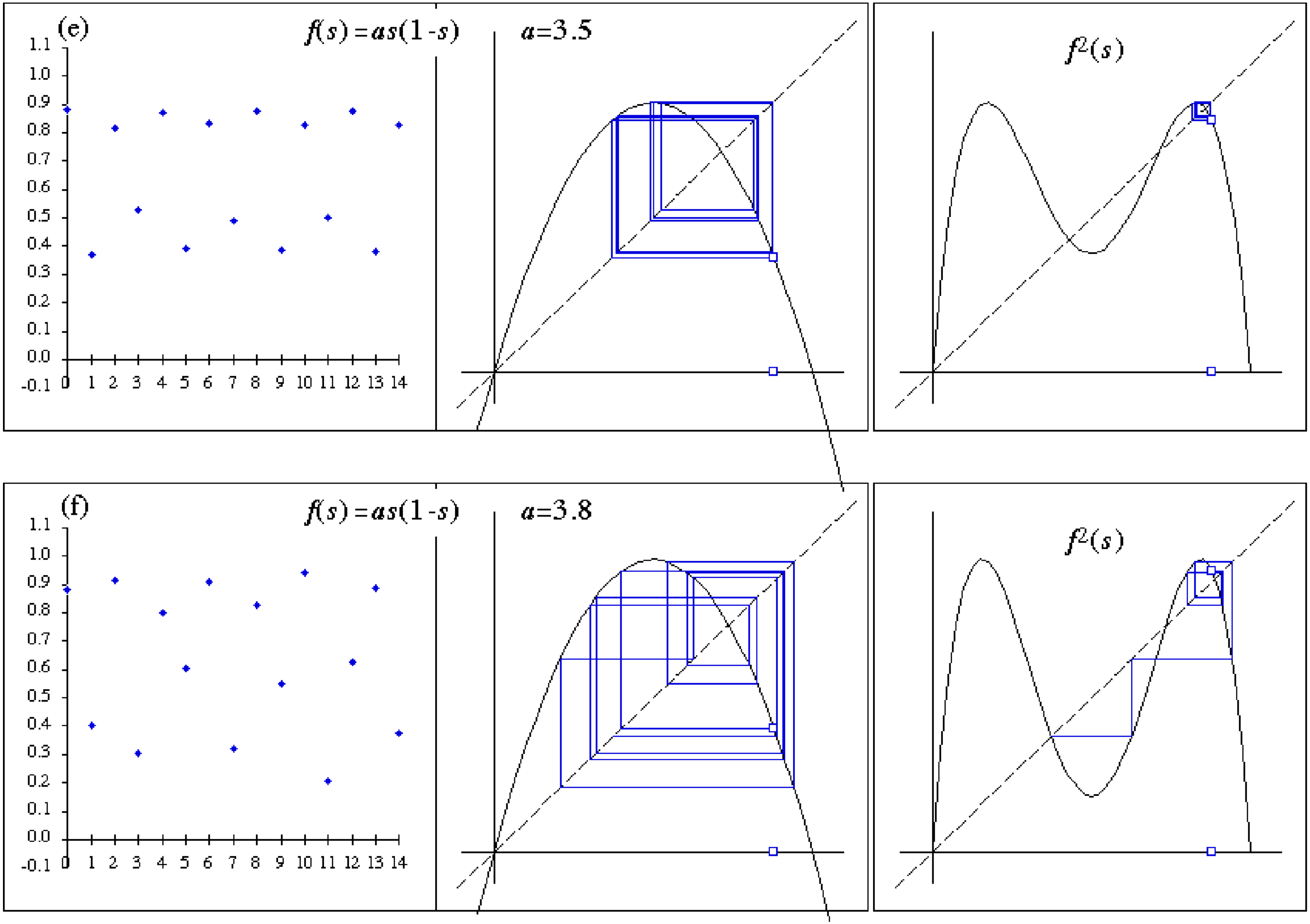

(e)
$f(s)=as(1-s)$
a=3.5
$f^2(s)$
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
(f)
$f(s)=as(1-s)$
a=3.8
$f^2(s)$
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

Figure 1.1.2 (pp. 28-30) Plots of the result of iterating the quadratic map $f(s) = as(1 - s)$ for different values of a. The left and center panels are similar to the left and right panels of Fig. 1.1.1. The left panels plot $s(t)$. The center panels describe the iteration of the map $f(s)$ on axes corresponding to $s(t)$ and $s(t - 1)$. The right panels are similar to the center panels but are for the function $f^2(s)$. The different values of a are indicated on the panels and show the changes from (a) convergence to $s = 0$ for $a = 0.5$, (b) convergence to $s = (a - 1)/a$ for $a = 1.5$, (c) alternating convergence to $s = (a - 1)/a$ for $a = 2.8$, (d) bifurcation — convergence to a 2-cycle for $a = 3.2$, (e) second bifurcation — convergence to a 4-cycle for $a = 3.5$, (f) chaotic behavior for $a = 3.8$. ▮

We can reduce the equation to a quadratic by dividing by $(s - s_1)$ as follows (we simplify the algebra by dividing by $a(s - s_1) = (as - (a - 1))$):

$$\begin{array}{r|l}
 & a^2s^2 - a(a+1)s + (a+1) \\ \hline
(as - (a-1)) & a^3s^3 - 2a^3s^2 + a^2(1+a)s + (1-a^2) \\
 & \underline{a^3s^3 - (a-1)a^2s^2} \\
 & \quad -(a+1)a^2s^2 + a^2(1+a)s + (1-a^2) \\
 & \quad \underline{-(a+1)a^2s^2 + a(1+a)(a-1)s} \\
 & \qquad\qquad + a(1+a)s + (1-a^2)
\end{array} \tag{1.1.25}$$

Now we can obtain the roots to the quadratic:

$$a^2s^2 - a(a+1)s + (a+1) = 0 \tag{1.1.26}$$

$$s_2 = \frac{(a+1) \pm \sqrt{(a+1)(a-3)}}{2a} \tag{1.1.27}$$

This has two solutions (as it must for a 2-cycle) for $a < -1$ or for $a > 3$. The former case is not of interest to us since we have assumed $0 < a < 4$. The latter case is the two roots that are promised. Notice that for exactly $a = 3$ the two roots that are the new 2-cycle are the same as the fixed point we have already found s_1. The 2-cycle splits off from the fixed point at $a = 3$ when the fixed point becomes unstable. The two attracting points continue to separate as a increases. For $a > 3$ we expect that the result of iteration eventually settles down to the 2-cycle. The system state alternates between the two roots Eq. (1.1.27). This is shown in Fig. 1.1.2(d).

As we continue to increase a beyond 3, the 2-cycle will itself become unstable at a point that can be calculated by setting

$$\left.\frac{df^2}{ds}\right|_{s_2} = -1 \tag{1.1.28}$$

to be $a = 1 + \sqrt{6} = 3.44949$. At this value of a the 2-cycle splits into a 4-cycle (Fig. 1.1.2(e)).Each of the fixed points of $f^2(s)$ simultaneously split into 2-cycles that together form a 4-cycle for the original map.

Question 1.1.9 Show that when f has a 2-cycle, both of the fixed points of f^2 must split simultaneously.

Solution 1.1.9 The split occurs when the fixed points become unstable—the derivative of f^2 equals –1. We can show that the derivative is equal at the two fixed points of Eq. (1.1.27), which we call $s_2^{\pm}$:

$$\left.\frac{df^2}{ds}\right|_{s_2} = \left.\frac{df(f(s))}{ds}\right|_{s_2} = \left.\frac{df(s)}{ds}\right|_{f(s_2)} \left.\frac{df(s)}{ds}\right|_{s_2} \qquad (1.1.29)$$

where we have made use of the chain rule. Since $f(s_2^+) = s_2^-$ and vice versa, we have shown this expression is the same whether $s_2 = s_2^+$ or $s_2 = s_2^-$.

Note: This can be generalized to show that the derivative of f^k is the same at all of its k fixed points corresponding to a k-cycle of f. ∎

The process of taking an n-cycle into a $2n$-cycle is called bifurcation. Bifurcation continues to replace the limiting behavior of the iterative map with progressively longer cycles of length 2^k. The bifurcations can be simulated. They occur at smaller and smaller intervals and there is a limit point to the bifurcations at $a_c = 3.56994567$. Fig. 1.1.3 shows the values that are reached by the iterative map at long times—the stable cycles—as a function of $a < a_c$. We will discuss an algebraic treatment of the bifurcation regime in Section 1.10.

Beyond the bifurcation regime $a > a_c$ (Fig. 1.1.2(f)) the behavior of the iterative map can no longer be described using simple cycles that attract the iterations. The behavior in this regime has been identified with chaos. Chaos has been characterized in many ways, but one property is quite generally agreed upon—the inherent lack of predictability of the system dynamics. This is often expressed more precisely by describing the sensitivity of the system's fate to the initial conditions. A possible definition is: There exists a distance d such that for any neighborhood V of any point s it is possible to find a point s' within the neighborhood and a number of iterations k so that $f^k(s')$ is further than d away from $f^k(s)$. This means that arbitrarily close to any point is a point that will be displaced a significant distance away by iteration. Qualitatively, there are two missing aspects of this definition,first that the points that move far away must not be too unlikely (otherwise the system is essentially predictable) and second that d is not too small (in which case the divergence of the dynamics may not be significant).

If we look at the definition of chaotic behavior, we see that the concept of scale plays an important role. A small distance between s and s' turns into a large distance between $f^k(s)$ and $f^k(s')$. Thus a fine-scale difference eventually becomes a large-scale difference. This is the essence of chaos as a model of complex system behavior. To understand it more fully, we can think about the state variable s not as one real variable,

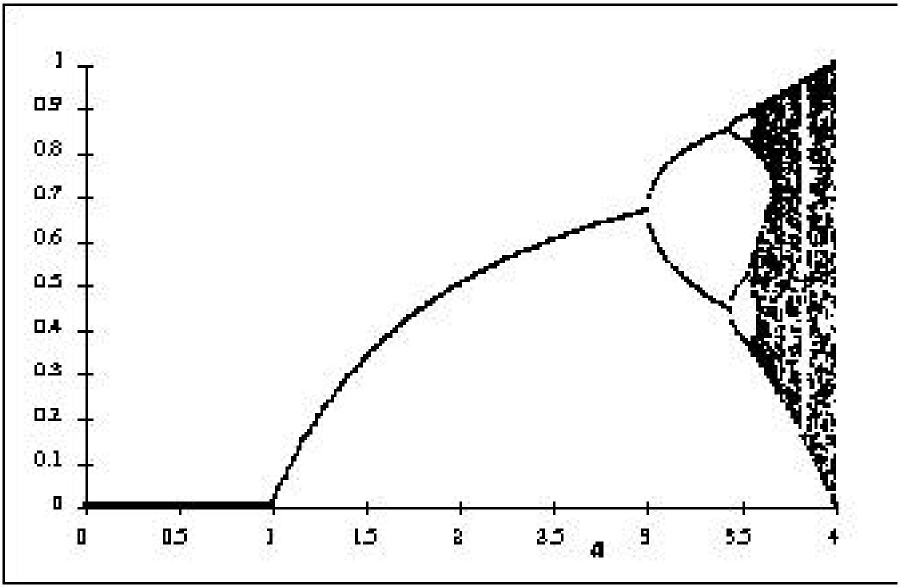


Figure 1.1.3 A plot of values of s visited by the quadratic map $f(s) = as(1 - s)$ after many iterations as a function of a, including stable points, cycles and chaotic behavior. The different regimes are readily apparent. For $a < 1$ the stable point is $s = 0$. For $1 < a < 3$ the stable point is at $s_0 = (a - 1)/a$. For $3 < a < a_c$ with $a_c = 3.56994567$, there is a bifurcation cascade with 2-cycles then 4-cycles, etc. 2^k-cycles for all values of k appear in progressively narrower regions of a. Beyond 4-cycles they cannot be seen in this plot. For $a > a_c$ there is chaotic behavior. There are regions of s values that are not visited and regions that are visited in the long time behavior of the quadratic map in the chaotic regime which this figure does not fully illustrate. ∎

but as an infinite sequence of binary variables that form its binary representation $s = 0.r_1r_2r_3r_4...$ Each of these binary variables represents the state of the system—the value of some quantity we can measure about the system—on a particular length scale. The higher order bits represent the larger scales and the lower order ones represent the finer scales. Chaotic behavior implies that the state of the first few binary variables, r_1r_2, at a particular time are determined by the value of fine scale variables at an earlier time. The farther back in time we look, the finer scale variables we have to consider in order to know the present values of r_1r_2. Because many different variables are relevant to the behavior of the system, we say that the system has a complex behavior. We will return to these issues in Chapter 8.

The influence of fine length scales on coarse ones makes iterative maps difficult to simulate by computer. Computer representations of real numbers always have finite precision. This must be taken into account if simulations of iterative maps or chaotic complex systems are performed.

Another significant point about the iterative map as a model of a complex system is that there is nothing outside of the system that is influencing it. All of the information we need to describe the behavior is contained in the precise value of s. The complex behavior arises from the way the different parts of the system—the fine and course scales—affect each other.

Question 1.1.10: Why isn't the iterative map in the chaotic regime equivalent to picking a number at random?

Solution 1.1.10: We can still predict the behavior of the iterative map over a few iterations. It is only when we iterate long enough that the map becomes unpredictable. More specifically, the continuity of the function $f(s)$ guarantees that for s and s' close together $f(s)$ and $f(s')$ will also be close together. Specifically, given an ε it is possible to find a δ such that for $|s - s'| < \delta$, $|f(s) - f(s')| < \varepsilon$. For the family of functions we have been considering, we only need to set $\delta < \varepsilon/a$, since then we have:

$$|f(s) - f(s')| = a|s(1-s) - s'(1-s')| = a|s - s'||1-(s+s')| < a|s-s'| < \varepsilon \tag{1.1.30}$$

Thus if we fix the number of cycles to be k, we can always find two points close enough so that $|f^k(s') - f^k(s)| < \varepsilon$ by setting $|s - s'| < \varepsilon/a^k$. ▮

The tuning of the parameter a leading from simple convergent behavior through cycle bifurcation to chaos has been identified as a universal description of the appearance of chaotic behavior from simple behavior of many systems. How do we take a complicated real system and map it onto a discrete time iterative map? We must define a system variable and then take snapshots of it at fixed intervals (or at least well-defined intervals). The snapshots correspond to an iterative map. Often there is a natural choice for the interval that simplifies the iterative behavior. We can then check to see if there is bifurcation and chaos in the real system when parameters that control the system behavior are varied.

One of the earliest examples of the application of iterative maps is to the study of heart attacks. Heart attacks occur in many different ways. One kind of heart attack is known as fibrillation. Fibrillation is characterized by chaotic and ineffective heart muscle contractions. It has been suggested that bifurcation may be observed in heartbeats as a period doubling (two heartbeats that are inequivalent). If correct, this may serve as a warning that the heart structure, due to various changes in heart tissue parameters, may be approaching fibrillation. Another system where more detailed studies have suggested that bifurcation occurs as a route to chaotic behavior is that of turbulent flows in hydrodynamic systems. A subtlety in the application of the ideas of bifurcation and chaos to physical systems is that physical systems are better modeled as having an increasing number of degrees of freedom at finer scales. This is to be contrasted with a system modeled by a single real number, which has the same number of degrees of freedom (represented by the binary variables above) at each length scale.

1.1.4 *Are all dynamical systems iterative maps?*

How general is the iterative map as a tool for describing the dynamics of systems? There are three apparent limitations of iterative maps that we will consider modifying later, Eq. (1.1.1):

a. describes the homogeneous evolution of a system since *f* itself does not depend on time,

b. describes a system where the state of the system at time *t* depends only on the state of the system at time $t - \delta t$, and

c. describes a deterministic evolution of a system.

We can,however, bypass these limitations and keep the same form of the iterative map if we are willing to let *s* describe not just the present state of the system but also

a. the state of the system and all other factors that might affect its evolution in time,

b. the state of the system at the present time and sufficiently many previous times, and

c. the probability that the system is in a particular state.

Taking these caveats together, all of the systems we will consider are iterative maps, which therefore appear to be quite general.Generality, however, can be quite useless, since we want to discard as much information as possible when describing a system.

Another way to argue the generality of the iterative map is through the laws of classical or quantum dynamics.If we consider *s* to be a variable that describes the positions and velocities of all particles in a system, all closed systems described by classical mechanics can be described as deterministic iterative maps.Quantum evolution of a closed system may also be described by an iterative map if *s* describes the wave function of the system. However, our intent is not necessarily to describe microscopic dynamics, but rather the dynamics of variables that we consider to be relevant in describing a system. In this case we are not always guaranteed that a deterministic iterative map is sufficient. We will discuss relevant generalizations, first to stochastic maps, in Section 1.2.

Extra Credit Question 1.1.11 Show that the system of quadratic iterative maps

$$s(t) = s(t-1)^2 + k \tag{1.1.31}$$

is essentially equivalent in its dynamical properties to the iterative maps we have considered in Eq. (1.1.16).

Solution 1.1.11 Two iterative maps are equivalent in their properties if we can perform a time-independent one-to-one map of the time-dependent system states from one case to the other. We will attempt to transform the family of quadratic maps given in this problem to the one of Eq.(1.1.16) using a linear map valid at all times

$$s(t) = ms'(t) + b \tag{1.1.32}$$

By direct substitution this leads to:

$$ms'(t)+b=(ms'(t-1)+b)^2+k \qquad (1.1.33)$$

We must now choose the values of m and b so as to obtain the form of Eq. (1.1.16).

$$s'(t)=ms'(t-1)(s'(t-1)+\frac{2b}{m})+\frac{1}{m}(b^2+k-b) \qquad (1.1.34)$$

For a correct placement of minus signs in the parenthesis we need:

$$s'(t)=(-m)s'(t-1)(\left(-\frac{2b}{m}\right)-s'(t-1))+\frac{1}{m}(b^2+k-b) \qquad (1.1.35)$$

or

$$b^2-b+k=0 \qquad (1.1.36)$$

$$\frac{2b}{m}=-1 \qquad (1.1.37)$$

giving

$$b=(1\pm\sqrt{1-4k})/2 \qquad (1.1.38)$$

$$a=-m=2b=(1\pm\sqrt{1-4k}) \qquad (1.1.39)$$

We see that for $k < 1/4$we have two solutions. These solutions give all possible (positive and negative) values of a.

What about $k > 1/4$? It turns out that this case is not very interesting compared to the rich behavior for $k < 1/4$, since there are no finite fixed points,and therefore by Question 1.1.8 no 2-cycles (it is not hard to generalize this to n-cycles). To confirm this, verify that iterations diverge to $+\infty$ from any initial condition.

Note: The system of equations of this question are the ones extensively analyzed by Devaney in his excellent textbook *A First Course in Chaotic Dynamical Systems.* ▮

Extra Credit Question 1.1.12 You are given a problem to solve which when reduced to mathematical form looks like

$$s=f_c(s) \qquad (1.1.40)$$

where f is a complicated function that depends on a parameter c. You know that there is a solution of this equation in the vicinity of s_0. To solve this equation you try to iterate it (Newton's method) and it works,since you find that $f^k(s_0)$ converges nicely to a solution. Now, however, you realize that you need to solve this problem for a slightly different value of the parameter c, and when you try to iterate the equation you can't get the value of s to converge. Instead the values start to oscillate and then behave in a completely erratic way. Suggest a solution for this problem and see if it works for the function $f_c(s) = cs(1 - s)$, $c = 3.8$, $s_0 = 0.5$. A solution is given in stages (a) - (c) below.

Solution 1.1.12(a) A common resolution of this problem is to consider iterating the function:

$$h_c(s) = \alpha s + (1 - \alpha) f_c(s) \tag{1.1.41}$$

where we can adjust α to obtain rapid convergence. Note that solutions of

$$s = h_c(s) \tag{1.1.42}$$

are the same as solutions of the original problem.

Question 1.1.12(b) Explain why this could work.

Solution 1.1.12(b) The derivative of this function at a fixed point can be controlled by the value of α. It is a linear interpolation between the fixed point derivative of f_c and 1. If the fixed point is unstable and oscillating, the derivative of f_c must be less than -1 and the interpolation should help.

We can also explain this result without appealing to our work on iterative maps by noting that if the iteration is causing us to overshoot the mark, it makes sense to mix the value s we start from with the value we get from $f_c(s)$ to get a better estimate.

Question 1.1.12(c) Explain how to pick α.

Solution 1.1.12(c) If the solution is oscillating, then it makes sense to assume that the fixed point is in between successive values and the distance is revealed by how much further it gets each time;i.e., we assume that the iteration is essentially a linear map near the fixed point and we adjust α so that we compensate exactly for the overshoot of f_c.

Using two trial iterations, a linear approximation to f_c at s_0 looks like:

$$\begin{aligned} s_2 &= f_c(s_1) \approx g(s_1 - s_0) + s_0 \\ s_3 &= f_c(s_2) \approx g(s_2 - s_0) + s_0 \end{aligned} \tag{1.1.43}$$

Adopting the linear approximation as a definition of g we have:

$$g \equiv (s_3 - s_2)/(s_2 - s_1) \tag{1.1.44}$$

Set up α so that the first iteration of the modified system will take you to the desired answer:

$$s_0 = \alpha s_1 + (1 - \alpha) f_c(s_1) \tag{1.1.45}$$

or

$$s_0 - s_1 = (1 - \alpha)(f_c(s_1) - s_1) = (1 - \alpha)(s_2 - s_1) \tag{1.1.46}$$

$$(1 - \alpha) = (s_0 - s_1)/(s_2 - s_1) \tag{1.1.47}$$

To eliminate the unknown s_0 we use Eq. (1.1.43) to obtain:

$$(s_2 - s_1) = g(s_1 - s_0) + (s_0 - s_1) \qquad (1.1.48)$$

$$(s_0 - s_1) = (s_2 - s_1)/(1 - g) \qquad (1.1.49)$$

or

$$1 - \alpha = 1/(1 - g) \qquad (1.1.50)$$

$$\alpha = -g/(1 - g) = (s_2 - s_3)/(2s_2 - s_1 - s_3) \qquad (1.1.51)$$

It is easy to check, using the formula in terms of g, that the modified iteration has a zero derivative at s_0 when we use the approximate linear forms for f_c. This means we have the best convergence possible using the information from two iterations of f_c. We then use the value of α to iterate to convergence. Try it! ▮

1.2 Stochastic Iterative Maps

Many of the systems we would like to consider are described by system variables whose value at the next time step we cannot predict with complete certainty. The uncertainty may arise from many sources,including the existence of interactions and parameters that are too complicated or not very relevant to our problem. We are then faced with describing a system in which the outcome of an iteration is probabilistic and not deterministic. Such systems are called stochastic systems. There are several ways to describe such systems mathematically. One of them is to consider the outcome of a particular update to be selected from a set of possible values. The probability of each of the possible values must be specified. This description is not really a model of a single system, because each realization of the system will do something different. Instead,this is a model of a collection of systems—an ensemble.Our task is to study the properties of this ensemble.

A stochastic system is generally described by the time evolution of random variables. We begin the discussion by defining a random variable.A random variable s is defined by its probability distribution $P_s(s')$, which describes the likelihood that s has the value s'. If s is a continuous variable,then $P_s(s')ds'$ is the probability that s resides between s' and $s' + ds'$. Note that the subscript is the variable name rather than an index. For example, s might be a binary variable that can have the value +1 or −1. $P_s(1)$ is the probability that $s = 1$ and $P_s(-1)$ is the probability that $s = -1$. If s is the outcome of an unbiased coin toss, with heads called 1 and tails called −1, both of these values are 1/2. When no confusion can arise,the notation $P_s(s')$ is abbreviated to $P(s)$, where s may be either the variable or the value. The sum over all possible values of the probability must be 1.

$$\sum_{s'} P_s(s') = 1 \qquad (1.2.1)$$

In the discussion of a system described by random variables, we often would like to know the average value of some quantity $Q(s)$ that depends in a definite way on the value of the stochastic variable s. This average is given by:

$$< Q(s) >= \sum_{s'} P_s(s')Q(s') \tag{1.2.2}$$

Note that the average is a linear operation.

We now consider the case of a time-dependent random variable. Rather than describing the time dependence of the variable $s(t)$, we describe the time dependence of the probability distribution $P_s(s';t)$. Similar to the iterative map, we can consider the case where the outcome only depends on the value of the system variable at a previous time,and the transition probabilities do not depend explicitly on time. Such systems are called Markov chains. The transition probabilities from a state at a particular time to the next discrete time are written:

$$P_s(s'(t)|s'(t-1)) \tag{1.2.3}$$

P_s is used as the notation for the transition probability, since it is also the probability distribution of s at time t, given a particular value $s'(t-1)$ at the previous time. The use of a time index for the arguments illustrates the use of the transition probability. $P_s(1|1)$ is the probability that when $s = 1$ at time $t - 1$ then $s = 1$ at time t. $P_s(-1|1)$ is the probability that when $s = 1$ at time $t - 1$ then $s = -1$ at time t. The transition probabilities,along with the initial probability distribution of the system $P_s(s'; t = 0)$, determine the time-dependent ensemble that we are interested in. Assuming that we don't lose systems on the way, the transition probabilities of Eq. (1.2.3) must satisfy:

$$\sum_{s''} P_s(s''|s') = 1 \tag{1.2.4}$$

This states that no matter what the value of the system variable is at a particular time, it must reach some value at the next time.

The stochastic system described by transition probabilities can be written as an iterative map on the probability distribution $P(s)$

$$P_s(s';t) = \sum_{s'} P_s(s'|s'')P_s(s'';t-1) \tag{1.2.5}$$

It may be more intuitive to write this using the notation

$$P_s(s'(t);t) = \sum_{s'(t-1)} P_s(s'(t)|s'(t-1))P_s(s'(t-1);t-1) \tag{1.2.6}$$

in which case it may be sufficient, though hazardous, to write the abbreviated form

$$P(s(t)) = \sum_{s(t-1)} P(s(t)|s(t-1))P(s(t-1)) \tag{1.2.7}$$

It is important to recognize that the time evolution equation for the probability is linear. The linear evolution of this system (Eq. (1.2.5)) guarantees that superposition applies. If we start with an initial distribution $P(s;0) = \frac{1}{2}P^1(s;0) + \frac{1}{2}P^2(s;0)$ at time $t = 0$, then we could find the result at time t by separately looking at the evolution of each of the probabilities $P^1(s;0)$ and $P^2(s;0)$. Explicitly we can write $P(s;t) = \frac{1}{2}P^1(s;t) + \frac{1}{2}P^2(s;t)$. The meaning of this equation should be well noted. The right side of the equation is the sum of the evolved probabilities $P^1(s;t)$ and $P^2(s;t)$. This linearity is a direct consequence of the independence of different members of the ensemble and says nothing about the complexity of the dynamics.

We note that ultimately we are interested in the behavior of a particular system $s(t)$ that only has one value of s at every time t. The ensemble describes how many such systems will behave. Analytically it is easier to describe the ensemble as a whole, however, simulations may also be used to observe the behavior of a single system.

1.2.1 *Random walk*

Stochastic systems with only one binary variable might seem to be trivial, but we will devote quite a bit of attention to this problem. We begin by considering the simplest possible binary stochastic system. This is the system which corresponds to a coin toss. Ideally, for each toss there is equal probability of heads ($s = +1$) or tails ($s = -1$), and there is no memory from one toss to the next. The ensemble at each time is independent of time and has an equal probability of ±1:

$$P(s;t) = \frac{1}{2}\delta_{s,1} + \frac{1}{2}\delta_{s,-1} \tag{1.2.8}$$

where the discrete delta function is defined by

$$\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{1.2.9}$$

Since Eq. (1.2.8) is independent of what happens at all previous times, the evolution of the state variable is given by the same expression

$$P(s' \mid s) = \frac{1}{2}\delta_{s',1} + \frac{1}{2}\delta_{s',-1} \tag{1.2.10}$$

We can illustrate the evaluation of the average of a function of s at time t:

$$< Q(s) >_t = \sum_{s'=\pm 1} Q(s')P_s(s';t) = \sum_{s'=\pm 1} Q(s')\left(\frac{1}{2}\delta_{s',1} + \frac{1}{2}\delta_{s',-1}\right) = \frac{1}{2}\sum_{s'=\pm 1} Q(s') \tag{1.2.11}$$

For example, if we just take $Q(s)$ to be s itself we have the average of the system variable:

$$< s >_t = \frac{1}{2}\sum_{s'=\pm 1} s' = 0 \tag{1.2.12}$$

Question 1.2.1 Will you win more fair coin tosses if (a) you pick heads every time, or if (b) you alternate heads and tails, or if (c) you pick heads or tails at random or if (d) you pick heads and tails by some other system? Explain why.

Solution 1.2.1 In general, we cannot predict the number of coin tosses that will be won, we can only estimate it based on the chance of winning. Assuming a fair coin means that this is the best that can be done. Any of the possibilities (a)–(c) give the same chance of winning. In none of these ways of gambling does the choice you make correlate with the result of the coin toss. The only system (d) that can help is if you have some information about what the result of the toss will be,like betting on the known result *after* the coin is tossed.A way to write this formally is to write the probability distribution of the choice that you are making. This choice is also a stochastic process. Calling the choice $c(t)$, the four possibilities mentioned are:

$$(a)\quad P(c;t) = \delta_{c,1} \tag{1.2.13}$$

$$(b)\quad P(c;t) = \frac{1+(-1)^t}{2}\delta_{c,1} + \frac{1-(-1)^t}{2}\delta_{c,-1} = \mathrm{mod}_2(t)\delta_{c,1} + \mathrm{mod}_2(t+1)\delta_{c,-1} \tag{1.2.14}$$

$$(c)\quad P(c;t) = \frac{1}{2}\delta_{c,1} + \frac{1}{2}\delta_{c,-1} \tag{1.2.15}$$

$$(d)\quad P(c;t) = \delta_{c,s(t)} \tag{1.2.16}$$

It is sufficient to show that the average probability of winning is the same in each of (a)–(c) and is just 1/2. We follow through the manipulations in order to illustrate some concepts in the treatment of more than one stochastic variable. We have to sum over the probabilities of each of the possible values of the coin toss and each of the values of the choices, adding up the probability that they coincide at a particular time t:

$$< \delta_{c,s} >= \sum_{s'}\sum_{c'} \delta_{c',s'} P_s(s';t) P_c(c',t) \tag{1.2.17}$$

This expression assumes that the values of the coin toss and the value of the choice are independent,so that the joint probability of having a particular value of s and a particular value of c is the product of the probabilities of each of the variables independently:

$$P_{s,c}(s',c';t) = P_s(s';t)P_c(c';t) \tag{1.2.18}$$

—the probabilities-of-independent-variables factor. This is valid in cases (a)–(c) and not in case (d), where the probability of c occurring is explicitly a function of the value of s.

We evaluate the probability of winning in each case (a) through (c) using

$$\begin{aligned} <\delta_{c,s}> &= \sum_{s'}\sum_{c'}\delta_{c',s'}\left(\tfrac{1}{2}\delta_{s',1}+\tfrac{1}{2}\delta_{s',-1}\right)P_c(c';t) \\ &= \sum_{c'}\left(\tfrac{1}{2}\delta_{c',1}+\tfrac{1}{2}\delta_{c',-1}\right)P_c(c';t) \\ &= \sum_{c'}\left(\tfrac{1}{2}P_c(1;t)+\tfrac{1}{2}P_c(-1;t)\right)=\tfrac{1}{2} \end{aligned} \qquad (1.2.19)$$

where the last equality follows from the normalization of the probability (the sum over all possibilities must be 1, Eq. (1.2.1)) and does not depend at all on the distribution. This shows that the independence of the variables guarantees that the probability of a win is just 1/2.

For the last case (d) the trivial answer, that a win is guaranteed by this method of gambling, can be arrived at formally by evaluating

$$<\delta_{c,s}> = \sum_{s'}\sum_{c'}\delta_{c',s'}P_{s,c}(s',c';t) \qquad (1.2.20)$$

The value of s at time t is independent of the value of c, but the value of c depends on the value of s. The joint probability $P_{s,c}(s',c';t)$ may be written as the product of the probability of a particular value of $s = s'$ times the conditional probability $P_c(c'|s';t)$ of a particular value of $c = c'$ given the assumed value of s:

$$\begin{aligned} <\delta_{c,s}> &= \sum_{s'}\sum_{c'}\delta_{c',s'}P_s(s';t)P_c(c'\,|s';t) \\ &= \sum_{s'}\sum_{c'}\delta_{c',s'}P_s(s';t)\delta_{c',s'} = \sum_{s'}P_s(s';t)=1 \end{aligned} \qquad (1.2.21) \blacksquare$$

The next step in our analysis of the binary stochastic system is to consider the behavior of the sum of $s(t)$ over a particular number of time steps. This sum is the difference between the total number of heads and the total number of tails. It is equivalent to asking how much you will win or lose if you gamble an equal amount of money on each coin toss after a certain number of bets. This problem is known as a random walk, and we will define it as a consideration of the state variable

$$d(t)=\sum_{t'=1}^{t}s(t') \qquad (1.2.22)$$

The way to write the evolution of the state variable is:

$$P(d'\,|d)=\tfrac{1}{2}\delta_{d',d+1}+\tfrac{1}{2}\delta_{d',d-1} \qquad (1.2.23)$$

Thus a random walk considers a state variable d that can take integer values d $\{\ldots, -1,0,1,\ldots\}$. At every time step, $d(t)$ can only move to a value one higher or one lower than where it is. We assume that the probability of a step to the right (higher) is equal to that of a step to the left (lower). For convenience, we assume (with no loss of generality) that the system starts at position $d(0) = 0$. This is built into Eq.(1.2.22). Because of the symmetry of the system under a shift of the origin, this is equivalent to considering any other starting point. Once we solve for the probability distribution of d at time t, because of superposition we can also find the result of evolving any initial probability distribution $P(d;t = 0)$.

We can picture the random walk as that of a drunk who has difficulty consistently moving forward. Our model of this walk assumes that the drunk is equally likely to take a step forward or backward. Starting at position 0, he moves to either +1 or −1. Let's say it was +1. Next he moves to +2 or back to 0. Let's say it was 0. Next to +1 or −1. Let's say it was +1. Next to +2 or 0. Let's say +2. Next to +3 or +1. Let's say +1. And so on.

What is the value of system variable $d(t)$ at time t? This is equivalent to asking how far has the walk progressed after t steps.Of course there is no way to know how far a particular system goes without watching it. The average distance over the ensemble of systems is the average over all possible values of $s(t)$. This average is given by applying Eq. (1.2.2) or Eq. (1.2.11) to all of the variables $s(t)$:

$$\begin{aligned} < d(t) > &= \frac{1}{2}\sum_{s(t)=\pm1} \ldots \frac{1}{2}\sum_{s(3)=\pm1}\frac{1}{2}\sum_{s(2)=\pm1}\frac{1}{2}\sum_{s(1)=\pm1} d(t) \\ &= \sum_{t'=1}^{t} < s(t') > = 0 \end{aligned} \qquad (1.2.24)$$

The average is written out explicitly on the first line using Eq.(1.2.11). The second line expression can be arrived at either directly or from the linearity of the average. The final answer is clear, since it is equally likely for the walker to move to the right as to the left.

We can also ask what is a typical distance traveled by a particular walker. By typical distance we mean how far from the starting point. This can either be defined by the average absolute value of the distance, or as is more commonly accepted,the root mean square (RMS) distance:

$$\sigma(t) = \sqrt{< d(t)^2 >} \qquad (1.2.25)$$

$$< d(t)^2 > = < \left(\sum_{t'=1}^{t} s(t')\right)^2 > = < \sum_{t',t''=1}^{t} s(t')s(t'') > = \sum_{t',t''=1}^{t} < s(t')s(t'') > \qquad (1.2.26)$$

To evaluate the average of the product of the two steps, we treat differently the case in which they are the same step and when they are different steps. When the two steps are the same one we use $s(t) = \pm1$ to obtain:

$$< s(t)^2 > = < 1 > = 1 \qquad (1.2.27)$$

Which follows from the normalization of the probability (or is obvious). To evaluate the average of the product of two steps at different times we need the joint probability of $s(t)$ and $s(t')$. This is the probability that each of them will take a particular value. Because we have assumed that the steps are *independent*, the joint probability is the product of the probabilities for each one separately:

$$P(s(t),s(t'))=P(s(t))P(s(t')) \qquad t \neq t' \qquad (1.2.28)$$

so that for example there is 1/4 chance that $s(t) = +1$ and $s(t) = -1$. The independence of the two steps leads the average of the product of the two steps to factor:

$$\begin{aligned} <s(t)s(t')> &= \sum_{s(t),s(t')} P(s(t),s(t'))s(t)s(t') \\ &= \sum_{s(t),s(t')} P(s(t))P(s(t'))s(t)s(t') \qquad t \neq t' \qquad (1.2.29) \\ &= <s(t)><s(t')> = 0 \end{aligned}$$

This is zero, since either of the averages are zero. We have the combined result:

$$<s(t)s(t')> = \delta_{t,t'} \qquad (1.2.30)$$

and finally:

$$<d(t)^2> = \sum_{t',t''=1}^{t} <s(t')s(t'')> = \sum_{t',t''=1}^{t} \delta_{t',t'} = \sum_{t'=1}^{t} 1 = t \qquad (1.2.31)$$

This gives the classic and important result that a random walk travels a typical distance that grows as the square root of the number of steps taken: $\sigma(t) = \sqrt{t}$.

We can now consider more completely the probability distribution of the position of the walker at time t. The probability distribution at $t = 0$ may be written:

$$P(d;0) = \delta_{d,0} \qquad (1.2.32)$$

After the first time step the probability distribution changes to

$$P(d;1) = \frac{1}{2}\delta_{d,1} + \frac{1}{2}\delta_{d,-1} \qquad (1.2.33)$$

this results from the definition $d(1) = s(1)$. After the second step $d(2) = s(1) + s(2)$ it is:

$$P(d;2) = \frac{1}{4}\delta_{d,2} + \frac{1}{2}\delta_{d,0} + \frac{1}{4}\delta_{d,-2} \qquad (1.2.34)$$

More generally it is not difficult to see that the probabilities are given by normalized binomial coefficients,since the number of ones chosen out of t steps is equivalent to the number of powers of x in $(1 + x)^t$. To reach a position d after t steps we must take $(t + d)/2$ steps to the right and $(t - d)/2$ steps to the left. The sum of these is the number of steps t and their difference is d. Since each choice has 1/2 probability we have:

$$P(d,t)=\frac{1}{2^t}\binom{t}{(d+t)/2}\delta_{t,d}^{oddeven}=\frac{1}{2^t}\frac{t!}{[(d+t)/2]![(t-d)/2]!}\delta_{t,d}^{oddeven}$$
$$\delta_{t,d}^{oddeven}=\frac{(1+(-1)^{t+d})}{2} \qquad (1.2.35)$$

where the unusual delta function imposes the condition that d takes only odd or only even values depending on whether t is odd or even.

Let us now consider what happens after a long time. The probability distribution spreads out, and a single step is a small distance compared to the typical distance traveled. We can consider s and t to be continuous variables where both conditions $d,t >> 1$ are satisfied. Moreover, we can also consider $|d| << t$, because the chance that all steps will be taken in one direction becomes very small. This enables us to use Sterling's approximation to the factorial

$$x! \sim \sqrt{2\pi x}\, e^{-x} x^{x}$$
$$\ln(x!) \sim x(\ln x - 1) + \ln(\sqrt{2\pi x}) \qquad (1.2.36)$$

For large t it also makes sense not to restrict d to be either odd or even. In order to allow both, we, in effect, interpolate and then take only half of the probability we have in Eq. (1.2.35). This leads to the expression:

$$P(d,t)=\frac{\sqrt{t}}{\sqrt{2\pi(t-d)(t+d)}\,2^t}\frac{t^t e^{-t}}{[(d+t)/2]^{[(d+t)/2]}[(t-d)/2]^{[(t-d)/2]}e^{-(d+t)/2-(t-d)/2}}$$
$$=\frac{(2\pi t(1-x^2))^{-1/2}}{(1+x)^{[(1+x)t/2]}(1-x)^{[(1-x)t/2]}} \qquad (1.2.37)$$

where we have defined $x = d / t$. To approximate this expression it is easier to consider it in logarithmic form:

$$\ln(P(d,t)) = -(t/2)[(1+x)\ln(1+x)+(1-x)\ln(1-x)]-(1/2)\ln(2\pi t(1-x^2))$$
$$\approx -(t/2)[(1+x)(x-x^2/2+\ldots)+(1-x)(-x-x^2/2+\ldots)]-(1/2)\ln(2\pi t+\ldots)$$
$$= -tx^2/2-\ln(\sqrt{2\pi t}) \qquad (1.2.38)$$

or exponentiating:

$$P(d,t)=\frac{1}{\sqrt{2\pi t}}e^{-d^2/2t}=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-d^2/2\sigma^2} \qquad (1.2.39)$$

The prefactor of the exponential, $1/\sqrt{2\pi}\sigma$, originates from the factor $\sqrt{2\pi x}$ in Eq. (1.2.36). It is independent of d and takes care of the normalization of the probability. The result is a Gaussian distribution. Questions 1.2.2–1.2.5 investigate higher-order corrections to the Gaussian distribution.

Question 1.2.2 In order to obtain a correction to the Gaussian distribution we must add a correction term to Sterling's approximation:

$$\begin{aligned} x! &\sim \sqrt{2\pi x}\, e^{-x} x^{x} \left(1+\frac{1}{12x}+ \quad \right) \\ \ln(x!) &\sim x(\ln x - 1) + \ln(\sqrt{2\pi x}) + \ln\left(1+\frac{1}{12x}+ \quad \right) \end{aligned} \tag{1.2.40}$$

Using this expression, find the first correction term to Eq. (1.2.37).

Solution 1.2.2 The correction term in Sterling's approximation contributes a factor to Eq. (1.2.37) which is (for convenience we write here $c = 1/12$):

$$\frac{(1+c/t)}{(1+2c/(t+d))(1+2c/(t-d))} = \left(1-\frac{3c}{t}+ \quad \right) = \left(1-\frac{1}{4t}+ \quad \right) \tag{1.2.41}$$

where we have only kept the largest correction term, neglecting d compared to t. Note that the correction term vanishes as t becomes large. ▮

Question 1.2.3 Keeping additional terms of the expansion in Eq.(1.2.38), and the result of Question 1.2.2, find the first order correction terms to the Gaussian distribution.

Solution 1.2.3 Correction terms in Eq. (1.2.38) arise from several places. We want to keep all terms that are of order $1/t$. To do this we must keep in mind that a typical distance traveled is $d \quad \sqrt{t}$, so that $x \sim 1/\sqrt{t}$. The next terms are obtained from:

$$\begin{aligned} \ln(P(d,t)) &= -(t/2)[(1+x)\ln(1+x) + (1-x)\ln(1-x)] \\ &\quad - (1/2)\ln(2\pi t(1-x^2)) + \ln(1-1/4t) \\ &\approx -(t/2)[(1+x)(x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 \ldots) \\ &\quad + (1-x)(-x - \tfrac{1}{2}x^2 - \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 \ldots)] \\ &\quad - \ln(\sqrt{2\pi t}) - (1/2)\ln(1-x^2) + \ln(1-1/4t) \\ &\approx -(t/2)[(x + x^2 - \tfrac{1}{2}x^2 - \tfrac{1}{2}x^3 + \tfrac{1}{3}x^3 + \tfrac{1}{3}x^4 - \tfrac{1}{4}x^4 \ldots) \\ &\quad + (-x + x^2 - \tfrac{1}{2}x^2 + \tfrac{1}{2}x^3 - \tfrac{1}{3}x^3 + \tfrac{1}{3}x^4 - \tfrac{1}{4}x^4 \ldots)] \\ &\quad - \ln(\sqrt{2\pi t}) + (x^2/2 + \quad) + (-1/4t + \quad) \\ &= -tx^2/2 - \ln(\sqrt{2\pi t}) - tx^4/12 + x^2/2 - 1/4t \end{aligned} \tag{1.2.42}$$

This gives us a distribution:

$$P(d,t) = \sqrt{\frac{1}{2\pi t}} e^{-d^2/2t} e^{-d^4/12t^3 + d^2/2t^2 - 1/4t} \qquad (1.2.43)$$ ▮

Question 1.2.4 What is the size of the additional factor? Estimate the size of this term as t becomes large.

Solution 1.2.4 The typical value of the variable d is its root mean square value $\sigma = \sqrt{t}$. At this value the additional term gives a factor

$$e^{1/6t} \qquad (1.2.44)$$

which approaches 1 as time increases. ▮

Question 1.2.5 What is the fraction error that we will make if we neglect this term after one hundred steps? After ten thousand steps?

Solution 1.2.5 After one hundred time steps the walker has traveled a typical distance of ten steps. We generally approximate the probability of arriving at this distance using Eq.(1.2.39). The fractional error in the probability of arriving at this distance according to Eq. (1.2.44) is $1 - e^{1/6t} \approx -1 / 6t = -0.00167$. So already at a distance of ten steps the error is less than 0.2%.

It is much less likely for the walker to arrive at the distance $2\sigma = 20$. The ratio of the probability to arrive at 20 compared to 10 is $e^{-2} / e^{-0.5}$ 0.22. If we want to know the error of this smaller probability case we would write $(1 - e^{-16/12t + 4/2t - 1/4t}) = (1 - e^{5/12t}) \approx -0.0042$, which is a larger but still small error.

After ten thousand steps the errors are smaller than the errors at one hundred steps by a factor of one hundred. ▮

1.2.2 *Generalized random walk and the central limit theorem*

We can generalize the random walk by allowing a variety of steps from the current location of the walker to sites nearby, not only to the adjacent sites and not only to integer locations. If we restrict ourselves to steps that on average are balanced left and right and are not too long ranged, we can show that all such systems have the same behavior as the simplest random walk at long enough times (and characteristically not even for very long times). This is the content of the central limit theorem. It says that summing any set of independent random variables eventually leads to a Gaussian distribution of probabilities, which is the same distribution as the one we arrived at for the random walk. The reason that the same distribution arises is that successive iteration of the probability update equation, Eq.(1.2.7),smoothes out the distribution, and the only relevant information that survives is the width of the distribution which is given by $\sigma(t)$. The proof given below makes use of a Fourier transform and can be skipped by readers who are not well acquainted with transforms. In the next section we will also include a bias in the random walk. For long times this can be described as an average motion superimposed on the unbiased random walk. We start with the unbiased random walk.

Each step of the random walk is described by the state variable $s(t)$ at time t. The probability of a particular step size is an unspecified function that is independent of time:

$$P(s;t) = f(s) \tag{1.2.45}$$

We treat the case of integer values of s. The continuum case is Question 1.2.6. The absence of bias in the random walk is described by setting the average displacement in a single step to zero:

$$< s > = \sum_{s} sf(s) = 0 \tag{1.2.46}$$

The statement above that each step is not too long ranged, is mathematically just that the mean square displacement in a single step has a well-defined value (i.e., is not infinite):

$$< s^2 > = \sum_{s} s^2 f(s) = \sigma_0^2 \tag{1.2.47}$$

Eqs. (1.2.45)–(1.2.47) hold at all times.

We can still evaluate the average of $d(t)$ and the RMS value of $d(t)$ directly using the linearity of the average:

$$< d(t) > = < \sum_{t'=1}^{t} s(t') > = t < s > = 0 \tag{1.2.48}$$

$$< d(t)^2 > = < \left(\sum_{t'=1}^{t} s(t') \right)^2 > = \sum_{t',t''=1}^{t} < s(t')s(t'') > \tag{1.2.49}$$

Since $s(t')$ and $s(t'')$ are independent for $t' \neq t''$, as in Eq. (1.2.29), the average factors:

$$< s(t')s(t'') > = < s(t') >< s(t'') > = 0 \qquad t' \neq t'' \tag{1.2.50}$$

Thus, all terms $t' \neq t''$ are zero by Eq. (1.2.46). We have:

$$< d(t)^2 > = \sum_{t'=1}^{t} < s(t')^2 > = t\sigma_0^2 \tag{1.2.51}$$

This means that the typical value of $d(t)$ is $\sigma_0\sqrt{t}$.

To obtain the full distribution of the random walk state variable $d(t)$ we have to sum the stochastic variables $s(t)$. Since $d(t) = d(t - 1) + s(t)$ the probability of transition from $d(t - 1)$ to $d(t)$ is $f(d(t) - d(t - 1))$ or:

$$P(d'|d) = f(d'-d) \tag{1.2.52}$$

We can now write the time evolution equation and iterate it t times to get $P(d;t)$.

$$P(d;t) = \sum_{d'} P(d\,|\,d')P(d';t-1) = \sum_{d'} f(d-d')P(d';t-1) \tag{1.2.53}$$

This is a convolution, so the most convenient way to effect a t fold iteration is in Fourier space. The Fourier representation of the probability and transition functions for integral d is:

$$\begin{aligned}\tilde{P}(k;t) &\equiv \sum_{d} e^{-ikd} P(d;t) \\ \tilde{f}(k) &\equiv \sum_{s} e^{-iks} f(s)\end{aligned} \tag{1.2.54}$$

We use a Fourier series because of the restriction to integer values of d. Once we solve the problem using the Fourier representation, the probability distribution is recovered from the inverse formula:

$$P(d;t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikd} \tilde{P}(k;t) \tag{1.2.55}$$

which is proved

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikd}\tilde{P}(k;t) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikd} \sum_{d'} e^{-ikd'} P(d';t) \\ = \frac{1}{2\pi}\sum_{d'} P(d';t)\int_{-\pi}^{\pi} dk e^{ik(d-d')} &= \sum_{d'} P(d';t)\delta_{d,d'} = P(d;t)\end{aligned} \tag{1.2.56}$$

using the expression:

$$\delta_{d,d'} = \frac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ik(d-d')} \tag{1.2.57}$$

Applying Eq. (1.2.54) to Eq. (1.2.53):

$$\begin{aligned}\tilde{P}(k;t) &= \sum_{d} e^{-ikd} \sum_{d'} f(d-d')P(d';t-1) \\ &= \sum_{d'} \sum_{d} e^{-ik(d-d')} e^{-ikd'} f(d-d')P(d';t-1) \\ &= \sum_{d'} \sum_{d''} e^{-ikd''} e^{-ikd'} f(d'')P(d';t-1) \\ &= \sum_{d'} e^{-ikd''} f(d'') \sum_{d'} e^{-ikd'} P(d';t-1) = \tilde{f}(k)\tilde{P}(k;t-1)\end{aligned} \tag{1.2.58}$$

we can iterate the equation to obtain:

$$\tilde{P}(k;t) = \tilde{f}(k)\tilde{P}(k;t-1) = \tilde{f}(k)^t \tag{1.2.59}$$

where we use the definition $d(1) = s(1)$ that ensures that $P(d;1) = P(s;1) = f(d)$.

For large t the walker has traveled a large distance,so we are interested in variations of the probability $P(d;t)$ over large distances.Thus,in Fourier space we are concerned with small values of k. To simplify Eq.(1.2.59) for large t we expand $\tilde{f}(k)$ near $k = 0$. From Eq.(1.2.54) we can directly evaluate the derivatives of $\tilde{f}(k)$ at $k = 0$ in terms of averages:

$$\left.\frac{d^n\tilde{f}(k)}{d^nk}\right|_{k=0} = \sum_s (-is)^n f(s) = (-i)^n < s^n > \tag{1.2.60}$$

We can use this expression to evaluate the terms of a Taylor expansion of $\tilde{f}(k)$:

$$\tilde{f}(k) = \tilde{f}(0) + \left.\frac{\partial\tilde{f}(k)}{\partial k}\right|_{k=0} k + \frac{1}{2}\left.\frac{\partial^2\tilde{f}(k)}{\partial k^2}\right|_{k=0} k^2 + \ldots \tag{1.2.61}$$

$$\tilde{f}(k) = <1> - i <s> k - \frac{1}{2} <s^2> k^2 + \ldots \tag{1.2.62}$$

Using the normalization of the probability $(< 1 > = 1)$,and Eqs.(1.2.46) and (1.2.47), gives us:

$$\tilde{P}(k;t) = \left(1 - \tfrac{1}{2}\sigma_0^2 k^2 + \ldots\right)^t \tag{1.2.63}$$

We must now remember that a typical value of $d(t)$,from its RMS value, is $\sigma_0\sqrt{t}$. By the properties of the Fourier transform,this implies that a typical value of k that we must consider in Eq.(1.2.63) varies with time as $1/\sqrt{t}$. The next term in the expansion,cubic in k, would give rise to a term that is smaller by this factor, and therefore becomes unimportant at long times. If we write $k = q/\sqrt{t}$, then it becomes clearer how to write Eq. (1.2.63) using a limiting expression for large t:

$$\tilde{P}(k;t) = \left(1 - \tfrac{1}{2}\frac{\sigma_0^2 q^2}{t} + \ldots\right)^t \sim e^{-\sigma_0^2 q^2/2} = e^{-t\sigma_0^2 k^2/2} \tag{1.2.64}$$

This Gaussian, when Fourier transformed back to an expression in d, gives us a Gaussian as follows:

$$P(d;t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikd} e^{-t\sigma_0^2 k^2/2} \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{ikd} e^{-t\sigma_0^2 k^2/2} \tag{1.2.65}$$

We have extended the integral because the decaying exponential becomes narrow as t increases. The integral is performed by completing the square in the exponent, giving:

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{-d^2/2t\sigma_0^2} e^{-(t\sigma_0^2 k^2 - 2ikd - d^2/t\sigma_0^2)/2} = \frac{1}{\sqrt{2\pi t\sigma_0^2}} e^{-d^2/2t\sigma_0^2} \tag{1.2.66}$$

or equivalently:

$$P(d;t) = \frac{1}{\sqrt{2\pi\sigma(t)^2}} e^{-d^2/2\sigma(t)^2} \tag{1.2.67}$$

which is the same as Eq. (1.2.39).

Question 1.2.6 Prove the central limit theorem when s takes a continuum of values.

Solution 1.2.6 The proof follows the same course as the integer valued case. We must define the appropriate averages,and the transform. The average of s is still zero, and the mean square displacement is defined similarly:

$$< s > = \int ds s f(s) = 0 \tag{1.2.46′}$$

$$< s^2 > = \int ds s^2 f(s) = \sigma_0^2 \tag{1.2.47′}$$

To avoid problems of notation we substitute the variable x for the state variable d:

$$< x(t) > = < \sum_{t'=1}^{t} s(t') > = t < s > = 0 \tag{1.2.48′}$$

Skipping steps that are the same we find:

$$< x(t)^2 > = < \left(\sum_{t'=1}^{t} s(t') \right)^2 > = \sum_{t'=1}^{t} < s(t')^2 > = t\sigma_0^2 \tag{1.2.51′}$$

since $s(t')$ and $s(t'')$ are still independent for $t' \neq t''$. Eq. (1.2.53) is also essentially unchanged:

$$P(x;t) = \int dx' f(x - x') P(x';t-1) \tag{1.2.53′}$$

The transform and inverse transform must now be defined using

$$\begin{gathered} \tilde{P}(k;t) \equiv \int dx e^{-ikx} P(x;t) \\ \tilde{f}(k) \equiv \int ds e^{-iks} f(s) \end{gathered} \tag{1.2.54′}$$

$$P(d;t) = \frac{1}{2\pi}\int dke^{ikd}\tilde{P}(k;t) \qquad (1.2.55')$$

The latter is proved using the properties of the Dirac (continuum) delta function:

$$\begin{aligned}\delta(x-x') &= \frac{1}{2\pi}\int dke^{ik(x-x')}\\ \int dx'\delta(x-x')g(x') &= g(x)\end{aligned} \qquad (1.2.56')$$

where the latter equation holds for an arbitrary function $g(x)$.

The remainder of the derivation carries forward unchanged. ▮

1.2.3 *Biased random walk*

We now return to the simple random walk with binary steps of ±1. The model we consider is a random walk that is biased in one direction.Each time a step is taken there is a probability P_+ for a step of +1, that is different from the probability P_- for a step of –1, or:

$$P(s;t) = P_+\delta_{s,1} + P_-\delta_{s,-1} \qquad (1.2.68)$$

$$P(d'|d) = P_+\delta_{d',d+1} + P_-\delta_{d',d-} \qquad (1.2.69)$$

where

$$P_+ + P_- = 1 \qquad (1.2.70)$$

What is the average distance traveled in time t?

$$<d(t)> = \sum_{t'=1}^{t} <s(t')> = \sum_{t'=1}^{t}(P_+ - P_-) = t(P_+ - P_-) \qquad (1.2.71)$$

This equation justifies defining the mean velocity as

$$v = P_+ - P_- \qquad (1.2.72)$$

Since we already have an average displacement it doesn't make sense to also ask for a typical displacement,as we did with the random walk—the typical displacement is the average one. However, we can ask about the spread of the displacements around the average displacement

$$\begin{aligned}\sigma(t)^2 &= <(d(t) - <d(t)>)^2> = <d(t)^2> -2<d(t)>^2 + <d(t)>^2\\ &= <d(t)^2> - <d(t)>^2\end{aligned} \qquad (1.2.73)$$

This is called the standard deviation and it reduces to the RMS distance in the unbiased case. For many purposes $\sigma(t)$ plays the same role in the biased random walk as in the unbiased random walk. From Eq. (1.2.71) and Eq. (1.2.72) the second term is $(vt)^2$. The first term is:

$$
\begin{aligned}
<d(t)^2> &= <\left(\sum_{t'=1}^{t} s(t')\right)^2> = \sum_{t',t''=1}^{t} <s(t')s(t'')> \\
&= \sum_{t',t''=1}^{t} \left(\delta_{t',t'} + (1-\delta_{t',t'})(P_+^{\;2} + P_-^{\;2} - 2P_+P_-)\right) \\
&= t + t(t-1)v^2 = t^2v^2 + t(1-v^2)
\end{aligned}
\tag{1.2.74}
$$

Substituting in Eq. (1.2.73):

$$\sigma^2 = t(1-v^2) \tag{1.2.75}$$

It is interesting to consider this expression in the two limits $v = 1$ and $v = 0$. For $v = 1$ the walk is deterministic, $P_+ = 1$ and $P_- = 0$, and there is no element of chance; the walker always walks to the right. This is equivalent to the iterative map Eq.(1.1.4).Our result Eq.(1.2.66) is that $\sigma = 0$, as it must be for a deterministic system. However, for smaller velocities,the spreading of the systems σ increases until at $v = 0$ we recover the case of the unbiased random walk.

The complete probability distribution is given by:

$$P(d;t) = P_+^{\;(d+t)/2} P_-^{\;(d-t)/2} \binom{t}{(d+t)/2} \delta_{t,d}^{oddeven} \tag{1.2.76}$$

For large t the distribution can be found as we did for the unbiased random walk. The work is left to Question 1.2.7.

Question 1.2.7 Find the long time (continuum) distribution for the biased random walk.

Solution 1.2.7 We use the Sterling approximation as before and take the logarithm of the probability. In addition to the expression from the first line of Eq. (1.2.38) we have an additional factor due to the coefficient of Eq.(1.2.76) which appears in place of the factor of $1/2^t$. We again define $x = d/t$, and divide by 2 to allow both odd and even integers. We obtain the expression:

$$
\begin{aligned}
\ln(P(d,t)) &= (t/2)[(1+x)\ln 2P_+ + (1-x)\ln 2P_-] \\
&\quad -(t/2)[(1+x)\ln(1+x) + (1-x)\ln(1-x)] - (1/2)\ln(2\pi t(1-x^2)
\end{aligned}
\tag{1.2.77}
$$

It makes the most sense to expand this around the mean of x, $<x> = v$. To simplify the notation we can use Eq. (1.2.70) and Eq. (1.2.72) to write:

$$
\begin{aligned}
P_+ &= (1+v)/2 \\
P_- &= (1-v)/2
\end{aligned}
\tag{1.2.78}
$$

With these substitutions we have:

$$
\begin{aligned}
\ln(P(d,t)) &= (t/2)[(1+x)\ln(1+v) + (1-x)\ln(1-v)] \\
&\quad -(t/2)[(1+x)\ln(1+x) + (1-x)\ln(1-x)] - (1/2)\ln(2\pi t(1-x^2))
\end{aligned}
\tag{1.2.79}
$$

We expand the first two terms in a Taylor expansion around the mean of x and expand the third term inside the logarithm. The first term of Eq.(1.2.79) has only a constant and linear term in a Taylor expansion. These cancel the constant and the first derivative of the Taylor expansion of the second term of Eq. (1.2.79) at $x = v$. Higher derivatives arise only from the second term:

$$\begin{aligned}\ln(P(d,t)) = &-(t/2)[\frac{1}{(1-v^2)}(x-v)^2 + \frac{2}{3(1-v^2)^2}(x-v)^3 + \ldots] \\ &-(1/2)\ln(2\pi t[(1-v^2) - 2v(x-v) + \ldots]) \qquad (1.2.80) \\ = &-[\frac{(d-vt)^2}{2\sigma(t)^2} + \frac{(d-vt)^3}{3\sigma(t)^4} + \ldots] - (1/2)\ln(2\pi(\sigma(t)^2 - 2v(d-vt) + \ldots))\end{aligned}$$

In the last line we have restored d and used Eq.(1.2.75). Keeping only the first terms in both expansions gives us:

$$P(d;t) = \frac{1}{\sqrt{2\pi\sigma(t)^2}} e^{-(d-vt)^2/2\sigma(t)^2} \qquad (1.2.81)$$

which is a Gaussian distribution around the mean we obtained before. This implies that aside from the constant velocity, and a slightly modified standard deviation, the distribution remains unchanged.

The second term in both expansions in Eq.(1.2.80) become small in the limit of large t, as long as we are not interested in the tail of the distribution. Values of $(d - vt)$ relevant to the main part of the distribution are given by the standard deviation, $\sigma(t)$. The second terms in Eq. (1.2.80) are thus reduced by a factor of $\sigma(t)$ compared to the first terms in the series. Since $\sigma(t)$ grows as the square root of the time, they become insignificant for long times. The convergence is slower, however, than in the unbiased random walk (Questions 1.2.2–1.2.5). ∎

Question 1.2.8 You are a manager of a casino and are told by the owner that you have a cash flow problem. In order to survive, you have to make sure that nine out of ten working days you have a profit. Assume that the only game in your casino is a roulette wheel. Bets are limited to only red or black with a 2:1 payoff. The roulette wheel has an equal number of red numbers and black numbers and one green number (the house always wins on green). Assume that people make a fixed number of 10^6 total \$1 bets on the roulette wheel in each day.

a. What is the maximum number of red numbers on the roulette wheel that will still allow you to achieve your objective?

b. With this number of red numbers, how much money do you make on average in each day?

Solution 1.2.8 The casino wins \$1 for every wrong bet and loses \$1 for every right bet. The results of bets at the casino are equivalent to a random walk with a bias given by:

$$P_+ = (N_{red} + 1)/(N_{red} + N_{black} + 1) \tag{1.2.82}$$

$$P_- = N_{black}/(N_{red} + N_{black} + 1) \tag{1.2.83}$$

where,as the manager, we consider positive the wins of the casino. The color subscripts can be used interchangeably, since the number of red and black is equal. The velocity of the random walk is given by:

$$v = 1/(2N_{red} + 1) \tag{1.2.84}$$

To calculate the probability that the casino will lose on a particular day we must sum the probability that the random walk after 10^6 steps will result in a negative number. We approximate the sum by an integral over the distribution of Eq. (1.2.81). To avoid problems of notation we replace d with y:

$$\begin{aligned} P_{loss} &= \int_{-\infty}^{0} dy P(y; t = 10^6) = \frac{1}{\sqrt{2\pi\sigma(t)^2}} \int_{-\infty}^{0} dy e^{-(y-vt)^2 / 2\sigma(t)^2} \\ &= \frac{1}{\sqrt{2\pi\sigma(t)^2}} \int_{-\infty}^{-vt} dy' e^{-(y')^2 / 2\sigma(t)^2} \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{z_0} dz e^{-z^2} = \frac{1}{2}(1 - \text{erf}(z_0)) \end{aligned} \tag{1.2.85}$$

$$\begin{aligned} z &= y' / \sqrt{2}\sigma(t) \\ z_0 &= -vt / \sqrt{2\sigma(t)^2} = -vt / \sqrt{2t(1 - v^2)} \end{aligned} \tag{1.2.86}$$

We have written the probability of loss in a day in terms of the error function erf(x)—the integral of a Gaussian defined by

$$\text{erf}(z_0) \equiv \frac{2}{\sqrt{\pi}} \int_{0}^{z_0} dz e^{-z^2} \tag{1.2.87}$$

Since

$$\text{erf}(\infty) = 1 \tag{1.2.88}$$

we have the expression

$$(1 - \text{erf}(z_0)) \equiv \frac{2}{\sqrt{\pi}} \int_{z_0}^{\infty} dz e^{-z^2} \tag{1.2.89}$$

which is also known as the complementary error function erfc(x).

To obtain the desired constraint on the number of red numbers, or equivalently on the velocity, we invert Eq. (1.2.85) to find a value of v that gives the desired $P_{loss} = 0.1$, or $\mathrm{erf}(z_0) = 0.8$. Looking up the error function or using iterative guessing on an appropriate computer gives $z_0 = 0.9062$. Inverting Eq. (1.2.86) gives:

$$v = \frac{1}{\sqrt{t/2z_0 - 1}} \approx \sqrt{2z_0/t} \tag{1.2.90}$$

The approximation holds because t is large. The numerical result is $v = 0.0013$. This gives us the desired number of each color (inverting Eq. (1.2.84)) of $N_{red} = 371$. Of course the result is a very large number and the problem of winning nine out of ten days is a very conservative problem for a casino. Even if we insist on winning ninety-nine out of one hundred days we would have $\mathrm{erf}(z_0) = 0.98$, $z_0 = 1.645$, $v = 0.0018$ and $N_{red} = 275$. The profits per day in each case are given by vt, which is approximately $1,300 and $1,800 respectively. Of course this is much less than for bets on a more realistic roulette wheel. Eventually as we reduce the chance of the casino losing and z_0 becomes larger, we might become concerned that we are describing the properties of the tail of the distribution when we calculate the fraction of days the casino might lose, and Eq. (1.2.85) will not be very accurate. However, it is not difficult to see that casinos do not have cash flow problems. ▮

In order to generalize the proof of the central limit theorem to the case of a biased random walk, we can treat the continuum case most simply by considering the system variable $\hat{x}$, where (using $d \rightarrow x$ for the continuum case):

$$\hat{x} = x - \langle x \rangle_t = x - t \langle s \rangle = x - vt \tag{1.2.91}$$

Only x is a stochastic variable on the right side, v and t are numbers. Since iterations of this variable would satisfy the conditions for the generalized random walk, the generalization of the Gaussian distribution to Eq. (1.2.81) is proved. The discrete case is more difficult to prove because we cannot shift the variable d by arbitrary amounts and continue to consider it as discrete. We can argue the discrete case to be valid on the basis of the result for the continuum case, but a separate proof can be constructed as well.

1.2.4 *Master equation approach*

The Master equation is an alternative approach to stochastic systems, an alternative to Eq. (1.2.5), that is usually applied when time is continuous. We develop it starting from the discrete time case. We can rewrite Eq. (1.2.5) in the form of a difference equation for a particular probability $P(s)$. Beginning from:

$$P(s;t) = P(s;t-1) + \left(\sum_{s'} P(s|s')P(s';t-1) - P(s;t-1) \right) \tag{1.2.92}$$

we extract the term where the system remains in the same state:

$$P(s;t) = P(s;t-1) + \left(\sum_{s' \neq s} P(s|s')P(s';t-1) + P(s|s)P(s;t-1) - P(s;t-1) \right) \quad (1.2.93)$$

We use the normalization of probability to write it in terms of the transitions away from this site:

$$P(s;t) = P(s;t-1) + \left(\sum_{s' \neq s} P(s|s')P(s';t-1) + \left(1 - \sum_{s' \neq s} P(s'|s) \right) P(s;t-1) - P(s;t-1) \right) \quad (1.2.94)$$

Canceling the terms in the bracket that refer only to the probability $P(s;t-1)$ we write this as a difference equation. On the right appear only the probabilities at different values of the state variable ($s' \neq s$):

$$P(s,t) - P(s;t-1) = \sum_{s' \neq s} \left(P(s|s')P(s';t-1) - P(s'|s)P(s;t-1) \right) \quad (1.2.95)$$

To write the continuum form we reintroduce the time difference between steps Δt.

$$\frac{P(s,t) - P(s;t-\Delta t)}{\Delta t} = \sum_{s' \neq s} \left(\frac{P(s|s')}{\Delta t} P(s';t-\Delta t) - \frac{P(s'|s)}{\Delta t} P(s;t-\Delta t) \right) \quad (1.2.96)$$

When the limit of $\Delta t \rightarrow 0$ is meaningful, it is possible to make the change to the equation

$$\dot{P}(s,t) = \sum_{s' \neq s} \left(R(s|s')P(s';t) - R(s'|s)P(s;t) \right) \quad (1.2.97)$$

Where the ratio $P(s|s')/\Delta t$ has been replaced by the rate of transition $R(s|s')$. Eq.(1.2.97) is called the Master equation and we can consider Eq.(1.2.95) as the discrete time analog.

The Master equation has a simple interpretation: The rate of change of the probability of a particular state is the total rate at which probability is being added into that state from all other states,minus the total rate at which probability is leaving the state. Probability is acting like a fluid that is flowing to or from a particular state and is being conserved,as it must be. Eq.(1.2.97) is very much like the continuity equation of fluid flow, where the density of the fluid at a particular place changes according to how much is flowing to that location or from it.We will construct and use the Master equation approach to discuss the problem of relaxation in activated processes in Section 1.4.

1.3 Thermodynamics and Statistical Mechanics

The field of thermodynamics is easiest to understand in the context of Newtonian mechanics. Newtonian mechanics describes the effect of forces on objects. Thermodynamics describes the effect of heat transfer on objects. When heat is transferred,the temperature of an object changes.Temperature and heat are also intimately related to energy. A hot gas in a piston has a high pressure and it can do mechanical work by applying a force to a piston. By Newtonian mechanics the work is directly related to a transfer of energy. The laws of Newtonian mechanics are simplest to describe using the abstract concept of a point object with mass but no internal structure. The analogous abstraction for thermodynamic laws are materials that are in equilibrium and (even better) are homogeneous. It turns out that even the description of the equilibrium properties of materials is so rich and varied that this is still a primary focus of active research today.

Statistical mechanics begins as an effort to explain the laws of thermodynamics by considering the microscopic application of Newton's laws. Microscopically, the temperature of a gas is found to be related to the kinetic motion of the gas molecules. Heat transfer is the transfer of Newtonian energy from one object to another. The statistical treatment of the many particles of a material, with a key set of assumptions, reveals that thermodynamic laws are a natural consequence of many microscopic particles interacting with each other. Our studies of complex systems will lead us to discuss the properties of systems composed of many interacting parts. The concepts and tools of statistical mechanics will play an important role in these studies, as will the laws of thermodynamics that emerge from them. Thermodynamics also begins to teach us how to think about systems interacting with each other.

1.3.1 *Thermodynamics*

Thermodynamics describes macroscopic pieces of material in equilibrium in terms of macroscopic parameters. Thermodynamics was developed as a result of experience/experiment and,like Newton's laws,is to be understood as a set of self-consistent definitions and equations. As with Newtonian mechanics, where in its simplest form objects are point particles and friction is ignored,the discussion assumes an idealization that is directly experienced only in special circumstances. However, the fundamental laws, once understood,can be widely applied. The central quantities that are to be defined and related are the energy U, temperature T, entropy S, pressure P, the mass (which we write as the number of particles) N, and volume V. For magnets,the quantities should include the magnetization M, and the magnetic field H. Other macroscopic quantities that are relevant may be added as necessary within the framework developed by thermodynamics.Like Newtonian mechanics,a key aspect of thermodynamics is to understand how systems can be acted upon or can act upon each other. In addition to the quantities that describe the state of a system, there are two quantities that describe actions that may be made on a system to change its state: work and heat transfer.

The equations that relate the macroscopic quantities are known as the zeroth, first and second laws of thermodynamics. Much of the difficulty in understanding thermodynamics arises from the way the entropy appears as an essential but counterintuitive quantity. It is more easily understood in the context of a statistical treatment included below. A second source of difficulty is that even a seemingly simple material system, such as a piece of metal in a room, is actually quite complicated thermodynamically. Under usual circumstances the metal is not in equilibrium but is emitting a vapor of its own atoms.A thermodynamic treatment of the metal requires consideration not only of the metal but also the vapor and even the air that applies a pressure upon the metal. It is therefore generally simplest to consider the thermodynamics of a gas confined in a closed (and inert) chamber as a model thermodynamic system. We will discuss this example in detail in Question 1.3.1. The translational motion of the whole system, treated by Newtonian mechanics, is ignored.

We begin by defining the concept of equilibrium.A system left in isolation for a long enough time achieves a macroscopic state that does not vary in time. The system in an unchanging state is said to be in equilibrium. Thermodynamics also relies upon a particular type of equilibrium known as thermal equilibrium. Two systems can be brought together in such a way that they interact only by transferring heat from one to the other. The systems are said to be in thermal contact. An example would be two gases separated by a fixed but thermally conducting wall. After a long enough time the system composed of the combination of the two original systems will be in equilibrium. We say that the two systems are in thermal equilibrium with each other. We can generalize the definition of thermal equilibrium to include systems that are not in contact. We say that any two systems are in thermal equilibrium with each other if they do not change their (macroscopic) state when they are brought into thermal contact. Thermal equilibrium does not imply that the system is homogeneous, for example, the two gases may be at different pressures.

The zeroth law of thermodynamics states that if two systems are in thermal equilibrium with a third they are in thermal equilibrium with each other. This is not obvious without experience with macroscopic objects. The zeroth law implies that the interaction that occurs during thermal contact is not specific to the materials,it is in some sense weak,and it matters not how many or how big are the systems that are in contact. It enables us to define the temperature T as a quantity which is the same for all systems in thermal equilibrium. A more specific definition of the temperature must wait till the second law of thermodynamics. We also define the concept of a thermal reservoir as a very large system such that any system that we are interested in, when brought into contact with the thermal reservoir, will change its state by transferring heat to or from the reservoir until it is in equilibrium with the reservoir, but the transfer of heat will not affect the temperature of the reservoir.

Quite basic to the formulation and assumptions of thermodynamics is that the macroscopic state of an isolated system in equilibrium is completely defined by a specification of three parameters: energy, mass and volume (U,N,V). For magnets we must add the magnetization M; we will leave this case for later. The confinement of the system to a volume V is understood to result from some form of containment. The state of a system can be characterized by the force per unit area—the pressure P—exerted by the system on the container or by the container on the system, which are the same. Since in equilibrium a system is uniquely described by the three quantities (U,N,V), these determine all the other quantities, such as the pressure P and temperature T. Strictly speaking, temperature and pressure are only defined for a system in equilibrium, while the quantities (U,N,V) have meaning both in and out of equilibrium.

It is assumed that for a homogeneous material, changing the size of the system by adding more material in equilibrium at the same pressure and temperature changes the mass, number of particles N, volume V and energy U, in direct proportion to each other. Equivalently, it is assumed that cutting the system into smaller parts results in each subpart retaining the same properties in proportion to each other (see Figs.1.3.1 and 1.3.2). This means that these quantities are additive for different parts of a system whether isolated or in thermal contact or full equilibrium:

$$\begin{aligned} N &= \sum_{\alpha} N^{\alpha} \\ V &= \sum_{\alpha} V^{\alpha} \\ U &= \sum_{\alpha} U^{\alpha} \end{aligned} \qquad (1.3.1)$$

where α indexes the parts of the system. This would not be true if the parts of the system were strongly interacting in such a way that the energy depended on the relative location of the parts. Properties such as (U,N,V) that are proportional to the size of the system are called extensive quantities. Intensive quantities are properties that do not change with the size of the system at a given pressure and temperature. The ratio of two extensive quantities is an intensive quantity. Examples are the particle density N/V and the energy density U/V. The assumption of the existence of extensive and intensive quantities is also far from trivial, and corresponds to the intuition that for a macroscopic object,the local properties of the system do not depend on the size of the system. Thus a material may be cut into two parts, or a small part may be separated from a large part, without affecting its local properties.

The simplest thermodynamic systems are homogeneous ones,like a gas in an inert container. However we can also use Eq.(1.3.1) for an inhomogeneous system. For example,a sealed container with water inside will reach a state where both water and vapor are in equilibrium with each other. The use of intensive quantities and the proportionality of extensive quantities to each other applies only within a single phase—a single homogeneous part of the system, either water or vapor. However, the additivity of extensive quantities in Eq. (1.3.1) still applies to the whole system. A homogeneous as well as a heterogeneous system may contain different chemical species. In this case the quantity N is replaced by the number of each chemical species N_i and the first line of Eq.(1.3.1) may be replaced by a similar equation for each species.

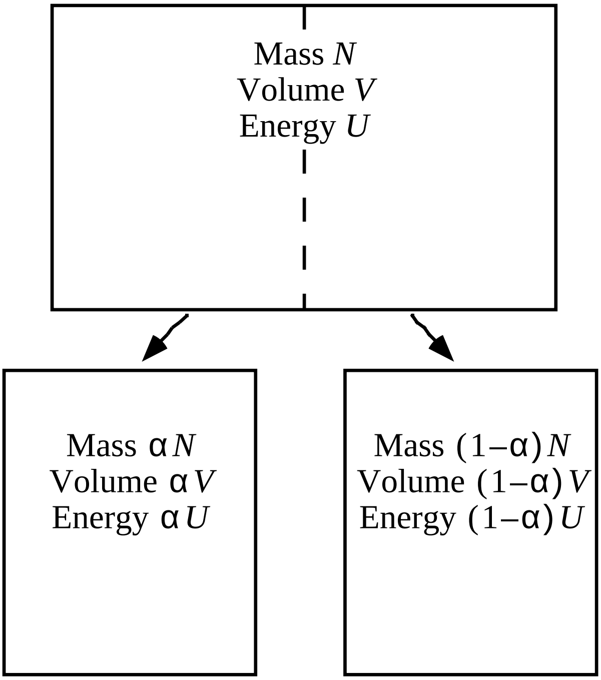


Figure 1.3.1 Thermodynamics considers macroscopic materials. A basic assumption is that cutting a system into two parts will not affect the local properties of the material and that the energy U, mass (or number of particles) N and the volume V will be divided in the same proportion. The process of separation is assumed to leave the materials under the same conditions of pressure and temperature. ▮

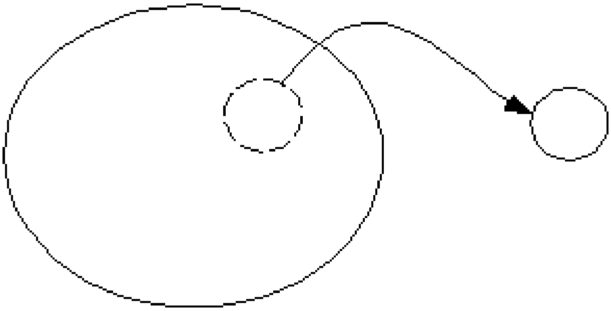


Figure 1.3.2 The assumption that the local properties of a system are unaffected by subdivision applies also to the case where a small part of a much larger system is removed. The local properties, both of the small system and of the large system are assumed to remain unchanged. Even though the small system is much smaller than the original system, the small system is understood to be a macroscopic piece of material. Thus it retains the same local properties it had as part of the larger system. ▮

The first law of thermodynamics describes how the energy of a system may change. The energy of an isolated system is conserved. There are two macroscopic processes that can change the energy of a system when the number of particles is fixed.

The first is work,in the sense of applying a force over a distance,such as driving a piston that compresses a gas. The second is heat transfer. This may be written as:

$$dU = q + w \tag{1.3.2}$$

where q is the heat transfer into the system, w is the work done on the system and U is the internal energy of the system. The differential d signifies the incremental change in the quantity U as a result of the incremental process of heat transfer and work. The work performed on a gas (or other system) is the force times the distance applied Fdx, where we write F as the magnitude of the force and dx as an incremental distance. Since the force is the pressure times the area $F = PA$, the work is equal to the pressure times the volume change or:

$$w = -PAdx = -PdV \tag{1.3.3}$$

The negative sign arises because positive work on the system,increasing the system's energy, occurs when the volume change is negative. Pressure is defined to be positive.

If two systems act upon each other, then the energy transferred consists of both the work and heat transfer. Each of these are separately equal in magnitude and opposite in sign:

$$\begin{aligned} dU_1 &= q_{21} + w_{21} \\ dU_2 &= q_{12} + w_{12} \\ q_{12} &= -q_{21} \\ w_{12} &= -w_{21} \end{aligned} \tag{1.3.4}$$

where q_{21} is the heat transfer from system 2 to system 1,and w_{21} is the work performed by system 2 on system 1. q_{12} and w_{12} are similarly defined. The last line of Eq.(1.3.4) follows from Newton's third law. The other equations follow from setting $dU = 0$ (Eq. (1.3.2)) for the total system, composed of both of the systems acting upon each other.

The second law of thermodynamics given in the following few paragraphs describes a few key aspects of the relationship of the equilibrium state with nonequilibrium states.The statement of the second law is essentially a definition and description of properties of the entropy. Entropy enables us to describe the process of approach to equilibrium. In the natural course of events,any system in isolation will change its state toward equilibrium. A system which is not in equilibrium must therefore undergo an irreversible process leading to equilibrium. The process is irreversible because the reverse process would take us away from equilibrium, which is impossible for a macroscopic system. Reversible change can occur if the state of a system in equilibrium is changed by transfer of heat or by work in such a way (slowly) that it always remains in equilibrium.

For every macroscopic state of a system (not necessarily in equilibrium) there exists a quantity S called the entropy of the system. The change in S is positive for any natural process (change toward equilibrium) of an isolated system

$$dS \geq 0 \tag{1.3.5}$$

For an isolated system, equality holds only in equilibrium when no change occurs. The converse is also true—any possible change that increases S is a natural process. Therefore, for an isolated system S achieves its maximum value for the equilibrium state.

The second property of the entropy describes how it is affected by the processes of work and heat transfer during reversible processes. The entropy is affected only by heat transfer and not by work. If we only perform work and do not transfer heat the entropy is constant. Such processes where $q = 0$ are called adiabatic processes. For adiabatic processes $dS = 0$.

The third property of the entropy is that it is extensive:

$$S = \sum_{\alpha} S^{\alpha} \qquad (1.3.6)$$

Since in equilibrium the state of the system is defined by the macroscopic quantities (U,N,V), S is a function of them—$S = S(U,N,V)$—in equilibrium. The fourth property of the entropy is that if we keep the size of the system constant by fixing both the number of particles N and the volume V, then the change in entropy S with increasing energy U is always positive:

$$\left(\frac{\partial S}{\partial U}\right)_{N,V} > 0 \qquad (1.3.7)$$

where the subscripts denote the (values of the) constant quantities. Because of this we can also invert the function $S = S(U,N,V)$ to obtain the energy U in terms of S, N and V: $U = U(S,N,V)$.

Finally, we mention that the zero of the entropy is arbitrary in classical treatments. The zero of entropy does attain significance in statistical treatments that include quantum effects.

Having described the properties of the entropy for a single system, we can now reconsider the problem of two interacting systems. Since the entropy describes the process of equilibration, we consider the process by which two systems equilibrate thermally. According to the zeroth law, when the two systems are in equilibrium they are at the same temperature. The two systems are assumed to be isolated from any other influence,so that together they form an isolated system with energy U_t and entropy S_t. Each of the subsystems is itself in equilibrium, but they are at different temperatures initially, and therefore heat is transferred to achieve equilibrium. The heat transfer is assumed to be performed in a reversible fashion—slowly. The two subsystems are also assumed to have a fixed number of particles N_1,N_2 and volume V_1,V_2. No work is done, only heat is transferred. The energies of the two systems U_1 and U_2 and entropies S_1 and S_2 are not fixed.

The transfer of heat results in a transfer of energy between the two systems according to Eq. (1.3.4), since the total energy

$$U_t = U_1 + U_2 \qquad (1.3.8)$$

is conserved, we have

$$dU_t = dU_1 + dU_2 = 0 \tag{1.3.9}$$

We will consider the processes of equilibration twice. The first time we will identify the equilibrium condition and the second time we will describe the equilibration. At equilibrium the entropy of the whole system is maximized. Variation of the entropy with respect to any internal parameter will give zero at equilibrium. We can consider the change in the entropy of the system as a function of how much of the energy is allocated to the first system:

$$\frac{dS_t}{dU_1} = \frac{dS_1}{dU_1} + \frac{dS_2}{dU_1} = 0 \tag{1.3.10}$$

in equilibrium. Since the total energy is fixed, using Eq. (1.3.9) we have:

$$\frac{dS_t}{dU_1} = \frac{dS_1}{dU_1} - \frac{dS_2}{dU_2} = 0 \tag{1.3.11}$$

or

$$\frac{dS_1}{dU_1} = \frac{dS_2}{dU_2} \tag{1.3.12}$$

in equilibrium. By the definition of the temperature,any function of the derivative of the entropy with respect to energy could be used as the temperature. It is conventional to define the temperature T using:

$$\frac{1}{T} = \left(\frac{dS}{dU}\right)_{N,V} \tag{1.3.13}$$

This definition corresponds to the Kelvin temperature scale. The units of temperature also define the units of the entropy. This definition has the advantage that heat always flows from the system at higher temperature to the system at lower temperature.

To prove this last statement, consider a natural small transfer of heat from one system to the other. The transfer must result in the two systems raising their collective entropy:

$$dS_t = dS_1 + dS_2 \geq 0 \tag{1.3.14}$$

We rewrite the change in entropy of each system in terms of the change in energy. We recall that N and V are fixed for each of the two systems and the entropy is a function only of the three macroscopic parameters (U,N,V). The change in S for each system may be written as:

$$\begin{aligned} dS_1 &= \left(\frac{\partial S}{\partial U}\right)_{N_1,V_1} dU_1 \\ dS_2 &= \left(\frac{\partial S}{\partial U}\right)_{N_2,V_2} dU_2 \end{aligned} \tag{1.3.15}$$

to arrive at:

$$\left(\frac{\partial S}{\partial U}\right)_{N_1,V_1} dU_1 + \left(\frac{\partial S}{\partial U}\right)_{N_2,V_2} dU_2 \geq 0 \tag{1.3.16}$$

or using Eq. (1.3.9) and the definition of the temperature (Eq. (1.3.13)) we have:

$$\left[\left(\frac{1}{T_1}\right) - \left(\frac{1}{T_2}\right)\right] dU_1 \geq 0 \tag{1.3.17}$$

or:

$$(T_2 - T_1)\, dU_1 \geq 0 \tag{1.3.18}$$

This implies that a natural process of heat transfer results in the energy of the first system increasing ($dU_1 > 0$) if the temperature of the second system is greater than the first ($(T_2 - T_1) > 0$), or conversely, ifthe temperature of the second system is less than the temperature of the first.

Using the definition of temperature, we can also rewrite the expression for the change in the energy of a system due to heat transfer or work, Eq.(1.3.2). The new expression is restricted to reversible processes. As in Eq. (1.3.2), N is still fixed. Considering only reversible processes means we consider only equilibrium states of the system, so we can write the energy as a function of the entropy $U = U(S,N,V)$. Since a reversible process changes the entropy and volume while keeping this function valid, we can write the change in energy for a reversible process as

$$\begin{aligned} dU &= \left(\frac{\partial U}{\partial S}\right)_{N,V} dS + \left(\frac{\partial U}{\partial V}\right)_{N,S} dV \\ &= TdS + \left(\frac{\partial U}{\partial V}\right)_{N,S} dV \end{aligned} \tag{1.3.19}$$

The first term reflects the effect of a change in entropy and the second reflects the change in volume. The change in entropy is related to heat transfer but not to work. If work is done and no heat is transferred,then the first term is zero. Comparing the second term to Eq. (1.3.2) we find

$$P = -\left(\frac{\partial U}{\partial V}\right)_{N,S} \tag{1.3.20}$$

and the incremental change in energy for a reversible process can be written:

$$dU = TdS - PdV \tag{1.3.21}$$

This relationship enables us to make direct experimental measurements of entropy changes. The work done on a system, in a reversible or irreversible process, changes the energy of the system by a known amount. This energy can then be extracted in a reversible process in the form of heat. When the system returns to its original state, we can quantify the amount of heat transferred as a form of energy. Measured heat transfer can then be related to entropy changes using $q = TdS$.

Our treatment of the fundamentals of thermodynamics was brief and does not contain the many applications necessary for a detailed understanding. The properties of S that we have described are sufficient to provide a systematic treatment of the thermodynamics of macroscopic bodies. However, the entropy is more understandable from a microscopic (statistical) description of matter. In the next section we introduce the statistical treatment that enables contact between a microscopic picture and the macroscopic thermodynamic treatment of matter. We will use it to give microscopic meaning to the entropy and temperature.Once we have developed the microscopic picture we will discuss two applications. The first application, the ideal gas, is discussed in Section 1.3.3. The discussion of the second application,the Ising model of magnetic systems, is postponed to Section 1.6.

1.3.2 *The macroscopic state from microscopic statistics*

In order to develop a microscopic understanding of the macroscopic properties of matter we must begin by restating the nature of the systems that thermodynamics describes. Even when developing a microscopic picture, the thermodynamic assumptions are relied upon as guides. Macroscopic systems are assumed to have an extremely large number N of individual particles (e.g.,at a scale of 10^{23}) in a volume V. Because the size of these systems is so large,they are typically investigated by considering the limit of $N \rightarrow \infty$ and $V \rightarrow \infty$, while the density $n = N/V$ remains constant. This is called the thermodynamic limit. Various properties of the system are separated into extensive and intensive quantities. Extensive quantities are proportional to the size of the system. Intensive quantities are independent of the size of the system. This reflects the intuition that local properties of a macroscopic object do not depend on the size of the system. As in Figs.1.3.1 and 1.3.2, the system may be cut into two parts, or a small part may be separated from a large part without affecting its local properties.

The total energy U of an isolated system in equilibrium, along with the number of particles N and volume V, defines the macroscopic state (macrostate) of an isolated system in equilibrium. Microscopically, the energy of the system E is given in classical mechanics in terms of the complete specification of the individual particle positions, momenta and interaction potentials. Together these define the microscopic state (microstate) of the system. The microstate is defined differently in quantum mechanics but similar considerations apply. When we describe the system microscopically we use the notation E rather than U to describe the energy. The reason for this difference is that macroscopically the energy U has some degree of fuzziness in its definition, though the degree of fuzziness will not enter into our considerations. Moreover, U may also be used to describe the energy of a system that is in thermal equilibrium with another system. However, thinking microscopically, the energy of such a system is not well defined,since thermal contact allows the exchange of energy between the two systems. We should also distinguish between the microscopic and macroscopic concepts of the number of particles and the volume,but since we will not make use of this distinction, we will not do so.

There are many possible microstates that correspond to a particular macrostate of the system specified only by U,N,V. We now make a key assumption of statistical mechanics—that all of the possible microstates of the system occur with equal probability. The number of these microstates $\Omega(U,N,V)$, which by definition depends on the macroscopic parameters, turns out to be central to statistical mechanics and is directly related to the entropy. Thus it determines many of the thermodynamic properties of the system, and can be discussed even though we are not always able to obtain it explicitly.

We consider again the problem of interacting systems. As before, we consider two systems (Fig. 1.3.3) that are in equilibrium separately, with state variables (U_1,N_1,V_1) and (U_2,N_2,V_2). The systems have a number of microstates $\Omega_1(U_1,N_1,V_1)$ and $\Omega_2(U_2,N_2,V_2)$ respectively. It is not necessary that the two systems be formed of the same material or have the same functional form of $\Omega(U,N,V)$, so the function Ω is also labeled by the system index. The two systems interact in a limited way, so that they can exchange only energy. The number of particles and volume of each system remains fixed. Conservation of energy requires that the total energy $U_t = U_1 + U_2$ remains fixed, but energy may be transferred from one system to the other. As before, our objective is to identify when energy transfer stops and equilibrium is reached.

Consider the number of microstates of the whole system Ω_t. This number is a function not only of the total energy of the system but also of how the energy is allocated between the systems. So, we write $\Omega_t(U_1,U_2)$, and we assume that at any time the energy of each of the two systems is well defined. Moreover, the interaction between the two systems is sufficiently weak so that the number of states of each system

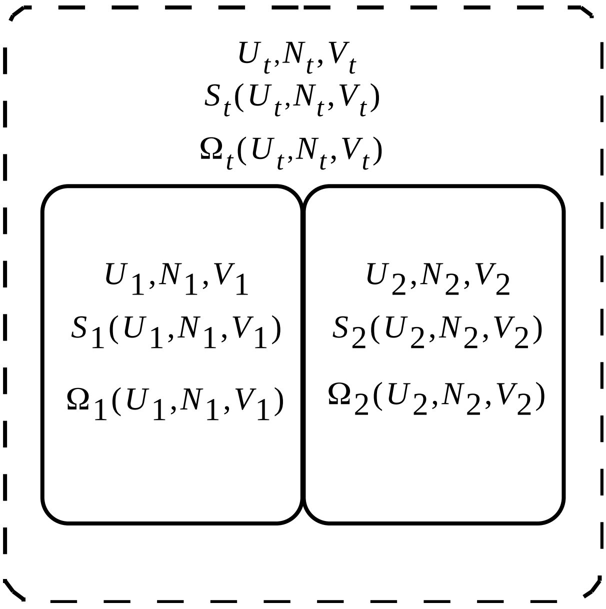


Figure 1.3.3 Illustration of a system formed out of two parts. The text discusses this system when energy is transferred from one part to the other. The transfer of energy on a microscopic scale is equivalent to the transfer of heat on a macroscopic scale, since the two systems are not allowed to change their number of particles or their volume. ▮

may be counted independently. Then the total number of microstates is the product of the number of microstates of each of the two systems separately.

$$\Omega_t(U_1,U_2) = \Omega_1(U_1)\Omega_2(U_2) \tag{1.3.22}$$

where we have dropped the arguments N and V, since they are fixed throughout this discussion. When energy is transferred,the number of microstates of each of the two systems is changed. When will the transfer of energy stop? Left on its own,the system will evolve until it reaches the most probable separation of energy. Since any particular state is equally likely, the most probable separation of energy is the separation that gives rise to the greatest possible number of states. When the number of particles is large,the greatest number of states corresponding to a particular energy separation is much larger than the number of states corresponding to any other possible separation. Thus any other possibility is completely negligible. No matter when we look at the system, it will be in a state with the most likely separation of the energy. For a macroscopic system,it is impossible for a spontaneous transfer of energy to occur that moves the system away from equilibrium.

The last paragraph implies that the transfer of energy from one system to the other stops when Ω_t reaches its maximum value. Since $U_t = U_1 + U_2$ we can find the maximum value of the number of microstates using:

$$\begin{gathered}\frac{\partial\Omega_t(U_1,U_t-U_1)}{\partial U_1}=0=\frac{\partial\Omega_1(U_1)}{\partial U_1}\Omega_2(U_t-U_1)+\Omega_1(U_1)\frac{\partial\Omega_2(U_t-U_1)}{\partial U_1}\\ 0=\frac{\partial\Omega_1(U_1)}{\partial U_1}\Omega_2(U_2)-\Omega_1(U_1)\frac{\partial\Omega_2(U_2)}{\partial U_2}\end{gathered} \tag{1.3.23}$$

or

$$\begin{gathered}\frac{1}{\Omega_1(U_1)}\frac{\partial\Omega_1(U_1)}{\partial U_1}=\frac{1}{\Omega_2(U_2)}\frac{\partial\Omega_2(U_2)}{\partial U_2}\\ \frac{\partial\ln\Omega_1(U_1)}{\partial U_1}=\frac{\partial\ln\Omega_2(U_2)}{\partial U_2}\end{gathered} \tag{1.3.24}$$

The equivalence of these quantities is analogous to the equivalence of the temperature of the two systems in equilibrium. Since the derivatives in the last equation are performed at constant N and V, it appears, by analogy to Eq. (1.3.12), that we can identify the entropy as:

$$S = k\ln(\Omega(E,N,V)). \tag{1.3.25}$$

The constant k, known as the Boltzmann constant, is needed to ensure correspondence of the microscopic counting of states with the macroscopic units of the entropy, as defined by the relationship of Eq. (1.3.13), once the units of temperature and energy are defined.

The entropy as defined by Eq.(1.3.25) can be shown to satisfy all of the properties of the thermodynamic entropy in the last section. We have argued that an isolated system evolves its macrostate in such a way that it maximizes the number of microstates that correspond to the macrostate. By Eq. (1.3.25), this is the same as the first property of the entropy in Eq. (1.3.5), the maximization of the entropy in equilibrium.

Interestingly, demonstrating the second property of the entropy, that it does not change during an adiabatic process, requires further formal developments relating entropy to information that will be discussed in Sections 1.7 and 1.8. We will connect the two discussions and thus be able to demonstrate the second property of the entropy in Chapter 8 (Section 8.3.2).

The extensive property of the entropy follows from Eq.(1.3.22). This also means that the number of states at a particular energy grows exponentially with the size of the system. More properly, we can say that experimental observation that the entropy is extensive suggests that the interaction between macroscopic materials, or parts of a single macroscopic material, is such that the microstates of each part of the system may be enumerated independently.

The number of microstates can be shown by simple examples to increase with the energy of the system. This corresponds to Eq.(1.3.7). There are also examples where this can be violated, though this will not enter into our discussions.

We consider next a second example of interacting systems that enables us to evaluate the meaning of a system in equilibrium with a reservoir at a temperature T. We consider a small part of a much larger system (Fig. 1.3.4). No assumption is necessary regarding the size of the small system; it may be either microscopic or macroscopic. Because of the contact of the small system with the large system, its energy is not

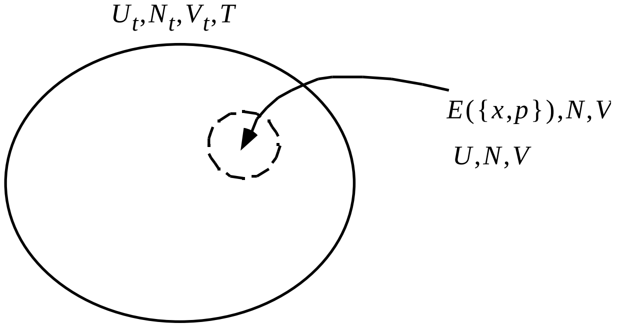


Figure 1.3.4 In order to understand temperature we consider a closed system composed of a large and small system, or equivalently a small system which is part of a much larger system. The larger system serves as a thermal reservoir transferring energy to and from the small system without affecting its own temperature. A microscopic description of this process in terms of a single microscopic state of the small system leads to the Boltzmann probability. An analysis in terms of the macroscopic state of the small system leads to the principle of minimization of the free energy to obtain the equilibrium state of a system at a fixed temperature. This principle replaces the principle of maximization of the entropy, which only applies for a closed system. ▮

always the same.Energy will be transferred back and forth between the small and large systems. The essential assumption is that the contact between the large and small system does not affect any other aspect of the description of the small system. This means that the small system is in some sense independent of the large system, despite the energy transfer. This is true if the small system is itself macroscopic, but it may also be valid for certain microscopic systems. We also assume that the small system and the large system have fixed numbers of particles and volumes.

Our objective is to consider the probability that a particular microstate of the small system will be realized. A microstate is identified by all of the microscopic parameters necessary to completely define this state. We use the notation $\{x,p\}$ to denote these coordinates. The probability that this particular state will be realized is given by the fraction of states of the whole system for which the small system attains this state. Because there is only one such state for the small system, the probability that this state will be realized is given by (proportional to) a count of the number of states of the rest of the system. Since the large system is macroscopic, we can count this number by using the macroscopic expression for the number of states of the large system:

$$P(\{x,p\}) \quad \Omega_R(U_t - \mathrm{E}(\{x,p\}),N_t - N,V_t - V) \tag{1.3.26}$$

where $E(\{x,p\}),N,V$ are the energy, number of particles and volume of the microscopic system respectively. $E(\{x,p\})$is a function of the microscopic parameters $\{x,p\}$. U_t,N_t,V_t are the energy, number of particles and volume of the whole system,including both the small and large systems. Ω_R is the entropy of the large subsystem (reservoir). Since the number of states generally grows faster than linearly as a function of the energy, we use a Taylor expansion of its logarithm (or equivalently a Taylor expansion of the entropy) to find

$$\begin{aligned}
&\ln\Omega_R(U_t - E(\{x,p\}),N_t - N,V_t - V)\\
&\quad = \ln\Omega_R(U_t,N_t - N,V_t - V) + \left(\frac{\partial \ln\Omega_R(U_t,N_t - N,V_t - V)}{\partial E_t}\right)_{N_t,V_t}(-E(\{x,p\}))\\
&\quad = \ln\Omega_R(U_t,N_t - N,V_t - V) + \frac{1}{kT}(-E(\{x,p\}))
\end{aligned} \tag{1.3.27}$$

where we have not expanded in the number of particles and the volume because they are unchanging. We take only the first term in the expansion, because the size of the small system is assumed to be much smaller than the size of the whole system. Exponentiating gives the relative probability of this particular microscopic state.

$$\Omega_R(U_t - E(\{x,p\}),N_t - N,V_t - V) = \Omega_R(U_t,N_t - N,V_t - V)e^{-E(\{x,p\})/kT} \tag{1.3.28}$$

The probability of this particular state must be normalized so that the sum over all states is one.Since we are normalizing the probability anyway, the constant coefficient does not affect the result. This gives us the Boltzmann probability distribution:

$$\begin{aligned} P(\{x,p\}) &= \frac{1}{Z} e^{-E(\{x,p\})/kT} \\ Z &= \sum_{\{x,p\}} e^{-E(\{x,p\})/kT} \end{aligned} \qquad (1.3.29)$$

Eq. (1.3.29) is independent of the states of the large system and depends only on the microscopic description of the states of the small system. It is this expression which generally provides the most convenient starting point for a connection between the microscopic description of a system and macroscopic thermodynamics. It identifies the probability that a particular microscopic state will be realized when the system has a well-defined temperature *T*. In this way it also provides a microscopic meaning to the macroscopic temperature *T*. It is emphasized that Eq.(1.3.29) describes both microscopic and macroscopic systems in equilibrium at a temperature *T*.

The probability of occurrence of a particular state should be related to the description of a system in terms of an ensemble. We have found by Eq. (1.3.29) that a system in thermal equilibrium at a temperature *T* is represented by an ensemble that is formed by taking each of the states in proportion to its Boltzmann probability. This ensemble is known as the canonical ensemble. The canonical ensemble should be contrasted with the assumption that each state has equal probability for isolated systems at a particular energy. The ensemble of fixed energy and equal a priori probability is known as the microcanonical ensemble. The canonical ensemble is both easier to discuss analytically and easier to connect with the physical world. It will be generally assumed in what follows.

We can use the Boltzmann probability and the definition of the canonical ensemble to obtain all of the thermodynamic quantities. The macroscopic energy is given by the average over the microscopic energy using:

$$U = \frac{1}{Z} \sum_{\{x,p\}} E(\{x,p\}) e^{-E(\{x,p\})/kT} \qquad (1.3.30)$$

For a macroscopic system,the average value of the energy will always be observed in any specific measurement, despite the Boltzmann probability that allows all energies. This is because the number of states of the system rises rapidly with the energy. This rapid growth and the exponential decrease of the probability with the energy results in a sharp peak in the probability distribution as a function of energy. The sharp peak in the probability distribution means that the probability of any other energy is negligible. This is discussed below in Question 1.3.1.

For an isolated macroscopic system, we were able to identify the equilibrium state from among other states of the system using the principle of the maximization of the entropy. There is a similar procedure for a macroscopic system in contact with a thermal reservoir at a fixed temperature *T*. The important point to recognize is that when we had a closed system,the energy was fixed. Now, however, the objective becomes to identify the energy at equilibrium. Of course, the energy is given by the average in

Eq.(1.3.30). However, to generalize the concept of maximizing the entropy, it is simplest to reconsider the problem of the system in contact with the reservoir when the small system is also macroscopic.

Instead of considering the probability of a particular microstate of well-defined energy E, we consider the probability of a macroscopic state of the system with an energy U. In this case, we find the equilibrium state of the system by maximizing the number of states of the whole system, or alternatively of the entropy:

$$\begin{aligned} &\ln\Omega(U,N,V)+\ln\Omega_R(U_t-U,N_t-N,V_t-V)\\ &\quad=S(U,N,V)/k+S_R(U_t-U,N_t-N,V_t-V)/k\\ &\quad=S(U,N,V)/k+S_R(U_t,N_t-N,V_t-V)/k+\frac{1}{kT}(-U)\end{aligned} \tag{1.3.31}$$

To find the equilibrium state, we must maximize this expression for the entropy of the whole system. We can again ignore the constant second term. This leaves us with quantities that are only characterizing the small system we are interested in, and the temperature of the reservoir. Thus we can find the equilibrium state by maximizing the quantity

$$S-U/T \tag{1.3.32}$$

It is conventional to rewrite this and, rather than maximizing the function in Eq. (1.3.32), to minimize the function known as the free energy:

$$F=U-TS \tag{1.3.33}$$

This suggests a simple physical significance of the process of change toward equilibrium. At a fixed temperature, the system seeks to minimize its energy and maximize its entropy at the same time. The relative importance of the entropy compared to the energy is set by the temperature. For high temperature, the entropy becomes more dominant, and the energy rises in order to increase the entropy. At low temperature, the energy becomes more dominant, and the energy is lowered at the expense of the entropy. This is the precise statement of the observation that "everything flows downhill." The energy entropy competition is a balance that is rightly considered as one of the most basic of physical phenomena.

We can obtain a microscopic expression for the free energy by an exercise that begins from a microscopic expression for the entropy:

$$S=k\ln(\Omega)=k\ln\left(\sum_{\{x,p\}}\delta_{E(\{x,p\}),U}\right) \tag{1.3.34}$$

The summation is over all microscopic states. The delta function is 1 only when $E(\{x,p\})=U$. Thus the sum counts all of the microscopic states with energy U. Strictly speaking, the δ function is assumed to be slightly "fuzzy," so that it gives 1 when $E(\{x,p\})$ differs from U by a small amount on a macroscopic scale, but by a large amount in terms of the differences between energies of microstates. We can then write

$$
\begin{aligned}
S = k\ln(\Omega) &= k\ln\left(\sum_{\{x,p\}} \delta_{E(\{x,p\}),U}\, e^{-E(\{x,p\})/kT} e^{U/kT}\right) \\
&= \frac{U}{T} + k\ln\left(\sum_{\{x,p\}} \delta_{E(\{x,p\}),U}\, e^{-E(\{x,p\})/kT}\right)
\end{aligned}
\tag{1.3.35}
$$

Let us compare the sum in the logarithm with the expression for Z in Eq.(1.3.29). We will argue that they are the same. This discussion hinges on the rapid increase in the number of states as the energy increases. Because of this rapid growth,the value of Z in Eq.(1.3.29) actually comes from only a narrow region of energy. We know from the expression for the energy average, Eq.(1.3.30),that this narrow region of energy must be at the energy U. This implies that for all intents and purposes the quantity in the brackets of Eq. (1.3.35) is equivalent to Z. This argument leads to the expression:

$$
S = \frac{U}{T} + k\ln Z \tag{1.3.36}
$$

Comparing with Eq. (1.3.33) we have

$$
F = -kT\ln Z \tag{1.3.37}
$$

Since the Boltzmann probability is a convenient starting point,this expression for the free energy is often simpler to evaluate than the expression for the entropy, Eq. (1.3.34).A calculation of the free energy using Eq.(1.3.37) provides contact between microscopic models and the macroscopic behavior of thermodynamic systems. The Boltzmann normalization Z, which is directly related to the free energy is also known as the partition function. We can obtain other thermodynamic quantities directly from the free energy. For example, we rewrite the expression for the energy Eq. (1.3.30) as:

$$
U = \frac{1}{Z}\sum_{\{x,p\}} E(\{x,p\}) e^{-\beta E(\{x,p\})} = -\frac{\partial \ln(Z)}{\partial \beta} = \frac{\partial \beta F}{\partial \beta} \tag{1.3.38}
$$

where we use the notation $\beta = 1/kT$. The entropy can be obtained using this expression for the energy and Eq. (1.3.33) or (1.3.36).

Question 1.3.1 Consider the possibility that the macroscopic energy of a system in contact with a thermal reservoir will deviate from its typical value U. To do this expand the probability distribution of macroscopic energies of a system in contact with a reservoir around this value. How large are the deviations that occur?

Solution 1.3.1 We considered Eq.(1.3.31) in order to optimize the entropy and find the typical value of the energy U. We now consider it again to find the distribution of probabilities of values of the energy around the value U similar to the way we discussed the distribution of microscopic states $\{x, p\}$ in Eq.(1.3.27). To do this we distinguish between the observed value of the energy U' and U. Note that we consider U' to be a macroscopic energy, though the same derivation could be used to obtain the distribution of microscopic energies. The probability of U' is given by:

$$P(U') \quad \Omega(U',N,V)\Omega_R(U_t-U',N_t-N,V_t-V)=e^{S(U')/k+S_R(U_t-U')/k} \tag{1.3.39}$$

In the latter form we ignore the fixed arguments N and V. We expand the logarithm of this expression around the expected value of energy U:

$$\begin{aligned} &S(U')+S_R(U_t-U') \\ &= S(U)/k+S_R(U_t-U)/k+\frac{1}{2k}\frac{d^2S(U)}{dU^2}(U-U')^2+\frac{1}{2k}\frac{d^2S(U_t-U)}{dU_t^2}(U-U')^2 \end{aligned} \tag{1.3.40}$$

where we have kept terms to second order. The first-order terms, which are of the form $(1/kT)(U' - U)$, have opposite signs and therefore cancel. This implies that the probability is a maximum at the expected energy U. The second derivative of the entropy can be evaluated using:

$$\frac{d^2S(U)}{dU^2}=\frac{d}{dU}\frac{1}{T}=-\frac{1}{T^2}\frac{1}{dU/dT}=-\frac{1}{T^2C_V} \tag{1.3.41}$$

where C_V is known as the specific heat at constant volume. For our purposes, its only relevant property is that it is an extensive quantity. We can obtain a similar expression for the reservoir and define the reservoir specific heat C_{VR}. Thus the probability is:

$$P(U') \quad e^{-(1/2kT^2)(1/C_V+1/C_{VR})(U-U')^2} \approx e^{-(1/2kT^2)(1/C_V)(U-U')^2} \tag{1.3.42}$$

where we have left out the (constant) terms that do not depend on U'. Because C_V and C_{VR} are extensive quantities and the reservoir is much bigger than the small system, we can neglect $1/C_{VR}$ compared to $1/C_V$. The result is a Gaussian distribution (Eq. (1.2.39)) with a standard deviation

$$\sigma = T\sqrt{kC_V} \tag{1.3.43}$$

This describes the characteristic deviation of the energy U' from the average or typical energy U. However, since C_V is extensive, the square root means that the deviation is proportional to $\sqrt{N}$. Note that the result is consistent with a random walk of N steps. So for a large system of $N \quad 10^{23}$ particles, the possible deviation in the energy is smaller than the energy by a factor of (we are neglecting everything but the N dependence) 10^{12}—i.e., it is undetectable. Thus the energy of a thermodynamic system is very well defined. ∎

1.3.3 *Kinetic theory of gases and pressure*

In the previous section, we described the microscopic analog of temperature and entropy. We assumed that the microscopic analog of energy was understood, and we developed the concept of free energy and its microscopic analog. One quantity that we have not discussed microscopically is the pressure. Pressure is a Newtonian concept—the force per unit area. For various reasons,it is helpful for us to consider the microscopic origin of pressure for the example of a simplified model of a gas called an ideal gas. In Question 1.3.2 we use the ideal gas as an example of the thermodynamic and statistical analysis of materials.

An ideal gas is composed of indistinguishable point particles with a mass m but with no internal structure or size. The interaction between the particles is neglected, so that the energy is just their kinetic energy. The particles do interact with the walls of the container in which the gas is confined. This interaction is simply that of reflection—when the particle is incident on a wall, the component of its velocity perpendicular to the wall is reversed.Energy is conserved. This is in accordance with the expectation from Newton's laws for the collision of a small mass with a much larger mass object.

To obtain an expression for the pressure, we must suffer with some notational hazards,as the pressure P, probability of a particular velocity $P(v)$ and momentum of a particular particle $\mathbf{p}_i$ are all designated by the letter P but with different case, arguments or subscripts.A bold letter $\mathbf{F}$ is used briefly for the force,and otherwise F is used for the free energy. We rely largely upon context to distinguish them. Since the objective of using an established notation is to make contact with known concepts,this situation is sometimes preferable to introducing a new notation.

Because of the absence of collisions between different particles of the gas, there is no communication between them, and each of the particles bounces around the container on its own course. The pressure on the container walls is given by the force per unit area exerted on the walls,as illustrated in Fig. 1.3.5. The force is given by the action of the wall on the gas that is needed to reverse the momenta of the incident particles between t and $t + \Delta t$:

$$P = \frac{|\mathbf{F}|}{A} = \frac{1}{A\Delta t}\left|\sum_i m\Delta\mathbf{v}_i\right| \qquad (1.3.44)$$

where $|\mathbf{F}|$ is the magnitude of the force on the wall. The latter expression relates the pressure to the change in the momenta of incident particles per unit area of the wall. A is a small but still macroscopic area,so that this part of the wall is flat. Microscopic roughness of the surface is neglected. The change in velocity Δv_i of the particles during the time Δt is zero for particles that are not incident on the wall. Particles that hit the wall between t and $t + \Delta t$ are moving in the direction of the wall at time t and are near enough to the wall to reach it during Δt. Faster particles can reach the wall from farther away, but only the velocity perpendicular to the wall matters. Denoting this velocity component as v , the maximum distance is $v \ \Delta t$ (see Fig. 1.3.5).

If the particles have velocity only perpendicular to the wall and no velocity parallel to the wall,then we could count the incident particles as those in a volume $Av \ \Delta t$. We can use the same expression even when particles have a velocity parallel to the surface, because the parallel velocity takes particles out of and into this volume equally.

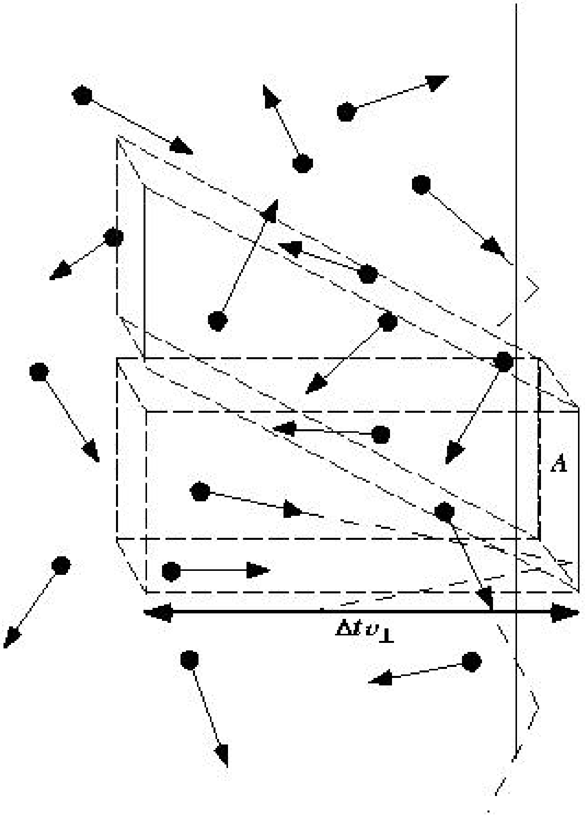


Figure 1.3.5 Illustration of a gas of ideal particles in a container near one of the walls. Particles incident on the wall are reflected, reversing their velocity perpendicular to the wall, and not affecting the other components of their velocity. The wall experiences a pressure due to the collisions and applies the same pressure to the gas. To calculate the pressure we must count the number of particles in a unit of time Δt with a particular perpendicular velocity v that hit an area A. This is equivalent to counting the number of particles with the velocity v in the box shown with one of its sides of length $\Delta t v$. Particles with velocity v will hit the wall if and only if they are in the box. The same volume of particles applies if the particles also have a velocity parallel to the surface, since this just skews the box, as shown, leaving its height and base area the same. ▮

Another way to say this is that for a particular parallel velocity we count the particles in a sheared box with the same height and base and therefore the same volume. The total number of particles in the volume, $(N / V)Av\ \Delta t$, is the volume times the density (N / V).

Within the volume $Av\ \Delta t$, the number of particles that have the velocity v is given by the number of particles in this volume times the probability $P(v\ )$ that a particle has its perpendicular velocity component equal to v . Thus the number of particles incident on the wall with a particular velocity perpendicular to the wall v is given by

$$\frac{N}{V} AP(v\ \ )v\ \ \Delta t \tag{1.3.45}$$

The total change in momentum is found by multiplying this by the change in momentum of a single particle reflected by the collision, $2mv$, and integrating over all velocities.

$$\left|\sum_i m\Delta \mathbf{v}_i\right| = \frac{1}{V} NA\Delta t \int_0^\infty dv\ \ P(v\ \ )v\ \ (2mv\ \ ) \tag{1.3.46}$$

Divide this by $A\Delta t$ to obtain the change in momentum per unit time per unit area, which is the pressure (Eq. (1.3.44)),

$$P = \frac{1}{V} N \int_0^\infty dv\ \ P(v\ \ )v\ \ (2mv\ \ ) \tag{1.3.47}$$

We rewrite this in terms of the average squared velocity perpendicular to the surface

$$P = \frac{N}{V} m 2 \int_0^\infty dv\ \ P(v\ \ )v\ \ ^2 = \frac{N}{V} m \int_{-\infty}^\infty dv\ \ P(v\ \ )v\ \ ^2 = \frac{N}{V} m < v\ \ ^2 > \tag{1.3.48}$$

where the equal probability of having positive and negative velocities enables us to extend the integral to $-\infty$ while eliminating the factor of two. We can rewrite Eq.(1.3.48) in terms of the average square magnitude of the total velocity. There are three components of the velocity (two parallel to the surface). The squares of the velocity components add to give the total velocity squared and the averages are equal:

$$< v^2 > = < v\ \ ^2 + v_2^2 + v_3^2 > = 3 < v\ \ ^2 > \tag{1.3.49}$$

where v is the magnitude of the particle velocity. The pressure is:

$$P = \frac{N}{V} m \frac{1}{3} < v^2 > \tag{1.3.50}$$

Note that the wall does not influence the probability of having a particular velocity nearby. Eq. (1.3.50) is a microscopic expression for the pressure, which we can calculate using the Boltzmann probability from Eq. (1.3.29). We do this as part of Question 1.3.2.

Question 1.3.2 Develop the statistical description of the ideal gas by obtaining expressions for the thermodynamic quantities Z, F, U, S and P, in terms of N, V, and T. For hints read the first three paragraphs of the solution.

Solution 1.3.2 The primary task of statistics is counting. To treat the ideal gas we must count the number of microscopic states to obtain the entropy, or sum over the Boltzmann probability to obtain Z and the free energy. The ideal gas presents us with two difficulties. The first is that each particle has a continuum of possible locations. The second is that we must treat the particles as microscopically indistinguishable. To solve the first problem, we have to set some interval of position at which we will call a particle here different from a particle there. Moreover, since a particle at any location may have many different velocities, we must also choose a difference of velocities that will be considered as distinct. We define the interval of position to be Δx and the interval of momentum to be Δp. In each spatial dimension,the positions between x and $x + \Delta x$ correspond to a single state,and the momenta between p and $p + \Delta p$ correspond to a single state. Thus we consider as one state of the system a particle which has position and momenta in a six-dimensional box of a size $\Delta x^3 \Delta p^3$. The size of this box enters only as a constant in classical statistical mechanics, and we will not be concerned with its value. Quantum mechanics identifies it with $\Delta x^3 \Delta p^3 = h^3$, where h is Planck's constant, and for convenience we adopt this notation for the unit volume for counting.

There is a subtle but important choice that we have made. We have chosen to make the counting intervals have a fixed width Δp in the momentum. From classical mechanics,it is not entirely clear that we should make the intervals of fixed width in the momentum or, for example,make them fixed in the energy ΔE. In the latter case we would count a single state between E and $E + \Delta E$. Since the energy is proportional to the square of the momentum,this would give a different counting. Quantum mechanics provides an unambiguous answer that the momentum intervals are fixed.

To solve the problem of the indistinguishability of the particles, we must remember every time we count the number of states of the system to divide by the number of possible ways there are to interchange the particles, which is $N!$.

The energy of the ideal gas is given by the kinetic energy of all of the particles:

$$E(\{x,p\}) = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = \sum_{i=1}^{N} \frac{p_i^{\,2}}{2m} \qquad (1.3.51)$$

where the velocity and momentum of a particle are three-dimensional vectors with magnitude v_i and p_i respectively. We start by calculating the partition function (Boltzmann normalization) Z from Eq. (1.3.29)

$$Z = \frac{1}{N!} \sum_{\{x,p\}} e^{-\sum\limits_{i=1}^{N} \frac{p_i^{\,2}}{2mkT}} = \frac{1}{N!} \int e^{-\sum\limits_{i=1}^{N} \frac{p_i^{\,2}}{2mkT}} \prod_{i=1}^{N} \frac{d^3x_i d^3p_i}{h^3} \qquad (1.3.52)$$

where the integral is to be evaluated over all possible locations of each of the N particles of the system. We have also included the correction to over-counting, $N!$. Since the particles do not see each other, the energy is a sum over each particle energy. The integrals separate and we have:

$$Z = \frac{1}{N!}\left(\frac{1}{h^3}\int e^{-\frac{p^2}{2mkT}} d^3x d^3p\right)^N \tag{1.3.53}$$

The position integral gives the volume V, immediately giving the dependence of Z on this macroscopic quantity. The integral over momentum can be evaluated giving:

$$\begin{aligned}\int e^{-\frac{p^2}{2mkT}} d^3p &= 4\pi\int_0^\infty p^2 dp e^{-\frac{p^2}{2mkT}} = 4\pi(2mkT)^{3/2}\int_0^\infty y^2 dy e^{-y^2}\\ &= 4\pi(2mkT)^{3/2}\left(-\frac{\partial}{\partial a}\Big|_{a=1}\right)\int_0^\infty dy e^{-ay^2} = 4\pi(2mkT)^{3/2}\left(-\frac{\partial}{\partial a}\Big|_{a=1}\right)\frac{1}{2}\sqrt{\frac{\pi}{a}}\\ &= (2\pi mkT)^{3/2}\end{aligned} \tag{1.3.54}$$

and we have that

$$Z(V,T,N) = \frac{V^N}{N!}\left(2\pi mkT/h^2\right)^{3N/2} \tag{1.3.55}$$

We could have simplified the integration by recognizing that each component of the momentum p_x,p_y and p_z can be integrated separately, giving $3N$ independent one-dimensional integrals and leading more succinctly to the result. The result can also be written in terms of a natural length λ(T) that depends on temperature (and mass):

$$\lambda(T) = (h^2/2\pi mkT)^{1/2} \tag{1.3.56}$$

$$Z(V,T,N) = \frac{V^N}{N!\lambda(T)^{3N}} \tag{1.3.57}$$

From the partition function we obtain the free energy, making use of Sterling's approximation (Eq. (1.2.36)):

$$F = kTN(\ln N - 1) - kTN\ln(V/\lambda(T)^3) \tag{1.3.58}$$

where we have neglected terms that grow less than linearly with N. Terms that vary as $\ln(N)$ vanish on a macroscopic scale. In this form it might appear that we have a problem,since the $N\ln(N)$ term from Sterling's approximation to the factorial does not scale proportional to the size of the system, and F is an extensive quantity. However, we must also note the $N\ln(V)$ term, which we can combine with the $N\ln(N)$ term so that the extensive nature is apparent:

$$F = kTN[\ln N\lambda(T)^3/V) - 1] \tag{1.3.59}$$

It is interesting that the factor of $N!$,and thus the indistinguishability of particles,is necessary for the free energy to be extensive. If the particles were distinguishable,then cutting the system in two would result in a different counting, since we would lose the states corresponding to particles switching from one part to the other. If we combined the two systems back together, there would be an effect due to the mixing of the distinguishable particles (Question 1.3.3).

The energy may be obtained from Eq. (1.3.38) (any of the forms) as:

$$U = \frac{3}{2}NkT \tag{1.3.60}$$

which provides an example of the equipartition theorem, which says that each degree of freedom (position-momentum pair) of the system carries $kT/2$ of energy in equilibrium.Each of the three spatial coordinates of each particle is one degree of freedom.

The expression for the entropy ($S = (U - F)/T$)

$$S = kN[\ln(V/N\lambda(T)^3) + 5/2] \tag{1.3.61}$$

shows that the entropy per particle S/N grows logarithmically with the volume per particleV/N. Using the expression for U, it may be written in a form $S(U,N,\mathrm{V})$.

Finally, the pressure may be obtained from Eq.(1.3.20), but we must be careful to keep N and S constant rather than T. We have

$$P = -\left.\frac{\partial U}{\partial V}\right|_{N,S} = -\frac{3}{2}Nk\left.\frac{\partial T}{\partial V}\right|_{N,S} \tag{1.3.62}$$

Taking the same derivative of the entropy Eq. (1.3.61) gives us (the derivative of S with S fixed is zero):

$$0 = \frac{1}{V} - \left.\frac{3}{2}\frac{\partial T}{\partial V}\right|_{N,S} \tag{1.3.63}$$

Substituting, we obtain the ideal gas equation of state:

$$PV = NkT \tag{1.3.64}$$

which we can also obtain from the microscopic expression for the pressure—Eq.(1.3.50). We describe two ways to do this.One way to obtain the pressure from the microscopic expression is to evaluate first the average of the energy

$$U = <E(\{x,p\})> = \sum_{i=1}^{N}\frac{1}{2}m<v_i^2> = N\frac{1}{2}m<v^2> \tag{1.3.65}$$

This may be substituted in to Eq. (1.3.60) to obtain

$$\frac{1}{2}m<v^2>=\frac{3}{2}kT \tag{1.3.66}$$

which may be substituted directly in to Eq. (1.3.50). Another way is to obtain the average squared velocity directly. In averaging the velocity, it doesn't matter which particle we choose. We choose the first particle:

$$<v_1^2>=3<v_1^2>=\frac{\frac{1}{N!}\int 3v_1^2\, e^{-\sum\limits_{i=1}^{N}\frac{p_i^2}{2mkT}}\prod\limits_{i=1}^{N}\frac{d^3x_i d^3p_i}{h^3}}{\frac{1}{N!}\int e^{-\sum\limits_{i=1}^{N}\frac{p_i^2}{2mkT}}\prod\limits_{i=1}^{N}\frac{d^3x_i d^3p_i}{h^3}} \tag{1.3.67}$$

where we have further chosen to average over only one of the components of the velocity of this particle and multiply by three. The denominator is the normalization constant Z. Note that the factor $1/N!$, due to the indistinguishability of particles, appears in the numerator in any ensemble average as well as in the denominator, and cancels. It does not affect the Boltzmann probability when issues of distinguishability are not involved.

There are $6N$ integrals in the numerator and in the denominator of Eq. (1.3.67). All integrals factor into one-dimensional integrals.Each integral in the numerator is the same as the corresponding one in the denominator, except for the one that involves the particular component of the velocity we are interested in. We cancel all other integrals and obtain:

$$<v_1^2>=3<v_1^2>=3\frac{\int v_1^2\, e^{-\frac{p_1^2}{2mkT}}dp_1}{\int e^{-\frac{p_1^2}{2mkT}}dp_1}=3\left(\frac{2kT}{m}\right)\frac{\int y^2 e^{-y^2}dy}{\int e^{-y^2}dy}=3\left(\frac{2kT}{m}\right)\left(\frac{1}{2}\right) \tag{1.3.68}$$

The integral is performed by the same technique as used in Eq.(1.3.54). The result is the same as by the other methods. ▮

Question 1.3.3 An insulated box is divided into two compartments by a partition. The two compartments contain two different ideal gases at the same pressure P and temperature T. The first gas has N_1 particles and the second has N_2 particles. The partition is punctured. Calculate the resulting change in thermodynamic parameters (N, V, U, P, S, T, F). What changes in the analysis if the two gases are the same, i.e., if they are composed of the same type of molecules?

Solution 1.3.3 By additivity the extrinsic properties of the whole system before the puncture are (Eq. (1.3.59)–Eq. (1.3.61)):

$$
\begin{aligned}
U_0 &= U_1 + U_2 = \frac{3}{2}(N_1 + N_2)kT \\
V_0 &= V_1 + V_2 \\
S_0 &= kN_1[\ln(V_1 / N_1\lambda(T)^3) + 5/2] + kN_2[\ln(V_2 / N_2\lambda(T)^3) + 5/2] \\
F_0 &= kTN_1[\ln(N_1\lambda(T)^3 / V_1) - 1] + kTN_2[\ln(N_2\lambda(T)^3 / V_2) - 1]
\end{aligned}
\tag{1.3.69}
$$

The pressure is intrinsic, so before the puncture it is (Eq. (1.3.64)):

$$P_0 = N_1kT / V_1 = N_2kT / V_2 \tag{1.3.70}$$

After the puncture, the total energy remains the same, because the whole system is isolated. Because the two gases do not interact with each other even when they are mixed, their properties continue to add after the puncture. However, each gas now occupies the whole volume, $V_1 + V_2$. The expression for the energy as a function of temperature remains the same,so the temperature is also unchanged. The pressure in the container is now additive: it is the sum of the pressure of each of the gases:

$$P = N_1kT / (V_1 + V_2) + N_2kT / (V_1 + V_2) = P_0 \tag{1.3.71}$$

i.e., the pressure is unchanged as well.

The only changes are in the entropy and the free energy. Because the two gases do not interact with each other, as with other quantities, we can write the total entropy as a sum over the entropy of each gas separately:

$$
\begin{aligned}
S &= kN_1[\ln((V_1 + V_2) / N_1\lambda(T)^3) + 5/2] \\
&\quad + kN_2[\ln((V_1 + V_2) / N_2\lambda(T)^3) + 5/2] \\
&= S_0 + (N_1 + N_2)k\ln(V_1 + V_2) - N_1k\ln(V_1) - N_2k\ln(V_2)
\end{aligned}
\tag{1.3.72}
$$

If we simplify to the case $V_1 = V_2$, we have $S = S_0 + (N_1 + N_2)k \ln(2)$. Since the energy is unchanged, by the relationship of free energy and entropy (Eq. (1.3.33)) we have:

$$F = F_0 - T(S - S_0) \tag{1.3.73}$$

If the two gases are composed of the same molecule,there is no change in thermodynamic parameters as a result of a puncture. Mathematically, the difference is that we replace Eq. (1.3.72) with:

$$S = k(N_1 + N_2)[\ln((V_1 + V_2) / (N_1 + N_2)\lambda(T)^3) + 5/2] = S_0 \tag{1.3.74}$$

where this is equal to the original entropy because of the relationship $N_1/V_1 = N_2 / V_2$ from Eq. (1.3.70). This example illustrates the effect of indistinguishability. The entropy increases after the puncture when the gases are different, but not when they are the same. ▮

Question 1.3.4 An ideal gas is in one compartment of a two-compartment sealed and thermally insulated box. The compartment it is in has a volume V_1. It has an energy U_0 and a number of particles N_0. The second compartment has volume V_2 and is empty. Write expressions for the changes in all thermodynamic parameters (N, V, U, P, S, T, F) if

a. the barrier between the two compartments is punctured and the gas expands to fill the box.

b. the barrier is moved slowly, like a piston, expanding the gas to fill the box.

Solution 1.3.4 Recognizing what is conserved simplifies the solution of this type of problem.

a. The energy U and the number of particles N are conserved. Since the volume change is given to us explicitly, the expressions for T (Eq. (1.3.60)), F (Eq. (1.3.59)), S (Eq. (1.3.61)), and P (Eq. (1.3.64)) in terms of these quantities can be used.

$$\begin{gathered} N = N_0 \\ U = U_0 \\ V = V_1 + V_2 \\ T = T_0 \\ F = kTN[\ln(N\lambda(T)^3/(V_1 + V_2)) - 1] = F_0 + kTN\ \ln(V_1 + V_2)) \\ S = kN[\ln((V_1 + V_2)/N\lambda T)^3) + 5/2] = S_0 + kN\ \ln((V_1 + V_2)/V_1) \\ P = NkT/V = NkT/(V_1 + V_2) = P_0 V_1/(V_1 + V_2) \end{gathered} \tag{1.3.75}$$

b. The process is reversible and no heat is transferred, thus it is adiabatic—the entropy is conserved. The number of particles is also conserved:

$$\begin{gathered} N = N_0 \\ S = S_0 \end{gathered} \tag{1.3.76}$$

Our main task is to calculate the effect of the work done by the gas pressure on the piston. This causes the energy of the gas to decrease, and the temperature decreases as well. One way to find the change in temperature is to use the conservation of entropy, and Eq. (1.3.61), to obtain that $V/\lambda(T)^3$ is a constant and therefore:

$$T \quad V^{-2/3} \tag{1.3.77}$$

Thus the temperature is given by:

$$T = T_0 \left(\frac{V_1 + V_2}{V_1}\right)^{-2/3} \tag{1.3.78}$$

Since the temperature and energy are proportional to each other (Eq. (1.3.60)), similarly:

$$U = U_0 \left(\frac{V_1 + V_2}{V_1}\right)^{-2/3} \tag{1.3.79}$$

The free-energy expression in Eq. (1.3.59) changes only through the temperature prefactor:

$$F = kTN[\ln(N\lambda(T)^3/V)-1] = F_0\frac{T}{T_0} = F_0\left(\frac{V_1+V_2}{V_1}\right)^{-2/3} \qquad (1.3.80)$$

Finally, the pressure (Eq. (1.6.64)):

$$P = NkT/V = P_0\frac{TV_0}{T_0V} = P_0\left(\frac{V_1+V_2}{V_1}\right)^{-5/3} \qquad (1.3.81)\ ■$$

The ideal gas illustrates the significance of the Boltzmann distribution. Consider a single particle. We can treat it either as part of the large system or as a subsystem in its own right. In the ideal gas, without any interactions, its energy would not change. Thus the particle would not be described by the Boltzmann probability in Eq. (1.3.29). However, we can allow the ideal gas model to include a weak or infrequent interaction (collision) between particles which changes the particle's energy. Over a long time compared to the time between collisions, the particle will explore all possible positions in space and all possible momenta. The probability of its being at a particular position and momentum (in a region d^3xd^3p) is given by the Boltzmann distribution:

$$\frac{e^{-\frac{p^2}{2mkT}}d^3pd^3x/h^3}{\int e^{-\frac{p^2}{2mkT}}d^3pd^3x/h^3} \qquad (1.3.82)$$

Instead of considering the trajectory of this particular particle and the effects of the (unspecified) collisions, we can think of an ensemble that represents this particular particle in contact with a thermal reservoir. The ensemble would be composed of many different particles in different boxes. There is no need to have more than one particle in the system. We do need to have some mechanism for energy to be transferred to and from the particle instead of collisions with other particles. This could happen as a result of the collisions with the walls of the box if the vibrations of the walls give energy to the particle or absorb energy from the particle. If the wall is at the temperature T, this would also give rise to the same Boltzmann distribution for the particle. The probability of a particular particle in a particular box being in a particular location with a particular momentum would be given by the same Boltzmann probability.

Using the Boltzmann probability distribution for the velocity, we could calculate the average velocity of the particle as:

$$<v^2>=3<v^2>=3\frac{\int v^2 e^{-\frac{p^2}{2mkT}} d^3p d^3x/h^3}{\int e^{-\frac{p^2}{2mkT}} d^3p d^3x/h^3} = \frac{\int v^2 e^{-\frac{p^2}{2mkT}} dp}{\int e^{-\frac{p^2}{2mkT}} dp} = \frac{3kT}{m} \quad (1.3.83)$$

which is the same result as we obtained for the ideal gas in the last part of Question 1.3.2. We could even consider one coordinate of one particle as a separate system and arrive at the same conclusion.Our description of systems is actually a description of coordinates.

There are differences when we consider the particle to be a member of an ensemble and as one particle of a gas. In the ensemble, we do not need to consider the distinguishability of particles. This does not affect any of the properties of a single particle.

This discussion shows that the ideal gas model may be viewed as quite close to the basic concept of an ensemble.Generalize the physical particle in three dimensions to a point with coordinates that describe a complete system. These coordinates change in time as the system evolves according to the rules of its dynamics.The ensemble represents this system in the same way as the ideal gas is the ensemble of the particle. The lack of interaction between the different members of the ensemble,and the existence of a transfer of energy to and from each of the systems to generate the Boltzmann probability, is the same in each of the cases. This analogy is helpful when thinking about the nature of the ensemble.

1.3.4 *Phase transitions—first and second order*

In the previous section we constructed some of the underpinnings of thermodynamics and their connection with microscopic descriptions of materials using statistical mechanics. One of the central conclusions was that by minimizing the free energy we can find the equilibrium state of a material that has a fixed number of particles, volume and temperature. Once the free energy is minimized to obtain the equilibrium state of the material, the energy, entropy and pressure are uniquely determined. The free energy is also a function of the temperature, the volume and the number of particles.

One of the important properties of materials is that they can change their properties suddenly when the temperature is changed by a small amount. Examples of this are the transition of a solid to a liquid, or a liquid to a gas. Such a change is known as a phase transition. Each well-defined state of the material is considered a particular phase. Let us consider the process of minimizing the free energy as we vary the temperature. Each of the properties of the material will, in general, change smoothly as the temperature is varied. However, special circumstances might occur when the minimization of the free energy at one temperature results in a very different set of properties of the material from this minimization at a slightly different temperature. This is illustrated in a series of frames in Fig. 1.3.6, where a schematic of the free energy as a function of some macroscopic parameter is illustrated.

The temperature at which the jump in properties of the material occurs is called the critical or transition temperature,T_c. In general,all of the properties of the material except for the free energy jump discontinuously at T_c. This kind of phase transition is known as a first-order phase transition. Some of the properties of a first-order phase transition are that the two phases can coexist at the transition temperature so that part of the material is in one phase and part in the other. An example is ice floating in water. If we start from a temperature below the transition temperature—with ice—and add heat to the system gradually, the temperature will rise until we reach the transition temperature. Then the temperature will stay fixed as the material converts from one phase to the other—from ice to water. Once the whole system is converted to the higher temperature phase, the temperature will start to increase again.

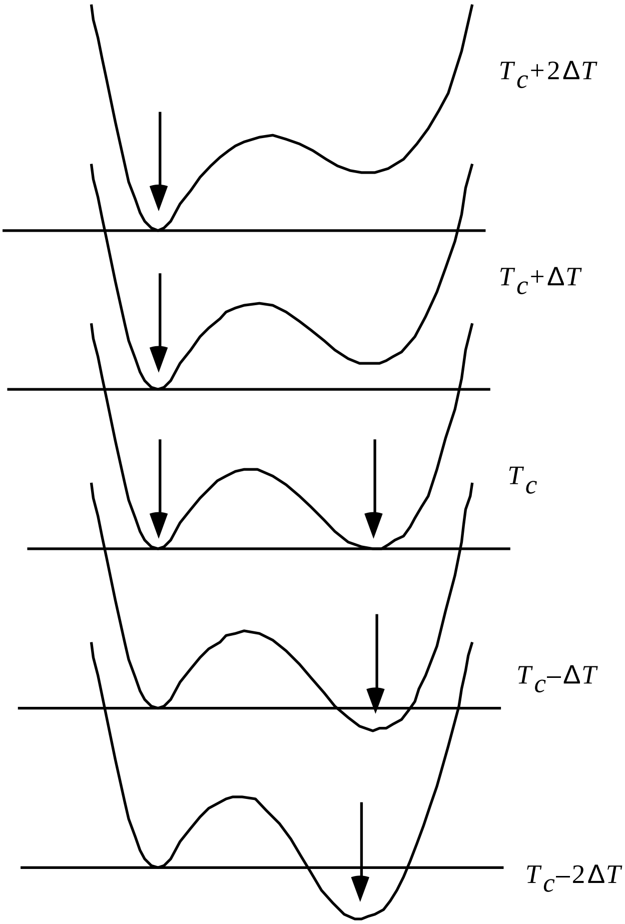


Figure 1.3.6 Each of the curves represents the variation of the free energy of a system as a function of macroscopic parameters. The different curves are for different temperatures. As the temperature is varied the minimum of the free energy all of a sudden switches from one set of macroscopic parameters to another. This is a first-order phase transition like the melting of ice to form water, or the boiling of water to form steam. Below the ice-to-water phase transition the macroscopic parameters that describe ice are the minimum of the free energy, while above the phase transition the macroscopic parameters that describe water are the minimum of the free energy. ∎

The temperature T_c at which a transition occurs depends on the number of particles and the volume of the system. Alternatively, it may be considered a function of the pressure. We can draw a phase-transition diagram (Fig. 1.3.7) that shows the transition temperature as a function of pressure. Each region of such a diagram corresponds to a particular phase.

There is another kind of phase transition, known as a second-order phase transition, where the energy and the pressure do not change discontinuously at the phase-transition point. Instead, they change continuously, but they are nonanalytic at the transition temperature.A common way that this can occur is illustrated in Fig. 1.3.8. In this case the single minimum of the free energy breaks into two minima as a function of temperature. The temperature at which the two minima appear is the transition temperature. Such a second-order transition is often coupled to the existence of first-order transitions. Below the second-order transition temperature, when the two minima exist, the variation of the pressure can change the relative energy of the two minima and cause a first-order transition to occur. The first-order transition occurs at a particular pressure $P_c(T)$ for each temperature below the second-order transition temperature. This gives rise to a line of first-order phase transitions. Above the second-order transition temperature, there is only one minimum, so that there are

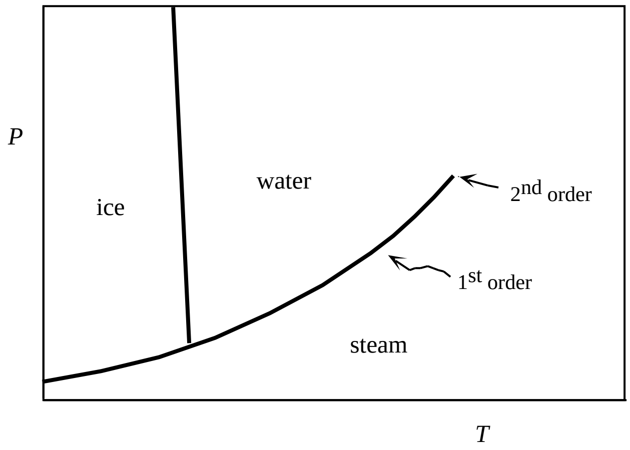


Figure 1.3.7 Schematic phase diagram of H_2O showing three phases — ice, water and steam. Each of the regions shows the domain of pressures and temperatures at which a pure phase is in equilibrium. The lines show phase transition temperatures, $T_c(P)$, or phase transition pressures, $P_c(T)$. The different ways of crossing lines have different names. Ice to water: melting; ice to steam: sublimation; water to steam: boiling; water to ice: freezing; steam to water: condensation; steam to ice: condensation to frost. The transition line from water to steam ends at a point of high pressure and temperature where the two become indistinguishable. At this high pressure steam is compressed till it has a density approaching that of water, and at this high temperature water molecules are energetic like a vapor. This special point is a second-order phase transition point (see Fig. 1.3.8). ∎

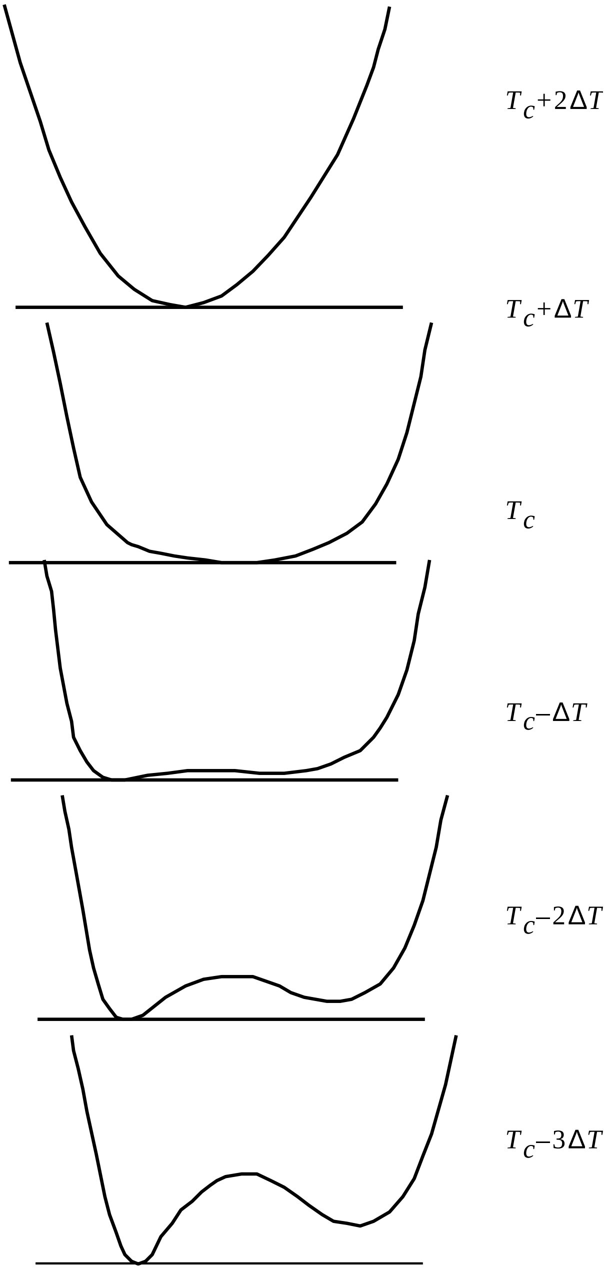


Figure 1.3.8 Similar to Fig. 1.3.6, each of the curves represents the variation of the free energy of a system as a function of macroscopic parameters. In this case, however, the phase transition occurs when two minima emerge from one. This is a second-order phase transition. Below the temperature at which the second-order phase transition occurs, varying the pressure can give rise to a first-order phase transition by changing the relative energies of the two minima (see Figs. 1.3.6 and 1.3.7). ▮

also no first-order transitions. Thus, the second-order transition point occurs as the end of a line of first-order transitions.A second-order transition is found at the end of the liquid-to-vapor phase line of water in Fig. 1.3.7.

The properties of second-order phase transitions have been extensively studied because of interesting phenomena that are associated with them. Unlike a first-order phase transition, there is no coexistence of two phases at the phase transition, because there is only one phase at that point. Instead, there exist large fluctuations in the local properties of the material at the phase transition. A suggestion of why this occurs can be seen from Fig. 1.3.8, where the free energy is seen to be very flat at the phase transition. This results in large excursions (fluctuations) of all the properties of the system except the free energy. These excursions, however, are not coherent over the whole material. Instead, they occur at every length scale from the microscopic on up. The closer a material is to the phase transition, the longer are the length scales that are affected. As the temperature is varied so that the system moves away from the transition temperature,the fluctuations disappear, first on the longest length scales and then on shorter and shorter length scales. Because at the phase transition itself even the macroscopic length scales are affected,thermodynamics itself had to be carefully rethought in the vicinity of second-order phase transitions. The methodology that has been developed, the renormalization group, is an important tool in the investigation of phase transitions. We will discuss it in Section 1.10. We note that, to be consistent with Question 1.3.1, the specific heat C_V must diverge at a second-order phase transition, where energy fluctuations can be large.

1.3.5 *Use of thermodynamics and statistical mechanics in describing the real world*

How do we generalize the notions of thermodynamics that we have just described to apply to more realistic situations? The assumptions of thermodynamics—that systems are in equilibrium and that dividing them into parts leads to unchanged local properties—do not generally apply. The breakdown of the assumptions of thermodynamics occurs for even simple materials, but are more radically violated when we consider biological organisms like trees or people. We still are able to measure their temperature. How do we extend thermodynamics to apply to these systems?

We can start by considering a system quite close to the thermodynamic ideal—a pure piece of material that is not in equilibrium. For example, a glass of water in a room. We generally have no trouble placing a thermometer in the glass and measuring the temperature of the water. We know it is not in equilibrium, because if we wait it will evaporate to become a vapor spread out throughout the room (even if we simplify by considering the room closed). Moreover, if we wait longer (a few hundred years to a few tens of thousands of years),the glass itself will flow and cover the table or flow down to the floor, and at least part of it will also sublime to a vapor. The table will undergo its own processes of deterioration. These effects will occur even in an idealized closed room without considerations of various external influences or traffic through the room. There is one essential concept that allows us to continue to apply thermodynamic principles to these materials,and measure the temperature of the water, glass or table, and generally to discover that they are at the same (or close to the same) temperature. The concept is the separation of time scales. This concept is as basic as the other principles of thermodynamics. It plays an essential role in discussions of the dynamics of physical systems and in particular of the dynamics of complex systems. The separation of time scales assumes that our observations of systems have a limited time resolution and are performed over a limited time. The processes that occur in a material are then separated into fast processes that are much faster than the time resolution of our observation, slow processes that occur on longer time scales than the duration of observation,and dynamic processes that occur on the time scale of our observation. Macroscopic averages are assumed to be averages over the fast processes. Thermodynamics allows us to deal with the slow and the fast processes but only in very limited ways with the dynamic processes. The dynamic processes are dealt with separately by Newtonian mechanics.

Slow processes establish the framework in which thermodynamics can be applied. In formal terms,the ensemble that we use in thermodynamics assumes that all the parameters of the system described by slow processes are fixed. To describe a system using statistical mechanics, we consider all of the slowly varying parameters of the system to be fixed and assume that equilibrium applies to all of the fast processes. Specifically, we assume that all possible arrangements of the fast coordinates exist in the ensemble with a probability given by the Boltzmann probability. Generally, though not always, it is the microscopic processes that are fast. To justify this we can consider that an atom in a solid vibrates at a rate of 10^{10}–10^{12} times per second,a gas molecule at room temperature travels five hundred meters per second. These are, however, only a couple of select examples.

Sometimes we may still choose to perform our analysis by averaging over many possible values of the slow coordinates. When we do this we have two kinds of ensembles—the ensemble of the fast coordinates and the ensemble of the different values of the slow coordinates. These ensembles are called the annealed and quenched ensembles. For example, say we have a glass of water in which there is an ice cube. There are fast processes that correspond to the motion of the water molecules and the vibrations of the ice molecules,and there are also slow processes corresponding to the movement of the ice in the water. Let's say we want to determine the average amount of ice. If we perform several measurements that determine the coordinates and size of the ice, we may want to average the size we find over all the measurements even though they are measurements corresponding to different locations of the ice. In contrast, if we wanted to measure the motion of the ice, averaging the measurements of location would be absurd.

Closely related to the discussion of fast coordinates is the ergodic theorem. The ergodic theorem states that a measurement performed on a system by averaging a property over a long time is the same as taking the average over the ensemble of the fast coordinates. This theorem is used to relate experimental measurements that are assumed to occur over long times to theoretically obtained averages over ensembles. The ergodic theorem is not a theorem in the sense that it has been proven in general, but rather a statement of a property that applies to some macroscopic systems and is known not to apply to others. The objective is to identify when it applies. When it does not apply, the solution is to identify which quantities may be averaged and which may not, often by separating fast and slow coordinates or equivalently by identifying quantities conserved by the fast dynamics of the system.

Experimental measurements also generally average properties over large regions of space compared to microscopic lengths. It is this spatial averaging rather than time averaging that often enables the ensemble average to stand for experimental measurements when the microscopic processes are not fast compared to the measurement time. For example, materials are often formed of microscopic grains and have many dislocations. The grain boundaries and dislocations do move, but they often change very slowly over time. When experiments are sensitive to their properties, they often average over the effects of many grains and dislocations because they do not have sufficient resolution to see a single grain boundary or dislocation.

In order to determine what is the relevant ensemble for a particular experiment, both the effect of time and space averaging must be considered. Technically, this requires an understanding of the correlation in space and time of the properties of an individual system. More conceptually, measurements that are made for particular quantities are in effect made over many independent systems both in space and in time, and therefore correspond to an ensemble average. The existence of correlation is the opposite of independence. The key question (like in the case of the ideal gas) becomes what is the interval of space and time that corresponds to an independent system. These quantities are known as the correlation length and the correlation time. If we are able to describe theoretically the ensemble over a correlation length and correlation time, then by appropriate averaging we can describe the measurement.

In summary, the program of use of thermodynamics in the real world is to use the separation of the different time scales to apply equilibrium concepts to the fast degrees of freedom and discuss their influence on the dynamic degrees of freedom while keeping fixed the slow degrees of freedom. The use of ensembles simplifies consideration of these systems by systematizing the use of equilibrium concepts to the fast degrees of freedom.

1.3.6 *From thermodynamics to complex systems*

Our objective in this book is to consider the dynamics of complex systems. While,as discussed in the previous section, we will use the principles of thermodynamics to help us in this analysis,another important reason to review thermodynamics is to recognize what complex systems are not. Thermodynamics describes macroscopic systems without structure or dynamics. The task of thermodynamics is to relate the very few macroscopic parameters to each other. It suggests that these are the only relevant parameters in the description of these systems. Materials and complex systems are both formed out of many interacting parts. The ideal gas example described a material where the interaction between the particles was weak. However, thermodynamics also describes solids, where the interaction is strong. Having decided that complex systems are not described fully by thermodynamics, we must ask, Where do the assumptions of thermodynamics break down? There are several ways the assumptions may break down, and each one is significant and plays a role in our investigation of complex systems. Since we have not yet examined particular examples of complex systems, this discussion must be quite abstract. However, it will be useful as we study complex systems to refer back to this discussion. The abstract statements will have concrete realizations when we construct models of complex systems.

The assumptions of thermodynamics separate into space-related and time-related assumptions. The first we discuss is the divisibility of a macroscopic material. Fig. 1.3.2 (page 61) illustrates the property of divisibility. In this process,a small part of a system is separated from a large part of the system without affecting the *local* properties of the material. This is inherent in the use of extensive and intensive quantities. Such divisibility is not true of systems typically considered to be complex systems. Consider, for example, a person as a complex system that cannot be separated and continue to have the same properties. In words, we would say that complex systems are formed out of not only interacting, but also interdependent parts. Since both thermodynamic and complex systems are formed out of interacting parts, it is the concept of interdependency that must distinguish them. We will dedicate a few paragraphs to defining a sense in which "interdependent" can have a more precise meaning.

We must first address a simple way in which a system may have a nonextensive energy and still not be a complex system. If we look closely at the properties of a material, say a piece of metal or a cup of water, we discover that its surface is different from the bulk. By separating the material into pieces, the surface area of the material is changed. For macroscopic materials,this generally does not affect the bulk properties of the material.A characteristic way to identify surface properties, such as the surface energy, is through their dependence on particle number. The surface energy scales as $N^{2/3}$, in contrast to the extensive bulk energy that is linear in N. This kind of correction can be incorporated directly in a slightly more detailed treatment of thermodynamics, where every macroscopic parameter has a surface term. The presence of such surface terms is not sufficient to identify a material as a complex system. For this reason, we are careful to identify complex systems by requiring that the scenario of Fig. 1.3.2 is violated by changes in the local (i.e., everywhere including the bulk) properties of the system, rather than just the surface.

It may be asked whether the notion of "local properties" is sufficiently well defined as we are using it. In principle,it is not. For now, we adopt this notion from thermodynamics. When only a few properties, like the energy and entropy, are relevant, "affect locally"is a precise concept.Later we would like to replace the use of local thermodynamic properties with a more general concept—the behavior of the system.

How is the scenario of Fig. 1.3.2 violated for a complex system? We can find that the local properties of the small part are affected without affecting the local properties of the large part.Or we can find that the local properties of the large part are affected as well. The distinction between these two ways of affecting the system is important, because it can enable us to distinguish between different kinds of complex systems. It will be helpful to name them for later reference. We call the first category of systems complex materials, the second category we call complex organisms.

Why don't we also include the possibility that the large part is affected but not the small part? At this point it makes sense to consider generic subdivision rather than special subdivision. By generic subdivision, we mean the ensemble of possible subdivisions rather than a particular one.Once we are considering complex systems,the effect of removal of part of a system may depend on which part is removed. However, when we are trying to understand whether or not we have a complex system, we can limit ourselves to considering the generic effects of removing a part of the system. For this reason we do not consider the possibility that subdivision affects the large system and not the small. This might be possible for the removal of a particular small part, but it would be surprising to discover a system where this is generically true.

Two examples may help to illustrate the different classes of complex systems. At least superficially, plants are complex materials, while animals are complex organisms. The reason that plants are complex materials is that the cutting of parts of a plant, such as leaves, a branch, or a root, typically does not affect the local properties of the rest of the plant, but does affect the excised part. For animals this is not generically the case. However, it would be better to argue that plants are in an intermediate category, where some divisions, such as cutting out a lateral section of a tree trunk, affect both small and large parts, while others affect only the smaller part. For animals, essentially all divisions affect both small and large parts. We believe that complex organisms play a special role in the study of complex system behavior. The essential quality of a complex organism is that its properties are tied to the existence of all of its parts.

How large is the small part we are talking about? Loss of a few cells from the skin of an animal will not generally affect it. As the size of the removed portion is decreased,it may be expected that the influence on the local properties of the larger system will be reduced. This leads to the concept of a robust complex system. Qualitatively, the larger the part that can be removed from a complex system without affecting its local properties,the more robust the system is. We see that a complex material is the limiting case of a highly robust complex system.

The flip side of subdivision of a system is aggregation. For thermodynamic systems, subdivision and aggregation are the same, but for complex systems they are quite different. One of the questions that will concern us is what happens when we place a few or many complex systems together. Generally we expect that the individual complex systems will interact with each other. However, one of the points we can make at this time is that just placing together many complex systems, trees or people, does not make a larger complex system by the criteria of subdivision. Thus, a collection of complex systems may result in a system that behaves as a thermodynamic system under subdivision—separating it into parts does not affect the behavior of the parts.

The topic of bringing together many pieces or subdividing into many parts is also quite distinct from the topic of subdivision by removal of a single part. This brings us to a second assumption we will discuss. Thermodynamic systems are assumed to be composed of a very large number of particles. What about complex systems? We know that the number of molecules in a cup of water is not greater than the number of molecules in a human being. And yet, we understand that this is not quite the right point. We should not be counting the number of water molecules in the person,instead we might count the number of cells, which is much smaller. Thus appears the problem of counting the number of components of a system. In the context of correlations in materials, this was briefly discussed at the end of the last section. Let us assume for the moment that we know how to count the number of components. It seems clear that systems with only a few components should not be treated by thermodynamics. One of the interesting questions we will discuss is whether in the limit of a very large number of components we will always have a thermodynamic system.Stated in a simpler way from the point of view of the study of complex systems, the question becomes how large is too large or how many is too many. From the thermodynamic perspective the question is, Under what circumstances do we end up with the thermodynamic limit?

We now switch to a discussion of time-related assumptions.One of the basic assumptions of thermodynamics is the ergodic theorem that enables the description of a single system using an ensemble. When the ergodic theorem breaks down, as discussed in the previous section, additional fixed or quenched variables become important. This is the same as saying that there are significant differences between different examples of the macroscopic system we are interested in. This is a necessary condition for the existence of a complex system. The alternative would be that all realizations of the system would be the same, which does not coincide with intuitive notions of complexity. We will discuss several examples of the breaking of the ergodic theorem later. The simplest example is a magnet. The orientation of the magnet is an additional parameter that must be specified, and therefore the ergodic theorem is violated for this system. Any system that breaks symmetry violates the ergodic theorem. However, we do not accept a magnet as a complex system. Therefore we can assume that the breaking of ergodicity is a necessary but not sufficient condition for complexity. All of the systems we will discuss break ergodicity, and therefore it is always necessary to specify which coordinates of the complex system are fixed and which are to be assumed to be so rapidly varying that they can be assigned equilibrium Boltzmann probabilities.

A special case of the breaking of the ergodic theorem, but one that strikes even more deeply at the assumptions of thermodynamics, is a violation of the separation of time scales. If there are dynamical processes that occur on every time scale, then it becomes impossible to treat the system using the conventional separation of scales into fast,slow and dynamic processes. As we will discuss in Section 1.10,the techniques of renormalization that are used in phase transitions to deal with the existence of many spatial scales may also be used to describe systems changing on many time scales.

Finally, inherent in thermodynamics,the concept of equilibrium and the ergodic theorem is the assumption that the initial condition of the system does not matter. For a complex system,the initial condition of the system does matter over the time scales relevant to our observation. This brings us back to the concept of correlation time. The correlation time describes the length of time over which the initial conditions are relevant to the dynamics. This means that our observation of a complex system must be shorter than a correlation time. The spatial analog, the correlation length, describes the effects of surfaces on the system. The discussion of the effects of subdivision also implies that the system must be smaller than a correlation length. This means that complex systems change their internal structure—adapt—to conditions at their boundaries. Thus, a suggestive though incomplete summary of our discussion of complexity in the context of thermodynamics is that a complex system is contained within a single correlation distance and correlation time.

1.4 Activated Processes (and Glasses)

In the last section we saw figures (Fig. 1.3.7) showing the free energy as a function of a macroscopic parameter with two minima. In this section we analyze a single particle system that has a potential energy with a similar shape (Fig. 1.4.1). The particle is in equilibrium with a thermal reservoir. If the average energy is lower than the energy of the barrier between the two wells, then the particle generally resides for a time in one well and then switches to the other. At very low temperatures, in equilibrium,it will be more and more likely to be in the lower well and less likely to be in the higher well. We use this model to think about a system with two possible states, where one state is higher in energy than the other. If we start the system in the higher energy state, the system will relax to the lower energy state. Because the process of relaxation is enabled or accelerated by energy from the thermal resevoir, we say that it is activated.

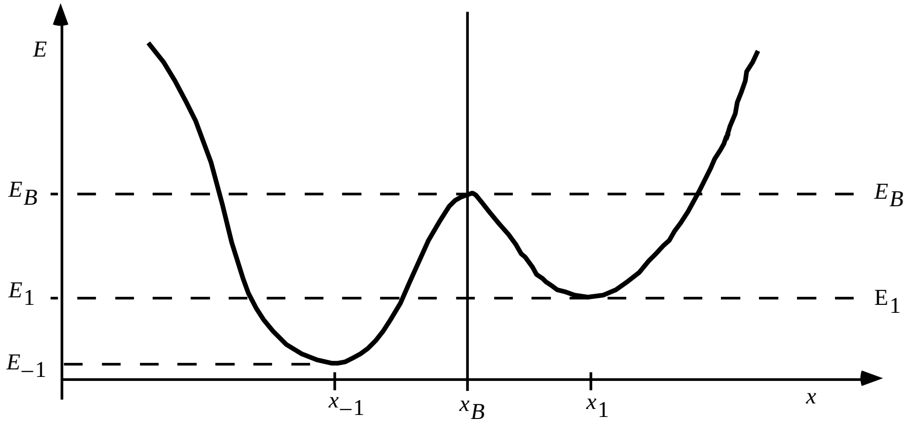


Figure 1.4.1 Illustration of the potential energy of a system that has two local minimum energy configurations x_1 and x_{-1}. When the temperature is lower than the energy barriers $E_B - E_{-1}$ and $E_B - E_1$, the system may be considered as a two-state system with transitions between them. The relative probability of the two states varies with temperature and the relative energy of the bottom of the two wells. The rate of transition also varies with temperature. When the system is cooled systematically the two-state system is a simple model of a glass (Fig. 1.4.2). At low temperatures the system can not move from one well to the other, but is in equilibrium within a single well. ▮

1.4.1 *Two-state systems*

It might seem that a system with only two different states would be easy to analyze. Eventually we will reach a simple problem. However, building the simple model will require us to identify some questions and approximations relevant to our understanding of the application of this model to physical systems (e.g. the problem of protein folding found in Chapter 4). Rather than jumping to the simple two-state problem (Eq. (1.4.40) below), we begin from a particle in a double-well potential. The kinetics and thermodynamics in this system give some additional content to the thermodynamic discussion of the previous section and introduce new concepts.

We consider Fig. 1.4.1 as describing the potential energy $V(x)$ experienced by a classical particle in one dimension. The region to the right of x_B is called the right well and to the left is called the left well.A classical trajectory of the particle with conserved energy would consist of the particle bouncing back and forth within the potential well between two points that are the solution of the equation $V(x) = E$, where E is the total energy of the particle. The kinetic energy at any time is given by

$$E(x,p)-V(x)=\frac{1}{2}mv^2 \tag{1.4.1}$$

which determines the magnitude of the velocity at any position but not the direction. The velocity switches direction every bounce.When the energy is larger than E_B, there is only one distinct trajectory at each energy. For energies larger than E_1 but smaller than E_B, there are two possible trajectories, one in the right well—to the right of x_B — and one in the left well. Below E_1, which is the minimum energy of the right well,there is again only one trajectory possible, in the left well. Below E_{-1} there are no possible locations for the particle.

If we consider this system in isolation,there is no possibility that the particle will change from one trajectory to another. Our first objective is to enable the particle to be in contact with some other system (or coordinate) with which it can transfer energy and momentum. For example, we could imagine that the particle is one of many moving in the double well—like the ideal gas. Sometimes there are collisions that change the energy and direction of the motion. The same effect would be found for many other ways we could imagine the particle interacting with other systems. The main approximation, however, is that the interaction of the particle with the rest of the universe occurs only over short times. Most of the time it acts as if it were by itself in the potential well. The particle follows a trajectory and has an energy that is the sum of its kinetic and potential energies (Eq.(1.4.1)). There is no need to describe the energy associated with the interaction with the other systems. All of the other particles of the gas (or whatever picture we imagine) form the thermal reservoir, which has a well-defined temperature T.

We can increase the rate of collisions between the system and the reservoir without changing our description. Then the particle does not go very far before it forgets the direction it was traveling in and the energy that it had. But as long as the collisions themselves occur over a short time compared to the time between collisions,any time we look at the particle, it has a well-defined energy and momentum. From moment to moment,the kinetic energy and momentum changes unpredictably. Still,the position of the particle must change continuously in time. This scenario is known as diffusive motion. The different times are related by:

collision (interaction) time << time between collisions << transit time

where the transit time is the time between bounces from the walls of the potential well if there were no collisions—the period of oscillation of a particle in the well. The particle undergoes a kind of random walk, with its direction and velocity changing randomly from moment to moment. We will assume this scenario in our treatment of this system.

When the particle is in contact with a thermal reservoir, the laws of thermodynamics apply. The Boltzmann probability gives the probability that the particle is found at position x with momentum p:

$$P(x,p)=e^{-E(x,p)/kT}/Z$$
$$Z=\sum_{x,p} e^{-E(x,p)/kT}=\frac{1}{h}\int dxdp\, e^{-E(x,p)/kT} \qquad (1.4.2)$$

Formally, this expression describes a large number of independent systems that make up a canonical ensemble.The ensemble of systems provides a formally precise way of describing probabilities as the number of systems in the ensemble with a particular value of the position and momentum. As in the previous section, Z guarantees that the sum over all probabilities is 1. The factor of h is not relevant in what follows, but for completeness we keep it and associate it with the momentum integral, so that $\Sigma_p \rightarrow \int dp/h$.

If we are interested in the position of the particle,and are not interested in its momentum, we can simplify this expression by integrating over all values of the momentum. Since the energy separates into kinetic and potential energy:

$$P(x)=\frac{e^{-V(x)/kT}\int (dp/h)e^{-p^2/2mkT}}{\int dx\, e^{-V(x)/kT}\int (dp/h)\,e^{-p^2/2mkT}}=\frac{e^{-V(x)/kT}}{\int dx\, e^{-V(x)/kT}} \qquad (1.4.3)$$

The resulting expression looks similar to our original expression. Its meaning is somewhat different,however, because $V(x)$ is only the potential energy of the system. Since the kinetic energy contributes equivalently to the probability at every location, $V(x)$ determines the probability at every x. An expression of the form $e^{-E/kT}$ is known as the Boltzmann factor of E. Thus Eq.(1.4.3) says that the probability $P(x)$ is proportional to the Boltzmann factor of $V(x)$. We will use this same trick to describe the probability of being to the right or being to the left of x_B in terms of the minimum energy of each well.

To simplify to a two-state system, we must define a variable that specifies only which of the two wells the particle is in. So we label the system by $s = \pm 1$, where $s = +1$ if $x > x_B$ and $s = -1$ if $x < x_B$ for a particular realization of the system at a particular time, or:

$$s = \mathrm{sign}(x - x_B) \tag{1.4.4}$$

Probabilistically, the case $x = x_B$ never happens and therefore does not have to be accounted for.

We can calculate the probability $P(s)$ of the system having a value of $s = +1$ using:

$$P(1) = \frac{\int_{x_B}^{\infty} dx\, e^{-V(x)/kT}}{\int dx\, e^{-V(x)/kT}} \tag{1.4.5}$$

The largest contribution to this probability occurs when $V(x)$ is smallest. We assume that kT is small compared to E_B, then the value of the integral is dominated by the region immediately in the vicinity of the minimum energy. Describing this as a two-state system is only meaningful when this is true. We simplify the integral by expanding it in the vicinity of the minimum energy and keeping only the quadratic term:

$$V(x) = E_1 + \tfrac{1}{2} m\omega_1^2 (x - x_1)^2 + \quad = E_1 + \tfrac{1}{2} k_1 (x - x_1)^2 + \tag{1.4.6}$$

where

$$k_1 = m\omega_1^2 = \left.\frac{d^2 V(x)}{dx^2}\right|_{x_1} \tag{1.4.7}$$

is the effective spring constant and ω_1 is the frequency of small oscillations in the right well. We can now write Eq. (1.4.5) in the form

$$P(1) = \frac{e^{-E_1/kT} \int_{x_B}^{\infty} dx\, e^{-k_1 (x - x_1)^2/2kT}}{\int dx\, e^{-V(x)/kT}} \tag{1.4.8}$$

Because the integrand in the numerator falls rapidly away from the point $x = x_1$, we could extend the lower limit to $-\infty$. Similarly, the probability of being in the left well is:

$$P(-1) = \frac{e^{-E_{-1}/kT} \int_{-\infty}^{x_B} dx\, e^{-k_{-1} (x - x_{-1})^2/2kT}}{\int dx\, e^{-V(x)/kT}} \tag{1.4.9}$$

Here the upper limit of the integral could be extended to ∞. It is simplest to assume that $k_1 = k_{-1}$. This assumption, that the shape of the wells are the same, does not significantly affect most of the discussion (Question 1.4.1–1.4.2). The two probabilities are proportional to a new constant times the Boltzmann factor $e^{-E/kT}$ of the energy at the bottom of the well. This can be seen even without performing the integrals in Eq. (1.4.8) and Eq. (1.4.9). We redefine Z for the two-state representation:

$$P(-1)=\frac{e^{-E_{-1}/kT}}{Z_s} \tag{1.4.10}$$

$$P(1)=\frac{e^{-E_1/kT}}{Z_s} \tag{1.4.11}$$

The new normalization Z_s can be obtained from:

$$P(1)+P(-1)=1 \tag{1.4.12}$$

giving

$$Z_s=e^{-E_1/kT}+e^{-E_{-1}/kT} \tag{1.4.13}$$

which is different from the value in Eq. (1.4.2). We arrive at the desired two-state result:

$$P(1)=\frac{e^{-E_1/kT}}{e^{-E_1/kT}+e^{-E_{-1}/kT}}=\frac{1}{1+e^{(E_1-E_{-1})/kT}}=f(E_1-E_{-1}) \tag{1.4.14}$$

where *f* is the Fermi probability or Fermi function:

$$f(x)=\frac{1}{1+e^{x/kT}} \tag{1.4.15}$$

For readers who were introduced to the Fermi function in quantum statistics,it is not unique to that field, it occurs anytime there are exactly two different possibilities. Similarly,

$$P(-1)=\frac{e^{-E_{-1}/kT}}{e^{-E_1/kT}+e^{-E_{-1}/kT}}=\frac{1}{1+e^{(E_{-1}-E_1)/kT}}=f(E_{-1}-E_1) \tag{1.4.16}$$

which is consistent with Eq. (1.4.12) above since

$$f(x)+f(-x)=1 \tag{1.4.17}$$

Question 1.4.1 Discuss how $k_1 \neq k_{-1}$ would affect the results for the two-state system in equilibrium. Obtain expressions for the probabilities in each of the wells.

Solution 1.4.1 Extending the integrals to ±∞, as described in the text after Eq. (1.4.8) and Eq. (1.4.9), we obtain:

$$P(1)=\frac{e^{-E_1/kT}\sqrt{2\pi kT/k_1}}{\int dx\, e^{-V(x)/kT}} \tag{1.4.18}$$

$$P(-1)=\frac{e^{-E_1/kT}\sqrt{2\pi kT/k_{-1}}}{\int dx\, e^{-V(x)/kT}} \tag{1.4.19}$$

Because of the approximate extension of the integrals, we are no longer guaranteed that the sum of these probabilities is 1. However, within the accuracy of the approximation, we can reimpose the normalization condition. Before we do so, we choose to rewrite $k_1 = m\omega_1{}^2 = m(2\pi\nu_1)^2$, where ν_1 is the natural frequency of the well. We then ignore all common factors in the two probabilities and write

$$P(1) = \frac{\nu_1^{-1} e^{-E_1/kT}}{Z'_s} \tag{1.4.20}$$

$$P(-1) = \frac{\nu_{-1}^{-1} e^{-E_{-1}/kT}}{Z'_s} \tag{1.4.21}$$

$$Z'_s = \nu_{-1}^{-1} e^{-E_1/kT} + \nu_{-1}^{-1} e^{-E_{-1}/kT} \tag{1.4.22}$$

Or we can write, as in Eq. (1.4.14)

$$P(1) = \frac{1}{1 + (\nu_1/\nu_{-1}) e^{(E_1 - E_{-1})/kT}} \tag{1.4.23}$$

and similarly for $P(-1)$. ∎

Question 1.4.2 Redefine the energies E_1 and E_{-1} to include the effect of the difference between k_1 and k_{-1} so that the probability $P(1)$ (Eq. (1.4.23)) can be written like Eq. (1.4.14) with the new energies. How is the result related to the concept of free energy and entropy?

Solution 1.4.2 We define the new energy of the right well as

$$F_1 = E_1 + kT\ln(\nu_1) \tag{1.4.24}$$

This definition can be seen to recover Eq. (1.4.23) from the form of Eq. (1.4.14) as

$$P(1) = f(F_1 - F_{-1}) \tag{1.4.25}$$

Eq. (1.4.24) is very reminiscent of the definition of the free energy Eq. (1.3.33) if we use the expression for the entropy:

$$S_1 = -k\ln(\nu_1) \tag{1.4.26}$$

Note that if we consider the temperature dependence, Eq. (1.4.25) is not identical in its behavior with Eq.(1.4.14). The free energy, F_1, depends on T, while the energy at the bottom of the well, E_1, does not. ∎

In Question 1.4.2, Eq. (1.4.24), we have defined what might be interpreted as a free energy of the right well. In Section 1.3 we defined only the free energy of the system as a whole. The new free energy is for part of the ensemble rather than the whole ensemble. We can do this quite generally. Start by identifying a certain subset of all possible states of a system. For example, $s = 1$ in Eq. (1.4.4). Then we define the free energy using the expression:

$$F_s(1) = -kT\ln\Big(\sum_{\{x,p\}}\delta_{s,1}\,e^{-E(\{x,p\})/kT}\Big) = -kT\ln(Z_1) \qquad (1.4.27)$$

This is similar to the usual expression for the free energy in terms of the partition function Z, but the sum is only over the subset of states. When there is no ambiguity, we often drop the subscript and write this as $F(1)$. From this definition we see that the probability of being in the subset of states is proportional to the Boltzmann factor of the free energy

$$P(1) \quad e^{-F_s(1)/kT} \qquad (1.4.28)$$

If we have several different subsets that account for all possibilities, then we can normalize Eq. (1.4.28) to find the probability itself. If we do this for the left and right wells, we immediately arrive at the expression for the probabilities in Eq.(1.4.14) and Eq. (1.4.16), with E_1 and E_{-1} replaced by $F_s(1)$ and $F_s(-1)$ respectively. From Eq.(1.4.28) we see that for a collection of states,the free energy plays the same role as the energy in the Boltzmann probability.

We note that Eq. (1.4.24) is not the same as Eq.(1.4.27). However, as long as the relative energy is the same, $F_1 - F_{-1} = F_s(1) - F_s(-1)$,the normalized probability is unchanged. When $k_1 = k_{-1}$, the entropic part of the free energy is the same for both wells. Then direct use of the energy instead of the free energy is valid,as in Eq.(1.4.14). We can evaluate the free energy of Eq. (1.4.27), including the momentum integral:

$$\begin{aligned}
Z_1 &= \int_{x_B}^{\infty} dx \int (dp/h)\, e^{-E(x,p)/kT} = \int_{x_B}^{\infty} dx e^{-V(x)/kT} \int (dp/h) e^{-p^2/2mkT} \\
&\approx e^{-E_1/kT} \int_{x_B}^{\infty} dx\, e^{-k_1(x-x_1)^2/2kT} \sqrt{2\pi mkT}/h \approx e^{-E_1/kT}\sqrt{m/k_1}\,2\pi kT/h \qquad (1.4.29) \\
&= e^{-E_1/kT} kT/h\nu_1
\end{aligned}$$

$$F_s(1) = E_1 + kT\ln(h\nu_1/kT) \qquad (1.4.30)$$

where we have used the definition of the well oscillation frequency above Eq.(1.4.20) to simplify the expression.A similar expression holds for Z_{-1}. The result would be exact for a pure harmonic well.

The new definition of the free energy of a set of states can also be used to understand the treatment of macroscopic systems,specifically to explain why the energy is determined by minimizing the free energy. Partition the possible microstates by the value of the energy, as in Eq. (1.3.35). Define the free energy as a function of the energy analogous to Eq. (1.4.27)

$$F(U) = -kT\ln\left(\sum_{\{x,p\}}\delta_{E(\{x,p\}),U}\,e^{-E(\{x,p\})/kT}\right) \qquad (1.4.31)$$

Since the relative probability of each value of the energy is given by

$$P(U) \quad e^{-F(U)/kT} \tag{1.4.32}$$

the most likely energy is given by the lowest free energy. For a macroscopic system, the most likely value is so much more likely than any other value that it is observed in any measurement. This can immediately be generalized. The minimization of the free energy gives not only the value of the energy but the value of any macroscopic parameter.

1.4.2 *Relaxation of a two-state system*

To investigate the kinetics of the two-state system, we assume an ensemble of systems that is not an equilibrium ensemble. Instead,the ensemble is characterized by a time-dependent probability of occupying the two wells:

$$\begin{aligned} P(1) &\rightarrow P(1;t) \\ P(-1) &\rightarrow P(-1;t) \end{aligned} \tag{1.4.33}$$

Normalization continues to hold at every time:

$$P(1;t)+P(-1;t)=1 \tag{1.4.34}$$

For example, we might consider starting a system in the upper well and see how the system evolves in time. Or we might consider starting a system in the lower well and see how the system evolves in time. We answer the question using the time-evolving probabilities that describe an ensemble of systems with the same starting condition. To achieve this objective, we construct a differential equation describing the rate of change of the probability of being in a particular well in terms of the rate at which systems move from one well to the other. This is just the Master equation approach from Section 1.2.4.

The systems that make transitions from the left to the right well are the ones that cross the point $x = x_B$. More precisely, the rate at which transitions occur is the probability current per unit time of systems at x_B, moving toward the right. Similar to Eq. (1.3.47) used to obtain the pressure of an ideal gas on a wall,the number of particles crossing x_B is the probability of systems at x_B with velocity v, times their velocity:

$$J(1|-1)=\int_0^\infty (dp/h)vP(x_B,p;t) \tag{1.4.35}$$

where $J(1|-1)$ is the number of systems per unit time moving from the left to the right. There is a hidden assumption in Eq. (1.4.35). We have adopted a notation that treats all systems on the left together. When we are considering transitions,this is only valid if a system that crosses $x = x_B$ from right to left makes it down into the well on the left, and thus does not immediately cross back over to the side it came from.

We further assume that in each well the systems are in equilibrium, even when the two wells are not in equilibrium with each other. This means that the probability of being in a particular location in the right well is given by:

$$\begin{aligned} P(x,p;t) &= P(1;t)e^{-E(x,p)/kT}/Z_1 \\ Z_1 &= \int_{x_B}^{\infty} dxdp\; e^{-E(x,p)/kT} \end{aligned} \qquad (1.4.36)$$

In equilibrium,this statement is true because then $P(1) = Z_1/Z$. Eq.(1.4.36) presumes that the rate of collisions between the particle and the thermal reservoir is faster than both the rate at which the system goes from one well to the other and the frequency of oscillation in a well.

In order to evaluate the transition rate Eq.(1.4.35), we need the probability at x_B. We assume that the systems that cross x_B moving from the left well to the right well (i.e.,moving to the right) are in equilibrium with systems in the left well from where they came. Systems that are moving from the right well to the left have the equilibrium distribution characteristic of the right well. With these assumptions, the rate at which systems hop from the left to the right is given by:

$$\begin{aligned} J(1|-1) &= \int_0^{\infty} (dp/h)(p/m)\left(P(-1;t)e^{-(E_B+p^2/2m)/kT}/Z_{-1}\right) \\ &= P(-1;t)e^{-E_B/kT}(kT/h)/Z_{-1} \end{aligned} \qquad (1.4.37)$$

We find using Eq. (1.4.29) that the current of systems can be written in terms of a transition rate per system:

$$\begin{aligned} J(1|-1) &= R(1|-1)P(-1;t) \\ R(1|-1) &= \nu_{-1}e^{-(E_B-E_{-1})/kT} \end{aligned} \qquad (1.4.38)$$

Similarly, the current and rate at which systems hop from the right to the left are given by:

$$\begin{aligned} J(-1|1) &= R(-1|1)P(1;t) \\ R(-1|1) &= \nu_1 e^{-(E_B-E_1)/kT} \end{aligned} \qquad (1.4.39)$$

When $k_1 = k_{-1}$ then $\nu_1 = \nu_{-1}$. We continue to deal with this case for simplicity and define $\nu = \nu_1 = \nu_{-1}$. The expressions for the rate of transition suggest the interpretation that the frequency ν is the rate of attempt to cross the barrier. The probability of crossing in each attempt is given by the Boltzmann factor, which gives the likelihood that the energy exceeds the barrier. While this interpretation is appealing, and is often given,it is misleading. It is better to consider the frequency as describing the width of the well in which the particle wanders. The wider the well is,the less likely is a barrier crossing. This interpretation survives better when more general cases are considered.

The transition rates enable us to construct the time variation of the probability of occupying each of the wells. This gives us the coupled equations for the two probabilities:

$$\begin{aligned} \dot{P}(1;t) &= R(1|-1)P(-1;t) - R(-1|1)P(1;t) \\ \dot{P}(-1;t) &= R(-1|1)P(1;t) - R(1|-1)P(-1;t) \end{aligned} \qquad (1.4.40)$$

These are the Master equations (Eq. (1.2.86)) for the two-state system. We have arrived at these equations by introducing a set of assumptions for treating the kinetics of a single particle. The equations are much more general, since they say only that there is a rate of transition between one state of the system and the other. It is the correspondence between the two-state system and the moving particle that we have established in Eqs. (1.4.38) and (1.4.39). This correspondence is approximate. Eq. (1.4.40) does not rely upon the relationship between E_B and the rate at which systems move from one well to the other. However, it does rely upon the assumption that we need to know only which well the system is in to specify its rate of transition to the other well. On average this is always true, but it would not be a good description of the system, for example, if energy is conserved and the key question determining the kinetics is whether the particle has more or less energy than the barrier E_B.

We can solve the coupled equations in Eq. (1.4.40) directly. Both equations are not necessary, given the normalization constraint Eq.(1.4.34). Substituting $P(-1;t) = 1 - P(1;t)$ we have the equation

$$\dot{P}(1;t) = R(-1|1) - P(1;t)(R(1|-1) + R(-1|1)) \tag{1.4.41}$$

We can rewrite this in terms of the equilibrium value of the probability. By definition this is the value at which the time derivative vanishes.

$$P(1;\infty) = R(-1|1)/(R(1|-1) + R(-1|1)) = f(E_1 - E_{-1}) \tag{1.4.42}$$

where the right-hand side follows from Eq.(1.4.38) and Eq.(1.4.39) and is consistent with Eq. (1.4.13), as it must be. Using this expression, Eq. (1.4.24) becomes

$$\dot{P}(1;t) = (P(1;\infty) - P(1;t))/\tau \tag{1.4.43}$$

where we have defined an additional quantity

$$1/\tau = (R(1|-1) + R(-1|1)) = \nu(e^{-(E_B - E_1)/kT} + e^{-(E_B - E_{-1})/kT}) \tag{1.4.44}$$

The solution of Eq. (1.4.43) is

$$P(1;t) = (P(1;0) - P(1;\infty))e^{-t/\tau} + P(1;\infty) \tag{1.4.45}$$

This solution describes a decaying exponential that changes the probability from the starting value to the equilibrium value. This explains the definition of τ, called the relaxation time. Since it is inversely related to the sum of the rates of transition between the wells,it is a typical time taken by a system to hop between the wells. The relaxation time does not depend on the starting probability. We note that the solution of Eq.(1.4.41) does not depend on the explicit form of $P(1;\infty)$ or τ. The definitions implied by the first equal signs in Eq.(1.4.42) and Eq.(1.4.44) are sufficient. Also, as can be quickly checked, we can replace the index 1 with the index −1 without changing anything else in Eq (1.4.45). The other equations are valid (by symmetry) after the substitution $1 \leftrightarrow -1$.

There are several intuitive relationships between the equilibrium probabilities and the transition rates that may be written down. The first is that the ratio of the equilibrium probabilities is the ratio of the transition rates:

$$P_1(\infty)\big/P_{-1}(\infty) = R(-1|1)/R(1|-1) \qquad (1.4.46)$$

The second is that the equilibrium probability divided by the relaxation time is the rate of transition:

$$P_1(\infty)\big/\tau = R(-1|1) \qquad (1.4.47)$$

Question 1.4.3 Eq. (1.4.45) implies that the relaxation time of the system depends largely on the smaller of the two energy barriers $E_B - E_1$ and $E_B - E_{-1}$. For Fig. 1.4.1 the smaller barrier is $E_B - E_1$. Since the relaxation time is independent of the starting probability, this barrier controls the rate of relaxation whether we start the system from the lower well or the upper well. Why does the barrier $E_B - E_1$ control the relaxation rate when we start from the lower well?

Solution 1.4.3 Even though the rate of transition from the lower well to the upper well is controlled by $E_B - E_{-1}$, the fraction of the ensemble that must make the transition in order to reach equilibrium depends on E_1. The higher it is,the fewer systems must make the transition from $s = -1$ to $s = 1$. Taking this into consideration implies that the time to reach equilibrium depends on $E_B - E_1$ rather than $E_B - E_{-1}$. ▮

1.4.3 *Glass transition*

Glasses are materials that when cooled from the liquid do not undergo a conventional transition to a solid. Instead their viscosity increases,and in the vicinity of a particular temperature it becomes so large that on a reasonable time scale they can be treated as solids. However, on long enough time scales,they flow as liquids. We will model the glass transition using a two-state system by considering what happens as we cool down the two-state system. At high enough temperatures, the system hops back and forth between the two minima with rates given by Eqs.(1.4.38) and (1.4.39). ν is a microscopic quantity; it might be a vibration rate in the material. Even if the barriers are higher than the temperature, $E_B - E_{\pm 1} >> kT$, the system will still be able to hop back and forth quite rapidly from a macroscopic perspective.

As the system is cooled down, the hopping back and forth slows down. At some point the rate of hopping will become longer than the time we are observing the system. Systems in the higher well will stay there. Systems in the lower well will stay there. This means that the population in each well becomes fixed. Even when we continue to cool the system down, there will be no change, and the ensemble will no longer be in equilibrium. Within each well the system will continue to have a probability distribution for its energy given by the Boltzmann probability, but the relative populations of the two wells will no longer be described by the equilibrium Boltzmann probability.

To gain a feeling for the numbers,a typical atomic vibration rate is 10^{12}/sec. For a barrier of 1eV, at twice room temperature, $kT \approx 0.05$eV (600°K), the transition rate would be of order 10^{3}/sec. This is quite slow from a microscopic perspective, but at room temperature it would be only 10^{-6}/sec, or one transition per year.

The rate at which we cool the system down plays an essential role. If we cool faster, then the temperature at which transitions stop is higher. If we cool at a slower rate, then the temperature at which the transitions stop is lower. This is found to be the case for glass transitions, where the cooling rate determines the departure point from the equilibrium trajectory of the system,and the eventual properties of the glass are also determined by the cooling rate. Rapid cooling is called quenching. If we raise the temperature and lower it slowly, the procedure is called annealing.

Using the model two-state system we can simulate what would happen if we perform an experiment of cooling a system that becomes a glass.Fig. 1.4.2 shows the probability of being in the upper well as a function of the temperature as the system is cooled down. The curves depart from the equilibrium curve in the vicinity of a transition temperature we might call a freezing transition, because the kinetics become frozen. The glass transition is not a transition like a first- or second-order transition (Section 1.3.4) because it is a transition of the kinetics rather than of the equilibrium structure of the system. Below the freezing transition, the relative probability of the system being in the upper well is given approximately by the equilibrium probability at the transition.

The freezing transition of the relative population of the upper state and the lower state is only a simple model of the glass transition;however, it is also more widely applicable. The freezing does not depend on cooperative effects of many particles. To find examples, a natural place to look is the dynamics of individual atoms in solids. Potential energies with two wells occur for impurities, defects and even bulk atoms in a solid. Impurities may have two different local configurations that differ in energy and are separated by a barrier. This is a direct analog of our model two-state system. When the temperature is lowered, the relative population of the two configurations becomes frozen. If we raise the temperature, the system can equilibrate again.

It is also possible to artificially cause impurity configurations to have unequal energies.One way is to apply uniaxial stress to a crystal—squeezing it along one axis. If an impurity resides in a bond between two bulk atoms, applying stress will raise the energy of impurities in bonds oriented with the stress axis compared to bonds perpendicular to the stress axis. If we start at a relatively high temperature, apply stress and then cool down the material, we can freeze unequal populations of the impurity. If we have a way of measuring relaxation, then by raising the temperature gradually and observing when the defects begin to equilibrate we can discover the barrier to relaxation. This is one of the few methods available to study the kinetics of impurity reorientation in solids.

The two-state system provides us with an example of how a simple system may not be able to equilibrate over experimental time scales. It also shows how an equi-

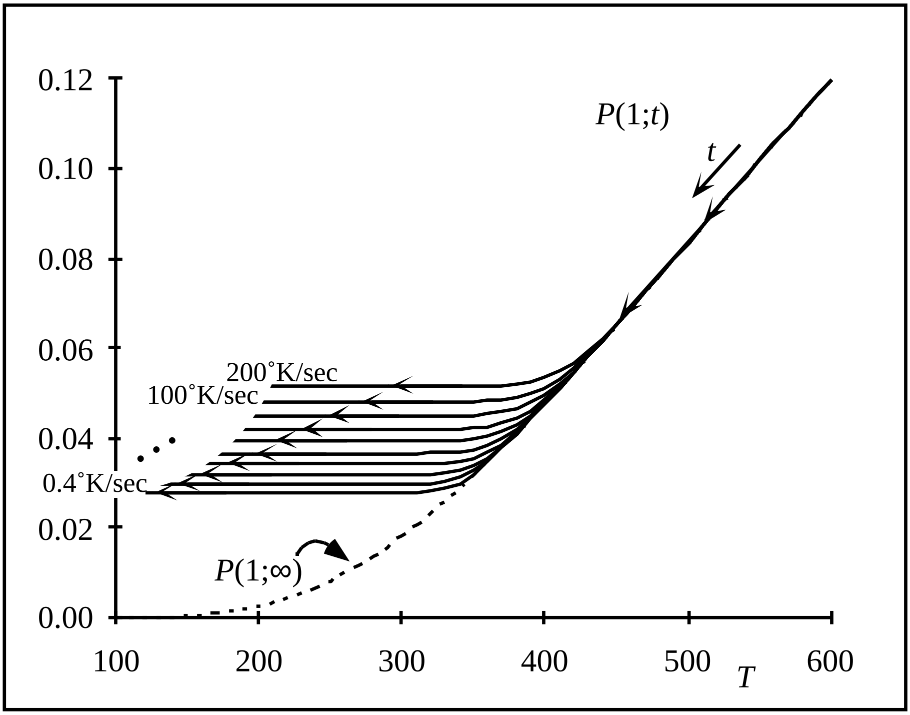


Figure 1.4.2 Plot of the fraction of the systems in the higher energy well as a function of temperature. The equilibrium value is shown with the dashed line. The solid lines show what happens when the system is cooled from a high temperature at a particular cooling rate. The example given uses $E_1 - E_{-1} = 0.1\text{eV}$ and $E_B - E_{-1} = 1.0\text{eV}$. Both wells have oscillation frequencies of $\nu = 10^{12}/\text{sec}$. The fastest cooling rate is 200°K/sec and each successive curve is cooled at a rate that is half as fast, with the slowest rate being 0.4°K/sec. For every cooling rate the system stops making transitions between the wells at a particular temperature that is analogous to a glass transition in this system. Below this temperature the probability becomes essentially fixed. ▮

librium ensemble can be used to treat relative probabilities within a subset of states. Because the motion within a particular well is fast,the relative probabilities of different positions or momenta within a well may be described using the Boltzmann probability. At the same time, the relative probability of finding a system in each of the two wells depends on the initial conditions and the history of the system—what temperature the system experienced and for how long. At sufficiently low temperatures, this relative probability may be treated as fixed. Systems that are in the higher well may be assumed to stay there. At intermediate temperatures, a treatment of the dynamics of the transition between the two wells can (and must) be included. This manifests a violation of the ergodic theorem due to the divergence of the time scale for equilibration between the two wells. Thus we have identified many of the features that are necessary in describing nonequilibrium systems: divergent time scales, violation of the ergodic theorem, frozen and dynamic coordinates. We have illustrated a method for treating systems where there is a separation of long time scales and short time scales.

Question 1.4.4 Write a program that can generate the time dependence of the two-state system for a specified time history. Reproduce Fig. 1.4.2. For an additional "experiment," try the following quenching and annealing sequence:

a. Starting from a high enough temperature to be in equilibrium, cool the system at a rate of 10°K/sec down to $T = 0$.

b. Heat the system up to temperature T_a and keep it there for one second.

c. Cool the system back down to $T = 0$ at rate of 100°K/sec.

Plot the results as a function of T_a. Describe and explain them in words. ∎

1.4.4 *Diffusion*

In this section we briefly consider a multiwell system. An example is illustrated in Fig. 1.4.3, where the potential well depths and barriers vary from site to site. A simpler case is found in Fig. 1.4.4, where all the well depths and barriers are the same. A concrete example would be an interstitial impurity in an ideal crystal. The impurity lives in a periodic energy that repeats every integral multiple of an elementary length a.

We can apply the same analysis from the previous section to describe what happens to a system that begins from a particular well at $x = 0$. Over time, the system makes transitions left and right at random,in a manner that is reminiscent of a random walk. We will see in a moment that the connection with the random walk is valid but requires some additional discussion.

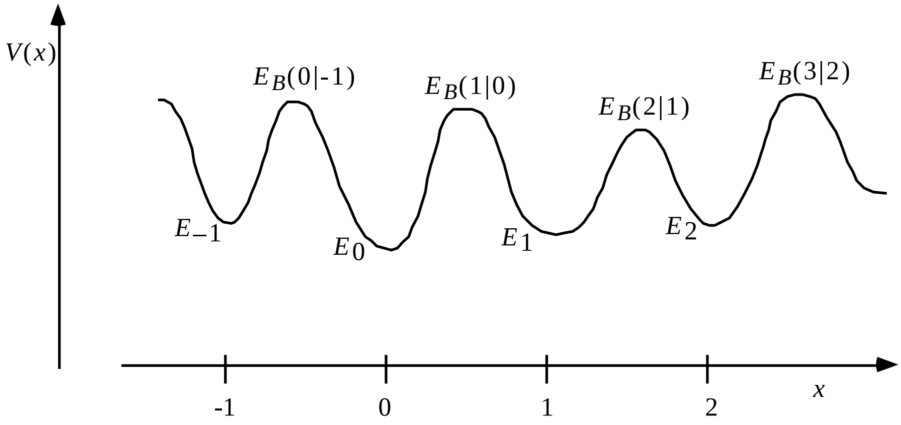


Figure 1.4.3 Illustration of a multiple-well system with barrier heights and well depths that vary from site to site. We focus on the uniform system in Fig. 1.4.4. ∎

The probability of the system being in a particular well is changed by probability currents into the well and out from the well. Systems can move to or from the well immediately to their right and immediately to their left. The Master equation for the *i*th well in Fig. 1.4.3 is:

$$\dot{P}(i;t) = R(i|i-1)P(i-1;t) + R(i|i+1)P(i+1;t) - (R(i+1|i) + R(i-1|i))P(i;t) \quad (1.4.48)$$

$$\begin{aligned} R(i+1|i) &= \nu_i e^{-(E_B(i+1|i)-E_i)/kT} \\ R(i-1|i) &= \nu_i e^{-(E_B(i|i-1)-E_i)/kT} \end{aligned} \quad (1.4.49)$$

where E_i is the depth of the *i*th well and $E_B(i + 1|i)$ is the barrier to its right. For the periodic system of Fig. 1.4.4 ($\nu_i \rightarrow \nu$, $E_B(i + 1|i) \rightarrow E_B$) this simplifies to:

$$\dot{P}(i;t) = R(P(i-1;t) + P(i+1;t) - 2P(i;t)) \quad (1.4.50)$$

$$R = \nu e^{-(E_B-E_0)/kT} \quad (1.4.51)$$

Since we are already describing a continuum differential equation in time,it is convenient to consider long times and write a continuum equation in space as well. Allowing a change in notation we write

$$P(i;t) \rightarrow P(x_i;t) \quad (1.4.52)$$

Introducing the elementary distance between wells *a* we can rewrite Eq. (1.4.50) using:

$$\begin{aligned} &\frac{(P(i-1;t) + P(i+1;t) - 2P(i;t))}{a^2} \\ &\rightarrow \frac{(P(x_i - a;t) + P(x_i + a;t) - 2P(x_i;t))}{a^2} \rightarrow \frac{\partial^2}{\partial x^2} P(x;t) \end{aligned} \quad (1.4.53)$$

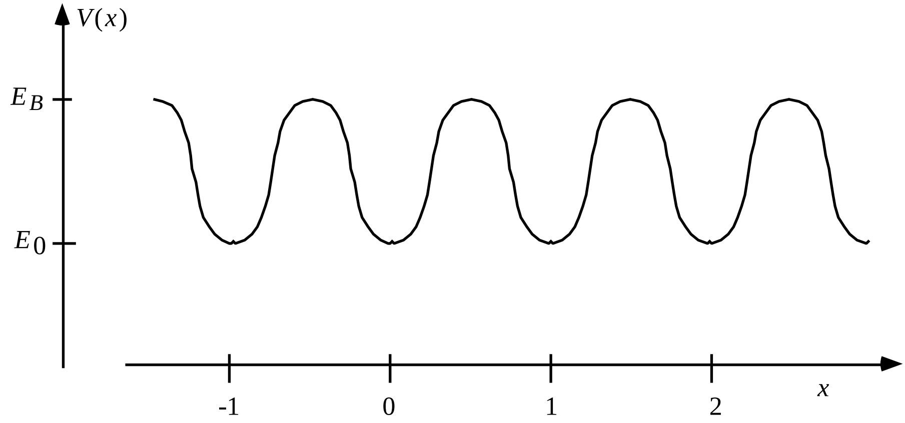


Figure 1.4.4 When the barrier heights and well depths are the same, as illustrated, the long time behavior of this system is described by the diffusion equation. The evolution of the system is controlled by hopping events from one well to the other. The net effect over long times is the same as for the random walk discussed in Section 1.2. ∎

where the last expression assumes a is small on the scale of interest. Thus the continuum version of Eq. (1.4.50) is the conventional diffusion equation:

$$\dot{P}(x;t) = D\frac{\partial^2}{\partial x^2}P(x;t) \tag{1.4.54}$$

The diffusion constant D is given by:

$$D = a^2 R = a^2 \nu e^{-(E_B - E_0)/kT} \tag{1.4.55}$$

The solution of the diffusion equation, Eq. (1.4.54), depends on the initial conditions that are chosen. If we consider an ensemble of a system that starts in one well and spreads out over time, the solution can be checked by substitution to be the Gaussian distribution found for the random walk in Section 1.2:

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} e^{-x^2/4Dt} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2} \qquad \sigma = \sqrt{2Dt} \tag{1.4.56}$$

We see that motion in a set of uniform wells after a long time reduces to that of a random walk.

How does the similarity to the random walk arise? This might appear to be a natural result,since we showed that the Gaussian distribution is quite general using the central limit theorem. The scenario here,h owever, is quite different. The central limit theorem was proven in Section 1.2.2 for the case of a distribution of probabilities of steps taken at specific time intervals. Here we have a time continuum. Hopping events may happen at any time. Consider the case where we start from a particular well. Our differential equation describes a system that might hop to the next well at any time. A hop is an event, and we might concern ourselves with the distribution of such events in time. We have assumed that these events are uncorrelated.There are unphysical consequences of this assumption. For example, no matter how small an interval of time we choose,the particle has some probability of traveling arbitrarily far away. This is not necessarily a correct microscopic picture, but it is the continuum model we have developed.

There is a procedure to convert the event-controlled hopping motion between wells into a random walk that takes steps with a œrtain probability at specific time intervals. We must select a time interval. For this time interval, we evaluate the total probability that hops move a system from its original position to all possible positions of the system. This would give us the function $f(s)$ in Eq.(1.2.34). As long as the mean square displacement is finite,the central limit theorem continues to apply to the probability distribution after a long enough time. The generality of the conclusion also implies that the result is more widely applicable than the assumptions indicate.However, there is a counter example in Question 1.4.5.

Question 1.4.5 Discuss the case of a particle that is not in contact with a thermal resevoir moving in the multiple well system (energy is conserved).

Solution 1.4.5 If the energy of the system is lower than E_B, the system stays in a single well bouncing back and forth. A model that describes how transitions occur between wells would just say there are none.

For the case where the energy is larger than E_B, the system will move with a periodically varying velocity in one direction. There is a problem in selecting an ensemble to describe it. If we choose the ensemble with only one system moving in one direction, then it is described as a deterministic walk. This description is consistent with the motion of the system. However, we might also think to describe the system using an ensemble consisting of particles with the same energy. In this case it would be one particle moving to the right and one moving to the left. Taking an interval of time to be the time needed to move to the next well, we would find a transition probability of 1/2 to move to the right and the same to the left. This would lead to a conventional random walk and will give us an incorrect result for all later times.

This example illustrates the need for an assumption that has not yet been explicitly mentioned. The ensemble must describe systems that can make transitions to each other. Since the energy-conserving systems cannot switch directions, the ensemble cannot include both directions. It is enough, however, for there to be a small nonzero probability for the system to switch directions for the central limit theorem to apply. This means that over long enough times, the distribution will be Gaussian. Over short times,however, the probability distribution from the random walk model and an almost ballistic system would not be very similar. ▮

We can generalize the multiple well picture to describe a biased random walk. The potential we would use is a "washboard potential," illustrated in Fig. 1.4.5. The Master equation is:

$$\dot{P}(i;t) = R_+ P(i-1;t) + R_- P(i+1;t) - (R_+ + R_-)P(i;t) \tag{1.4.57}$$

$$\begin{aligned} R_+ &= \nu_i e^{-\Delta E_+ /kT} \\ R_- &= \nu_i e^{-\Delta E_- /kT} \end{aligned} \tag{1.4.58}$$

To obtain the continuum limit, replace $i \rightarrow x$: $P(i+1;t) \rightarrow P(x+a,t)$, and $P(i-1;t) \rightarrow P(x-a,t)$, and expand in a Taylor series to second order in a to obtain:

$$\dot{P}(x;t) = -v\frac{\partial}{\partial x}P(x;t) + D\frac{\partial^2}{\partial x^2}P(x;t) \tag{1.4.59}$$

$$\begin{aligned} v &= a(R_+ - R_-) \\ D &= a^2(R_+ + R_-)/2 \end{aligned} \tag{1.4.60}$$

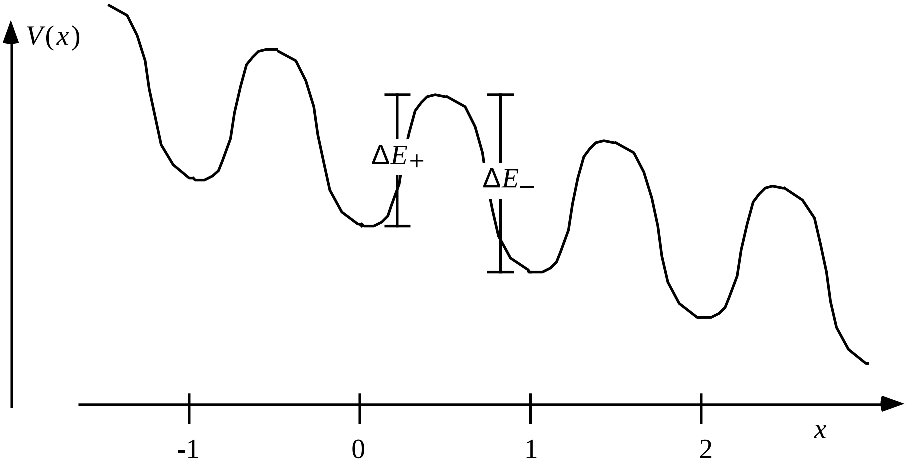


Figure 1.4.5 The biased random walk is also found in a multiple-well system when the illustrated washboard potential is used. The velocity of the system is given by the difference in hopping rates to the right and to the left. ▮

The solution is a moving Gaussian:

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} e^{-(x-vt)^2/4Dt} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-vt)^2/2\sigma^2} \qquad (1.4.61)$$
$$\sigma = \sqrt{2Dt}$$

Since the description of diffusive motion always allows the system to stay where it is, there is a limit to the degree of bias that can occur in the random walk. For this limit set $R_- = 0$. Then $D = av/2$ and the spreading of the probability is given by $\sigma = \sqrt{avt}$. This shows that unlike the biased random walk in Section 1.2, diffusive motion on a washboard with a given spacing a cannot describe ballistic or deterministic motion in a single direction.

1.5 Cellular Automata

The first four sections of this chapter were dedicated to systems in which the existence of many parameters (degrees of freedom) describing the system is hidden in one way or another. In this section we begin to describe systems where many degrees of freedom are explicitly represented. Cellular automata (CA) form a general class of models of dynamical systems which are appealingly simple and yet capture a rich variety of behavior. This has made them a favorite tool for studying the generic behavior of and modeling complex dynamical systems. Historically CA are also intimately related to the development of concepts of computers and computation. This connection continues to be a theme often found in discussions of CA. Moreover, despite the wide differences between CA and conventional computer architectures,CA are convenient for computer simulations in general and parallel computer simulations in particular. Thus CA have gained importance with the increasing use of simulations in the development of our understanding of complex systems and their behavior.

1.5.1 *Deterministic cellular automata*

The concept of cellular automata begins from the concept of space and the locality of influence. We assume that the system we would like to represent is distributed in space,and that nearby regions of space have more to do with each other than regions far apart. The idea that regions nearby have greater influence upon each other is often associated with a limit (such as the speed of light) to how fast information about what is happening in one place can move to another place.*

Once we have a system spread out in space, we mark off the space into cells. We then use a set of variables to describe what is happening at a given instant of time in a particular cell.

$$s(i, j, k;t) = s(x_i, y_j, z_k;t) \tag{1.5.1}$$

where i, j, k are integers ($i, j, k \quad$ Z),and this notation is for a three-dimensional space (3-d). We can also describe automata in one or two dimensions (1-d or 2-d) or higher than three dimensions. The time dependence of the cell variables is given by an iterative rule:

$$s(i, j, k;t) = R(\{s(i' - i, j' - j, k' - k;t - 1)\}_{i',j',k' \quad Z}) \tag{1.5.2}$$

where the rule R is shown as a function of the values of all the variables at the previous time,at positions relative to that of the cell $s(i, j, k;t - 1)$. The rule is assumed to be the same everywhere in the space—there is no space index on the rule. Differences between what is happening at different locations in the space are due only to the values of the variables, not the update rule. The rule is also homogeneous in time; i.e., the rule is the same at different times.

The locality of the rule shows up in the form of the rule. It is assumed to give the value of a particular cell variable at the next time only in terms of the values of cells in the vicinity of the cell at the previous time. The set of these cells is known as its neighborhood. For example, the rule might depend only on the values of twenty-seven cells in a cube centered on the location of the cell itself. The indices of these cells are obtained by independently incrementing or decrementing once, or leaving the same, each of the indices:

$$s(i, j, k;t) = R(s(i \pm 1{,}0, j \pm 1, 0, k \pm 1, 0;t - 1)) \tag{1.5.3}$$

*These assumptions are both reasonable and valid for many systems. However, there are systems where this is not the most natural set of assumptions. For example, when there are widely divergent speeds of propagation of different quantities (e.g.,light and sound) it may be convenient to represent one as instantaneous (light) and the other as propagating (sound). On a fundamental level, Einstein, Podalsky and Rosen carefully formulated the simple assumptions of local influence and found that quantum mechanics violates these simple assumptions.A complete understanding of the nature of their paradox has yet to be reached.

where the informal notation $i \pm 1,0$ is the set $\{i - 1,i,i + 1\}$. In this case there are a total of twenty-seven cells upon which the update rule $R(s)$ depends. The neighborhood could be smaller or larger than this example.

CA can be usefully simplified to the point where each cell is a single binary variable. As usual, the binary variable may use the notation $\{0,1\}$, $\{-1,1\}$, {ON,OFF} or $\{\uparrow,\downarrow\}$. The terminology is often suggested by the system to be described. Two 1-d examples are given in Question 1.5.1 and Fig. 1.5.1. For these 1-d cases we can show the time evolution of a CA in a single figure,where the time axis runs vertically down the page and the horizontal axis is the space axis.Each figure is a CA space-time diagram that illustrates a particular history.

In these examples, a finite space is used rather than an infinite space. We can define various boundary conditions at the edges.The most common is to use a periodic boundary condition where the space wraps around to itself. The one-dimensional examples can be described as circles.A two-dimensional example would be a torus and a three-dimensional example would be a generalized torus. Periodic boundary conditions are convenient, because there is no special position in the space. Some care must be taken in considering the boundary conditions even in this case, because there are rules where the behavior depends on the size of the space. Another standard kind of boundary condition arises from setting all of the values of the variables outside the finite space of interest to a particular value such as 0.

Question 1.5.1 Fill in the evolution of the two rules of Fig. 1.5.1. The first CA (Fig. 1.5.1(a)) is the majority rule that sets a cell to the majority of the three cells consisting of itself and its two neighbors in the previous time. This can be written using $s(i;t) = \pm 1$ as:

$$s(i;t + 1) = \text{sign}(s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad (1.5.4)$$

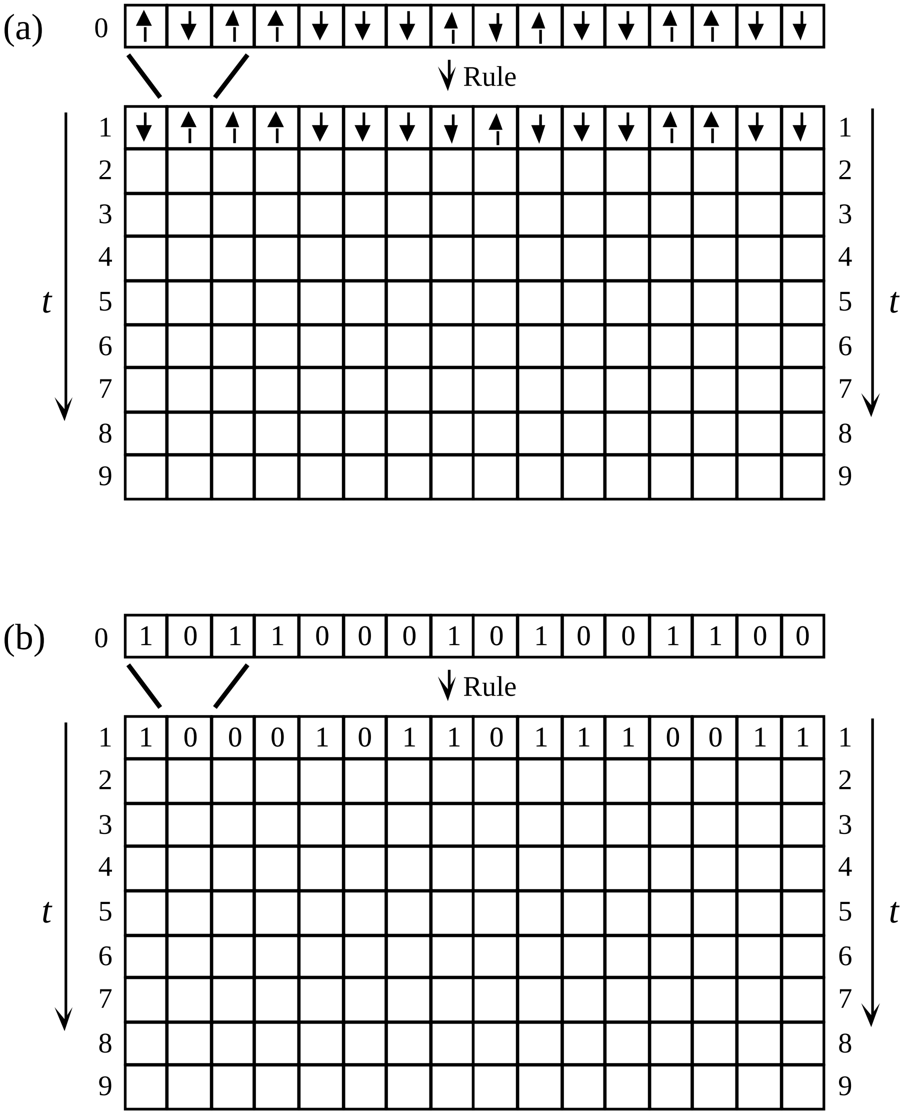
In the figure $\{-1, + 1\}$ are represented by $\{\uparrow, \downarrow\}$ respectively.

The second CA (Fig. 1.5.1(b)), called the mod2 rule,is obtained by setting the ith cell to be OFF if the number of ON squares in the neighborhood is even, and ON if this number is odd. To write this in a simple form use $s(i;t) = \{0, 1\}$. Then:

$$s(i;t + 1) = \text{mod}_2\,(s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad (1.5.5)$$

Solution 1.5.1 Notes:

1. The first rule (a) becomes trivial almost immediately, since it achieves a fixed state after only two updates. Many CA, as well as many physical systems on a macroscopic scale, behave this way.
2. Be careful about the boundary conditions when updating the rules,particularly for rule (b).
3. The second rule (b) goes through a sequence of states very different from each other. Surprisingly, it will recover the initial configuration after eight updates. ∎



0	↑	↓	↑	↑	↓	↓	↓	↑	↓	↑	↓	↓	↑	↑	↓	↓
1	↓	↑	↑	↑	↓	↓	↓	↓	↑	↓	↓	↓	↑	↑	↓	↓
2																
3																
4																
5																
6																
7																
8																
9																

0	1	0	1	1	0	0	0	1	0	1	0	0	1	1	0	0
1	1	0	0	0	1	0	1	1	0	1	1	1	0	0	1	1
2																
3																
4																
5																
6																
7																
8																
9																

Figure 1.5.1 Two examples of one dimensional (1-d) cellular automata. The top row in each case gives the initial conditions. The value of a cell at a particular time is given by a rule that depends on the values of the cells in its neighborhood at the previous time. For these rules the neighborhood consists of three cells: the cell itself and the two cells on either side. The first time step is shown below the initial conditions for (a) the majority rule, where each cell is equal to the value of the majority of the cells in its neighborhood at the previous time and (b) the mod2 rule which sums the value of the cells in the neighborhood modulo two to obtain the value of the cell in the next time. The rules are written in Question 1.5.1. The rest of the time steps are to be filled in as part of this question. ∎

Question 1.5.2 The evolution of the mod2 rule is periodic in time. After eight updates, the initial state of the system is recovered in Fig. 1.5.1(b). Because the state of the system at a particular time determines uniquely the state at every succeeding time, this is an 8-cycle that will repeat itself. There are sixteen cells in the space shown in Fig. 1.5.1(b). Is the number of cells connected with the length of the cycle? Try a space that has eight cells (Fig. 1.5.2(a)).

Solution 1.5.2 For a space with eight cells, the maximum length of a cycle is four. We could also use an initial condition that has a space periodicity of four in a space with eight cells (Fig. 1.5.2(b)). Then the cycle length would only be two. From these examples we see that the mod2 rule returns to the initial value after a time that depends upon the size of the space. More precisely, it depends on the periodicity of the initial conditions. The time periodicity (cycle length) for these examples is simply related to the space periodicity. ▮

Question 1.5.3 Look at the mod2 rule in a space with six cells (Fig. 1.5.2(c)) and in a space with five cells (Fig. 1.5.2(d)) . What can you conclude from these trials?

Solution 1.5.3 The mod2 rule can behave quite differently depending on the periodicity of the space it is in. The examples in Question 1.5.1 and 1.5.2 considered only spaces with a periodicity given by 2^k for some k. The new examples in this question show that the evolution of the rule may lead to a fixed point much like the majority rule. More than one initial condition leads to the same fixed point. Both the example shown and the fixed point itself does. Systematic analyses of the cycles and fixed points (cycles of period one) for this and other rules of this type,and various boundary conditions have been performed. ▮

The choice of initial conditions is an important aspect of the operation of many CA. Computer investigations of CA often begin by assuming a "seed" consisting of a single cell with the value +1 (a single ON cell) and all the rest −1 (OFF). Alternatively, the initial conditions may be chosen to be random: $s(i, j, k;0) = \pm 1$ with equal probability. The behavior of the system with a particular initial condition may be assumed to be generic, or some quantity may be averaged over different choices of initial conditions.

Like the iterative maps we considered in Section 1.1,the CA dynamics may be described in terms of cycles and attractors. As long as we consider only binary variables and a finite space, the dynamics must repeat itself after no more than a number of steps equal to the number of possible states of the system. This number grows exponentially with the size of the space. There are 2^N states of the system when there are a total of N cells. For 100 cells the length of the longest possible cycle would be of order 10^{30}. To consider such a long time for a small space may seem an unusual model of space-time. For most analogies of CA with physical systems,this model of space-time is not the most appropriate. We might restrict the notion of cycles to apply only when their length does not grow exponentially with the size of the system.

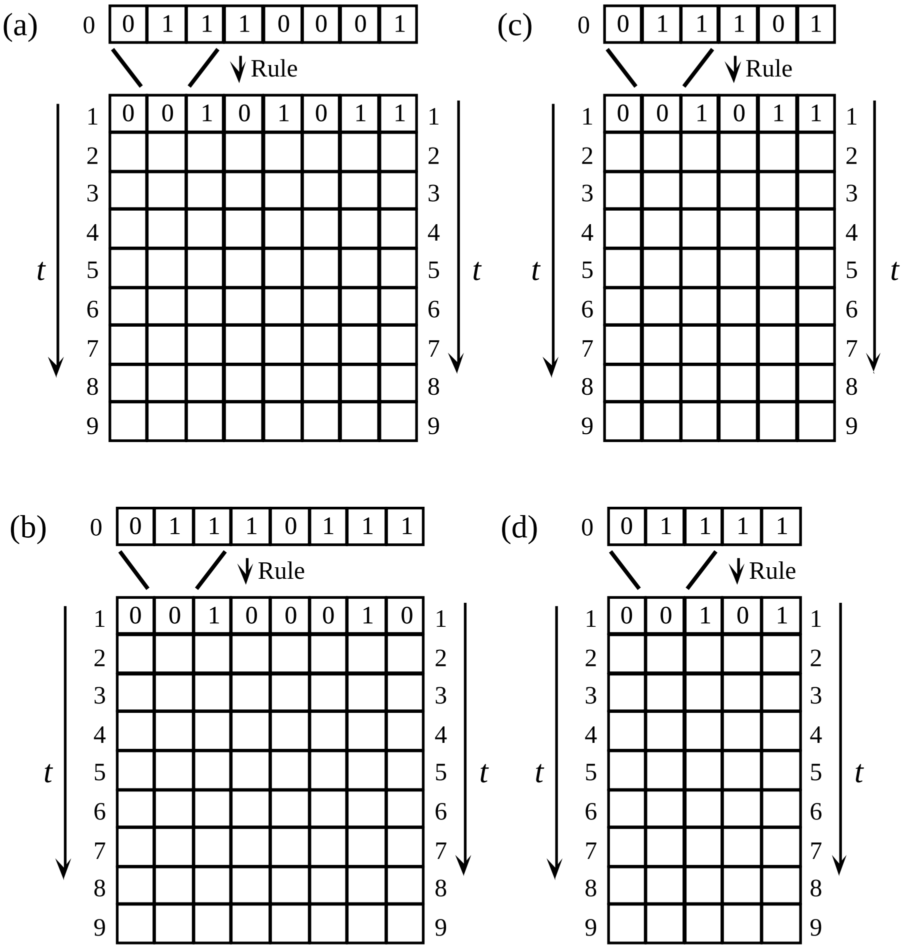


Figure 1.5.2 Four additional examples for the mod2 rule that have different initial conditions with specific periodicity: (a) is periodic in 8 cells, (b) is periodic in 4 cells, though it is shown embedded in a space of periodicity 8, (c) is periodic in 6 cells, (d) is periodic in 5 cells. By filling in the spaces it is possible to learn about the effect of different periodicities on the iterative properties of the mod2 rule. In particular, the length of the repeat time (cycle length) depends on the spatial periodicity. The cycle length may also depend on the specific initial conditions. ▮

Rules can be distinguished from each other and classified according to a variety of features they may possess. For example, some rules are reversible and others are not. Any reversible rule takes each state onto a unique successor. Otherwise it would be impossible to construct a single valued inverse mapping. Even when a rule is reversible,it is not guaranteed that the inverse rule is itself a CA,since it may not depend only on the local values of the variables. An example is given in question 1.5.5.

Question 1.5.4 Which if any of the two rules in Fig 1.5.1 is reversible?

Solution 1.5.4 The majority rule is not reversible, because locally we cannot identify in the next time step the difference between sequences that contain (11111) and (11011), since both result in a middle three of (111).

A discussion of the mod2 rule is more involved,since we must take into consideration the size of the space. In the examples of Questions 1.5.1–1.5.3 we see that in the space of six cells the rule is not reversible. In this case several initial conditions lead to the same result. The other examples all appear to be reversible, since each initial condition is part of a cycle that can be run backward to invert the rule. It turns out to be possible to construct explicitly the inverse of the mod2 rule. This is done in Question 1.5.5. ▮

Extra Credit Question 1.5.5 Find the inverse of the mod2 rule,when this is possible. This question involves some careful algebraic manipulation and may be skipped.

Solution 1.5.5 To find the inverse of the mod2 rule,it is useful to recall that equality modulo 2 satisfies simple addition properties including:

$$s_1 = s_2 \qquad s_1 + s = s_2 + s \qquad\qquad \text{mod}_2 \qquad (1.5.6)$$

as well as the special property:

$$2s = 0 \qquad\qquad \text{mod}_2 \qquad (1.5.7)$$

Together these imply that variables may be moved from one side of the equality to the other:

$$s_1 + s = s_2 \qquad s_1 = s_2 + s \qquad\qquad \text{mod}_2 \qquad (1.5.8)$$

Our task is to find the value of all $s(i;t)$ from the values of $s(j;t + 1)$ that are assumed known. Using Eq. (1.5.8), the mod2 update rule (Eq. (1.5.5))

$$s(i;t + 1) = (s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad\qquad \text{mod}_2 \qquad (1.5.9)$$

can be rewritten to give us the value of a cell in a layer in terms of the next layer and its own neighbors:

$$s(i - 1;t) = s(i;t + 1) + s(i;t) + s(i + 1;t\,) \qquad\qquad \text{mod}_2 \qquad (1.5.10)$$

Substitute the same equation for the second term on the right (using one higher index) to obtain

$$s(i - 1;t) = s(i;t + 1) + [s(i + 1;t + 1) + s(i + 1;t) + s(i + 2;t)] + s(i + 1;t) \qquad \text{mod}_2 \qquad (1.5.11)$$

the last term cancels against the middle term of the parenthesis and we have:

$$s(i - 1;t) = s(i;t + 1) + s(i + 1;t + 1) + s(i + 2;t) \qquad\qquad \text{mod}_2 \qquad (1.5.12)$$

It is convenient to rewrite this with one higher index:

$$s(i;t) = s(i + 1;t + 1) + s(i + 2;t + 1) + s(i + 3;t) \qquad\qquad \text{mod}_2 \qquad (1.5.13)$$

Interestingly, this is actually the solution we have been looking for, though some discussion is necessary to show this. On the right side of the equation appear three cell values. Two of them are from the time $t + 1$, and one from the time t that we are trying to reconstruct. Since the two cell values from $t + 1$ are assumed known, we must know only $s(i + 3;\ t)$ in order to obtain $s(i;t)$. We can iterate this expression and see that instead we need to know $s(i + 6;t)$ as follows:

$$\begin{aligned} s(i;t) &= s(i+1;t+1) + s(i+2;t+1) \\ &+ s(i+4;t+1) + s(i+5;t+1) + s(i+6;t) \end{aligned} \quad \mathrm{mod}_2 \qquad (1.5.14)$$

There are two possible cases that we must deal with at this point. The first is that the number of cells is divisible by three,and the second is that it is not. If the number of cells N is divisible by three, then after iterating Eq. (1.5.13) a total of $N/3$ times we will have an expression that looks like

$$\begin{aligned} s(i;t) &= s(i+1;t+1) + s(i+2;t+1) \\ &+ s(i+4;t+1) + s(i+5;t+1) + s(i+6;t) \\ &+ \ldots \\ &+ s(i+N-2;t+1) + s(i+N-1;t+1) + s(i;\ t) \end{aligned} \quad \mathrm{mod}_2 \qquad (1.5.15)$$

where we have used the property of the periodic boundary conditions to set $s(i + n;t) = s(i;t)$. We can cancel this value from both sides of the equation. What is left is an equation that states that the sum over particular values of the cell variables at time $t + 1$ must be zero.

$$\begin{aligned} 0 &= s(i+1;\ t+1) + s(i+2;\ t+1) \\ &+ s(i+4;\ t+1) + s(i+5;\ t+1) + s(i+6;\ t) \\ &+ \ldots \\ &+ s(i+N-2;\ t+1) + s(i+N-1;\ t+1) \end{aligned} \quad \mathrm{mod}_2 \qquad (1.5.16)$$

This means that any set of cell values that is the result of the mod2 rule update must satisfy this condition. Consequently, not all possible sets of cell values can be a result of mod2 updates. Thus the rule is not one-to-one and is not invertible when N is divisible by 3.

When N is not divisible by three, this problem does not arise, because we must go around the cell ring three times before we get back to $s(i;t)$. In this case,the analogous equation to Eq.(1.5.16) would have every cell value appearing exactly twice on the right of the equation. This is because each cell appears in two out of the three travels around the ring. Since the cell values all appear twice,they cancel,and the equation is the tautology $0 = 0$. Thus in this case there is no restriction on the result of the mod2 rule.

We almost have a full procedure for reconstructing $s(i;\ t)$. Choose the value of one particular cell variable, say $s(1;t) = 0$. From Eq.(1.5.13), obtain in sequence each of the cell variables $s(N - 2;t)$, $s(N - 5,t)$, . . . By going around the ring three times we can find uniquely all of the values. We now have to decide whether our original choice was correct. This can be done by directly applying the mod2 rule to find the value of say, $s(1; t + 1)$. If we obtain the right value, then we have the right choice; if the wrong value, then all we have to do is switch all of the cell values to their opposites. How do we know this is correct?

There was only one other possible choice for the value of $s(1; t) = 1$. If we were to choose this case we would find that each cell value was the opposite, or one's complement, $1 - s(i; t)$ of the value we found. This can be seen from Eq. (1.5.13). Moreover, the mod2 rule preserves complementation. Which means that if we complement all of the values of $s(i; t)$ we will find the complements of the values of $s(1; t + 1)$. The proof is direct:

$$\begin{aligned} &1 - s(i;t + 1) = 1 - (s(i - 1;t) + s(i;t) + s(i + 1;t)) \\ &= (1 - s(i - 1;t)) + (1 - s(i;t)) + (1 - s(i + 1;t))) - 2 \qquad \mathrm{mod}_2 \quad (1.5.17) \\ &= (1 - s(i - 1;t)) + (1 - s(i;t)) + (1 - s(i + 1;t))) \end{aligned}$$

Thus we can find the unique predecessor for the cell values $s(i;t + 1)$. With some care it is possible to write down a fully algebraic expression for the value of $s(i;t)$ by implementing this procedure algebraically. The result for $N = 3k + 1$ is:

$$s(i;t) = s(i;t + 1) + \sum_{j=1}^{(N-1)/3} (s(i + 3j - 2;t + 1) + s(i + 3j;t + 1)) \quad \mathrm{mod}_2 \quad (1.5.18)$$

A similar result for $N = 3k + 2$ can also be found.

Note that the inverse of the mod2 rule is not a CA because it is not a local rule. ▮

One of the interesting ways to classify CA—introduced by Wolfram—separates them into four classes depending on the nature of their limiting behavior. This scheme is particularly interesting for us,since it begins to identify the concept of complex behavior, which we will address more fully in a later chapter. The notion of complex behavior in a spatially distributed system is at least in part distinct from the concept of chaotic behavior that we have discussed previously. Specifically, the classification scheme is:

Class-one CA: evolve to a fixed homogeneous state

Class-two CA: evolve to fixed inhomogeneous states or cycles

Class-three CA: evolve to chaotic or aperiodic behavior

Class-four CA: evolve to complex localized structures

One example of each class is given in Fig. 1.5.3. It is assumed that the length of the cycles in class-two automata does not grow as the size of the space increases. This classification scheme has not yet found a firm foundation in analytical work and is supported largely by observation of simulations of various CA.

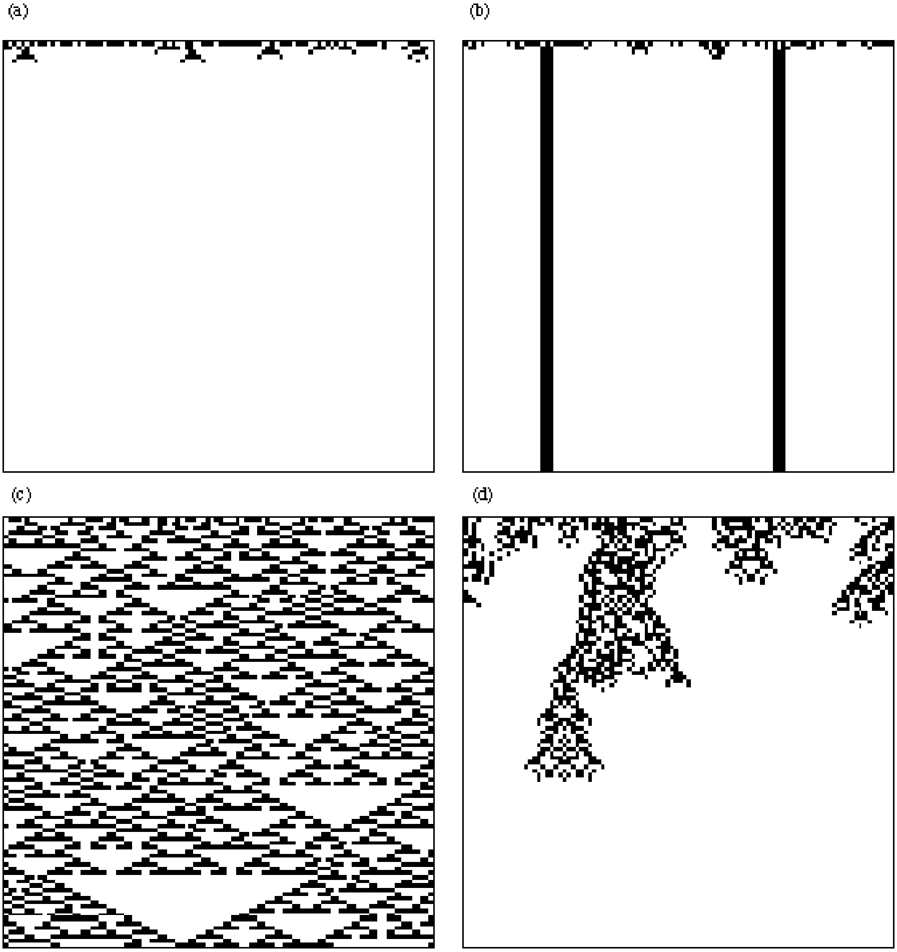


Figure 1.5.3 Illustration of four CA update rules with random initial conditions that are in a periodic space with a period of 100 cells. The initial conditions are shown at the top and time proceeds downward. Each is updated for 100 steps. ON cells are indicated as filled squares. OFF cells are not shown. Each of the rules gives the value of a cell in terms of a neighborhood of five cells at the previous time. The neighborhood consists of the cell itself and the two cells to the left and to the right. The rules are known as "totalistic" rules since they depend only on the sum of the variables in the neighborhood. Using the notation $s_i = 0,1$, the rules may be represented using $\Sigma_i(t) = s_{i-2}(t-1) + s_{i-1}(t-1) + s_i(t-1) + s_{i+1}(t-1) + s_{i+2}(t-1)$ by specifying the values of $\Sigma_i(t)$ for which $s_i(t)$ is ON. These are (a) only $\Sigma_i(t) = 2$, (b) only $\Sigma_i(t) = 3$, (c) $\Sigma_i(t) = 1$ and 2, and (d) $\Sigma_i(t) = 2$ and 4. See paper 1.3 in Wolfram's collection of articles on CA. ▮

It has been suggested that class-four automata have properties that enable them to be used as computers.Or, more precisely, to simulate a computer by setting the initial conditions to a set of data representing both the program and the input to the program. The result of the computation is to be obtained by looking some time later at the state of the system. A criteria that is clearly necessary for an automaton to be able to act as a computer is that the result of the dynamics is sensitive to the initial conditions. We will discuss the topic of computation further in Section 1.8.

The flip side of the use of a CA as a model of computation is to design a computer that will simulate CA with high efficiency. Such machines have been built, and are called cellular automaton machines (CAMs).

1.5.2 *2-d cellular automata*

Two- and three-dimensional CA provide more opportunities for contact with physical systems. We illustrate by describing an example of a 2-d CA that might serve as a simple model of droplet growth during condensation. The rule,illustrated in part pictorially in Fig. 1.5.4, may be described by saying that a particular cell with four or

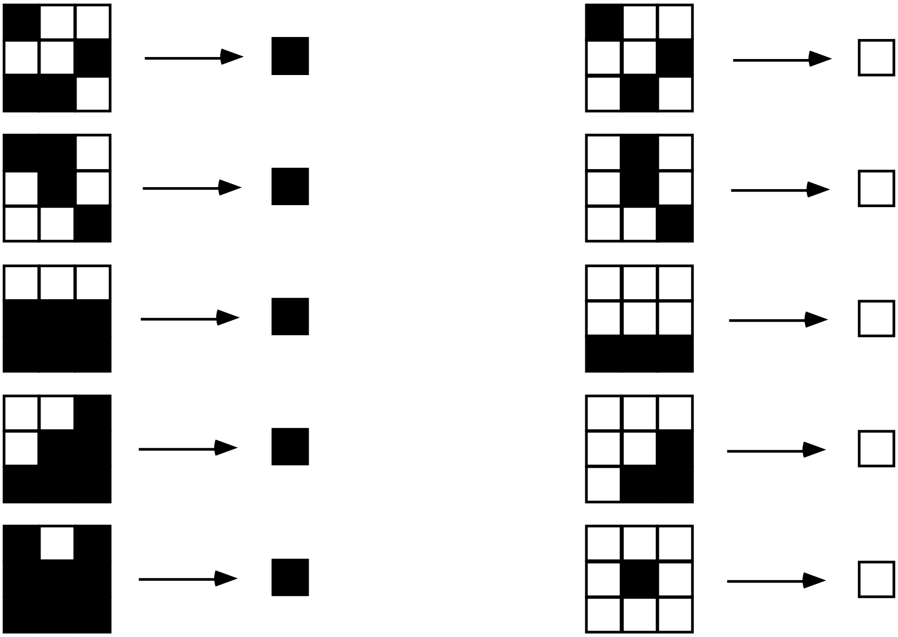


Figure 1.5.4 Illustration of a 2-d CA that may be thought of as a simple model of droplet condensation. The rule sets a cell to be ON (condensed) if four or more of its neighbors are condensed in the previous time, and OFF (uncondensed) otherwise. There are a total of 2^9=512 possible initial configurations; of these only 10 are shown. The ones on the left have 4 or more cells condensed and the ones on the right have less than 4 condensed. This rule is explained further by Fig. 1.5.5 and simulated in Fig. 1.5.6. ∎

more “condensed” neighbors at time t is condensed at time $t + 1$. Neighbors are counted from the 3 × 3 square region surrounding the cell, including the cell itself.

Fig. 1.5.5 shows a simulation of this rule starting from a random initial starting point of approximately 25% condensed (ON) and 75% uncondensed (OFF) cells. Over the first few updates, the random arrangement of dots resolves into droplets, where isolated condensed cells disappear and regions of higher density become the droplets. Then over a longer time, the droplets grow and reach a stable configuration.

The characteristics of this rule may be understood by considering the properties of boundaries between condensed and uncondensed regions,as shown in Fig. 1.5.6. Boundaries that are vertical,horizontal or at a 45° diagonal are stable. Other boundaries will move,increasing the size of the condensed region. Moreover, a concave corner of stable edges is not stable. It will grow to increase the condensed region.On the other hand,a convex corner is stable. This means that convex droplets are stable when they are formed of the stable edges.

It can be shown that for this size space,the 25% initial filling is a transition density, where sometimes the result will fill the space and sometimes it will not. For higher densities, the system almost always reaches an end point where the whole space is condensed. For lower densities, the system almost always reaches a stable set of droplets.

This example illustrates an important point about the dynamics of many systems, which is the existence of phase transitions in the kinetics of the system. Such phase transitions are similar in some ways to the thermodynamic phase transitions that describe the equilibrium state of a system changing from, for example,a solid to a liquid. The kinetic phase transitions may arise from the choice of initial conditions, as they did in this example. Alternatively, the phase transition may occur when we consider the behavior of a class of CA as a function of a parameter. The parameter gradually changes the local kinetics of the system; however, measures of its behavior may change abruptly at a particular value. Such transitions are also common in CA when the outcome of a particular update is not deterministic but stochastic, as discussed in Section 1.5.4.

1.5.3 *Conway’s Game of Life*

One of the most popular CA is known as Conway’s Game of Life. Conceptually, it is designed to capture in a simple way the reproduction and death of biological organisms. It is based on a model where,locally, if there are too few organisms or too many organisms the organisms will disappear. On the other hand,if the number of organisms is just right,they will multiply. Quite surprisingly, the model takes on a life of its own with a rich dynamical behavior that is best understood by direct observation.

The specific rule is defined in terms of the 3 × 3 neighborhood that was used in the last section. The rule,illustrated in Fig. 1.5.7,specifies that when there are less than three or more than four ON (populated) cells in the neighborhood,the central cell will be OFF (unpopulated) at the next time. If there are three ON cells,the central cell will be ON at the next time. If there are four ON cells,then the central cell will keep its previous state—ON if it was ON and OFF if it was OFF.

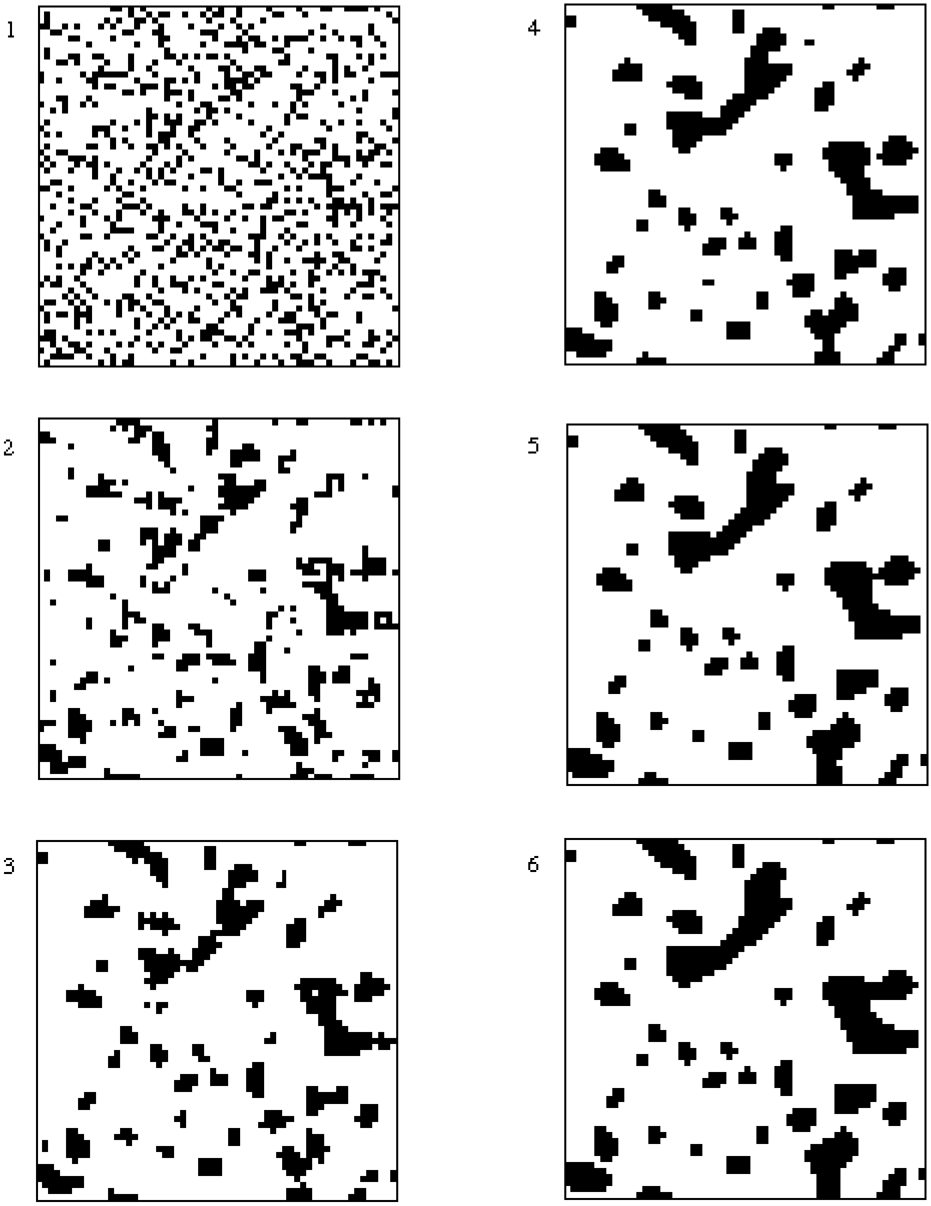


Figure 1.5.5 Simulation of the condensation CA described in Fig. 1.5.4. The initial conditions are chosen by setting randomly each site ON with a probability of 1 in 4. The initial few steps result in isolated ON sites disappearing and small ragged droplets of ON sites forming in higher-density regions. The droplets grow and smoothen their boundaries until at the sixtieth frame a static arrangement of convex droplets is reached. The first few steps are shown on the first page. Every tenth step is shown on the second page up to the sixtieth.

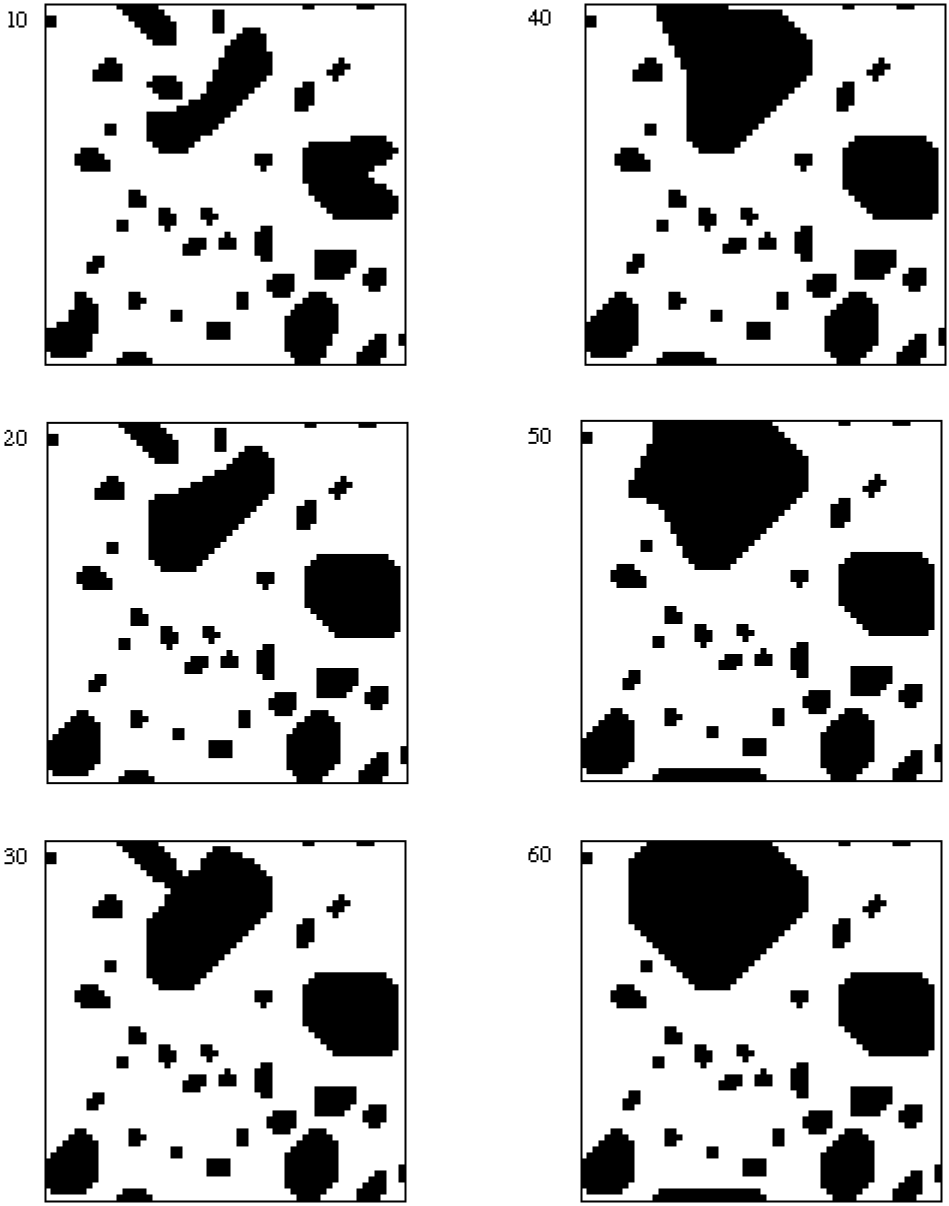


Figure 1.5.5 *Continued.* The initial occupation probability of 1 in 4 is near a phase transition in the kinetics of this model for a space of this size. For slightly higher densities the final configuration consists of a droplet covering the whole space. For slightly lower densities the final configuration is of isolated droplets. At a probability of 1 in 4 either may occur depending on the specific initial state. ❚

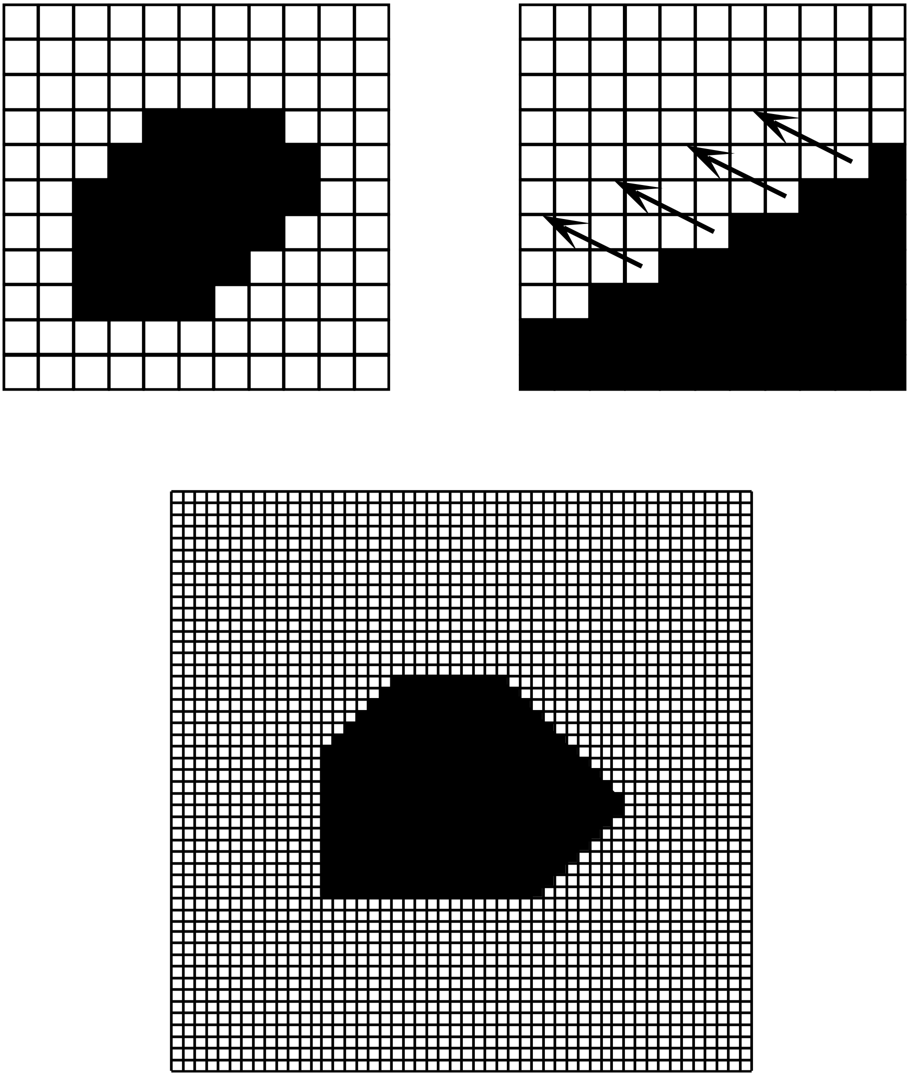


Figure 1.5.6 The droplet condensation model of Fig. 1.5.4 may be understood by noting that certain boundaries between condensed and uncondensed regions are stable. A completely stable shape is illustrated in the upper left. It is composed of boundaries that are horizontal, vertical or diagonal at 45°. A boundary that is at a different angle, such as shown on the upper right, will move, causing the droplet to grow. On a longer length scale a stable shape (droplet) is illustrated in the bottom figure. A simulation of this rule starting from a random initial condition is shown in Fig. 1.5.5. ▮

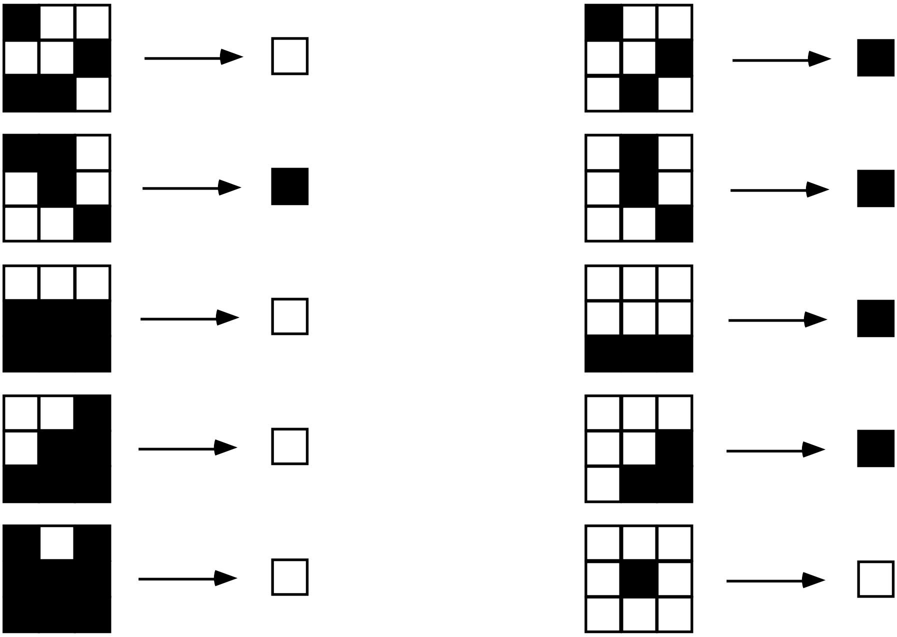


Figure 1.5.7 The CA rule Conway's Game of Life is illustrated for a few cases. When there are fewer than three or more than four neighbors in the 3 × 3 region the central cell is OFF in the next step. When there are three neighbors the central cell is ON in the next step. When there are four neighbors the central cell retains its current value in the next step. This rule was designed to capture some ideas about biological organism reproduction and death where too few organisms would lead to disappearance because of lack of reproduction and too many would lead to overpopulation and death due to exhaustion of resources. The rule is simulated in Fig. 1.5.8 and 1.5.9. ❚

Fig. 1.5.8 shows a simulation of the rule starting from the same initial conditions used for the condensation rule in the last section. Three sequential frames are shown, then after 100 steps an additional three frames are shown. Frames are also shown after 200 and 300 steps. After this amount of time the rule still has dynamic activity from frame to frame in some regions of the system, while others are apparently static or undergo simple cyclic behavior. An example of cyclic behavior may be seen in several places where there are horizontal bars of three ON cells that switch every time step between horizontal and vertical. There are many more complex local structures that repeat cyclically with much longer repeat cycles. Moreover, there are special structures called gliders that translate in space as they cycle through a set of configurations. The simplest glider is shown in Fig. 1.5.9,along with a structure called a glider gun, which creates them periodically.

We can make a connection between Conway's Game of Life and the quadratic iterative map considered in Section 1.1. The rich behavior of the iterative map was found because, for low values of the variable the iteration would increase its value, while for

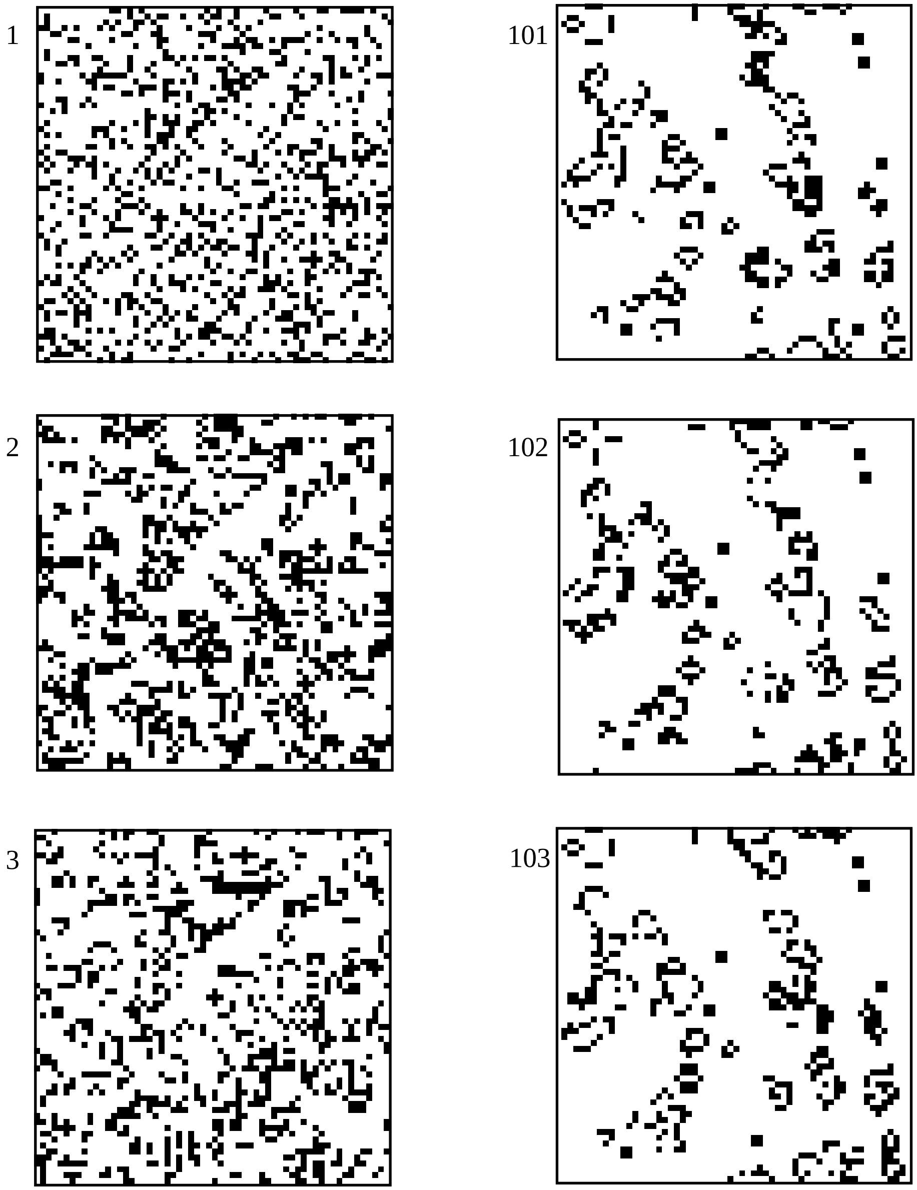


Figure 1.5.8 Simulation of Conway's Game of Life starting from the same initial conditions as used in Fig. 1.5.6 for the condensation rule where 1 in 4 cells are ON. Unlike the condensation rule there remains an active step-by-step evolution of the population of ON cells for many cycles. Illustrated are the three initial steps, and three successive steps each starting at steps 100, 200 and 300.

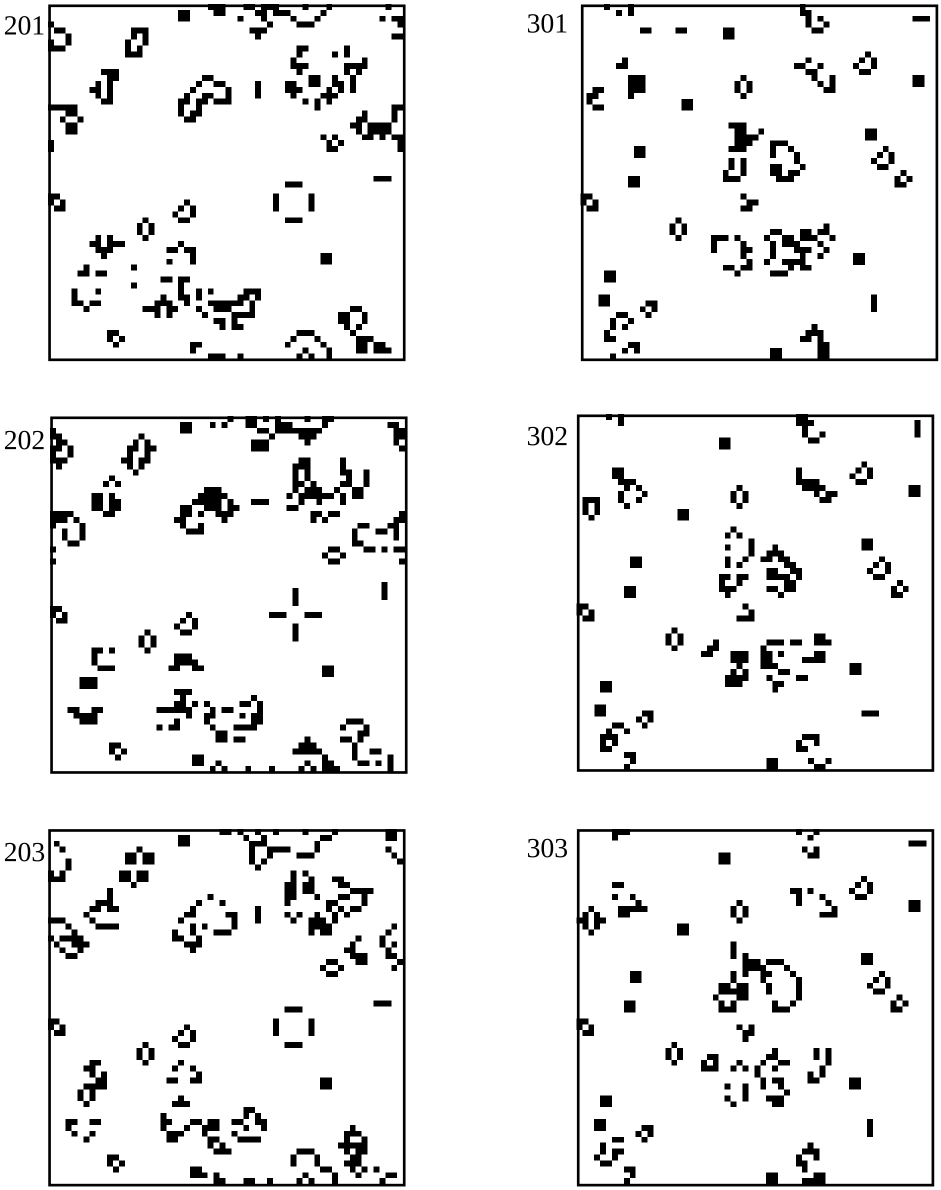


Figure 1.5.8 *Continued.* After the initial activity that occurs everywhere, the pattern of activity consists of regions that are active and regions that are static or have short cyclical activity. However, the active regions move over time around the whole space leading to changes everywhere. Eventually, after a longer time than illustrated here, the whole space becomes either static or has short cyclical activity. The time taken to relax to this state increases with the size of the space. ∎

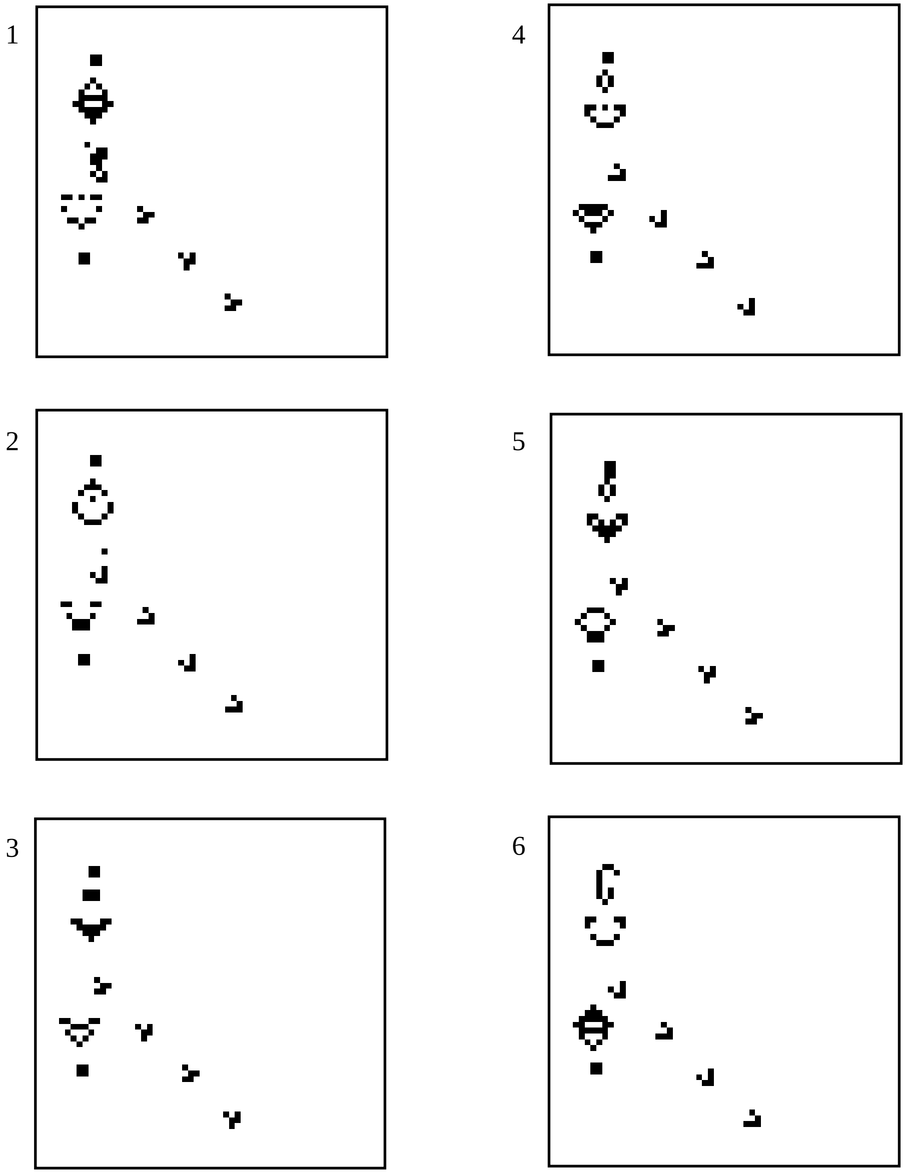


Figure 1.5.9 Special initial conditions simulated using Conway's Game of Life result in structures of ON cells called gliders that travel in space while progressing cyclically through a set of configurations. Several of the simplest type of gliders are shown moving toward the lower right. The more complex set of ON cells on the left, bounded by a 2 × 2 square of ON cells on top and bottom, is a glider gun. The glider gun cycles through 30 configurations during which a single glider is emitted. The stream of gliders moving to the lower right resulted from the activity of the glider gun. ▮

high values the iteration would decrease its value. Conway's Game of Life and other CA that exhibit interesting behavior also contain similar nonlinear feedback. Moreover, the spatial arrangement and coupling of the cells gives rise to a variety of new behaviors.

1.5.4 *Stochastic cellular automata*

In addition to the deterministic automaton of Eq. (1.5.3), we can define a stochastic automaton by the probabilities of transition from one state of the system to another:

$$P(\{s(i,j,k;\,t)\}|\{s(i,j,k;\,t-1)\}) \tag{1.5.19}$$

This general stochastic rule for the 2^N states of the system may be simplified. We have assumed for the deterministic rule that the rule for updating one cell may be performed independently of others. The analog for the stochastic rule is that the update probabilities for each of the cells is independent. If this is the case, then the total probability may be written as the product of probabilities of each cell value. Moreover, if the rule is local, the probability for the update of a particular cell will depend only on the values of the cell variables in the neighborhood of the cell we are considering.

$$P(\{s(i,j,k;t)\}\,|\,\{s(i,j,k;t-1)\}) = \prod_{i,j,k} P_0(s(i,j,k;t)\,|\,N(i,j,k;t-1)) \tag{1.5.20}$$

where we have used the notation $N(i,j,k;\,t)$ to indicate the values of the cell variables in the neighborhood of $(i,\,j,\,k)$. For example, we might consider modifying the droplet condensation model so that a cell value is set to be ON with a certain probability (depending on the number of ON neighbors) and OFF otherwise.

Stochastic automata can be thought of as modeling the effects of noise and more specifically the ensemble of a dynamic system that is subject to thermal noise. There is another way to make the analogy between the dynamics of a CA and a thermodynamic system that is exact—if we consider not the space of the automaton but the $d+1$ dimensional space-time. Consider the ensemble of all possible histories of the CA. If we have a three-dimensional space, then the histories are a set of variables with four indices $\{s(i,j,k,t)\}$. The probability of a particular set of these variables occurring (the probability of this history) is given by

$$P(\{s(i,j,k,t)\}) = \prod_t \prod_{i,j,k} P_0(s(i,j,k;t)\,|\,N(i,j,k;t-1))P(\{s(i,j,k;0)\}) \tag{1.5.21}$$

This expression is the product of the probabilities of each update occurring in the history. The first factor on the right is the probability of a particular initial state in the ensemble we are considering. If we consider only one starting configuration, its probability would be one and the others zero.

We can relate the probability in Eq. (1.5.21) to thermodynamics using Boltzmann probability. We simply set it to the expression for the Boltzmann probability at a particular temperature T.

$$P(\{s(i,j,k,t)\}) = e^{-E(\{s(i,j,k,t)\})/kT} \tag{1.5.22}$$

There is no need to include the normalization constant Z because the probabilities are automatically normalized. What we have done is to define the energy of the particular state as:

$$E(\{s(i,j,k,t)\}) = kT\ln\,(P(\{s(i,j,k,t)\})) \tag{1.5.23}$$

This expression shows that any d dimensional automaton can be related to a $d + 1$ dimensional system described by equilibrium Boltzmann probabilities. The ensemble of the $d + 1$ dimensional system is the set of time histories of the automaton.

There is an important cautionary note about the conclusion reached in the last paragraph. While it is true that time histories are directly related to the ensemble of a thermodynamic system,there is a hidden danger in this analogy. These are not typical thermodynamic systems, and therefore our intuition about how they should behave is not trustworthy. For example, the time direction may be very different from any of the space directions. For the $d + 1$ dimensional thermodynamic system, this means that one of the directions must be singled out. This kind of asymmetry does occur in thermodynamic systems, but it is not standard. Another example of the difference between thermodynamic systems and CA is in their sensitivity to boundary conditions. We have seen that many CA are quite sensitive to their initial conditions. While we have shown this for deterministic automata,it continues to be true for many stochastic automata as well. The analog of the initial conditions in a $d + 1$ dimensional thermodynamic system is the surface or boundary conditions. Thermodynamic systems are typically insensitive to their boundary conditions. However, the relationship in Eq.(1.5.23) suggests that at least some thermodynamic systems are quite sensitive to their boundary conditions. An interesting use of this analogy is to attempt to discover special thermodynamic systems whose behavior mimics the interesting behavior of CA.

1.5.5 *CA generalizations*

There are a variety of generalizations of the simplest version of CA which are useful in developing models of particular systems. In this section we briefly describe a few of them as illustrated in Fig. 1.5.10.

It is often convenient to consider more than one variable at a particular site. One way to think about this is as multiple spaces (planes in 2-d,lines in 1-d) that are coupled to each other. We could think about each space as a different physical quantity. For example, one might represent a magnetic field and the other an electric field. Another possibility is that we might use one space as a thermal reservoir. The system we are actually interested in might be simulated in one space and the thermal reservoir in another. By considering various combinations of multiple spaces representing a physical system, the nature of the physical system can become quite rich in its structure.

We can also consider the update rule to be a compound rule formed of a sequence of steps.Each of the steps updates the cells. The whole rule consists of cycling through the set of individual step rules. For example,our update rule might consist of two different steps. The first one is performed on every odd step and the second is performed on every even step. We could reduce this to the previous single update step case by looking at the composite of the first and second steps. This is the same as looking at only every even state of the system. We could also reduce this to a multiple space rule, where both the odd and even states are combined together to be a single step.

However, it may be more convenient at times to think about the system as performing a cycle of update steps.

Finally, we can allow the state of the system at a particular time to depend on the state of the system at several previous times,not just on the state of the system at the previous time.A rule might depend on the most recent state of the system and the previous one as well. Such a rule is also equivalent to a rule with multiple spaces, by considering both the present state of the system and its predecessor as two spaces. One use of considering rules that depend on more than one time is to enable systematic construction of reversible deterministic rules from nonreversible rules. Let the original (not necessarily invertible) rule be $R(N(i, j, k;\ t))$. A new invertible rule can be written using the form

$$s(i, j, k;\ t) = \text{mod}_2(R(N(i, j, k; t-1)) + s(i, j, k;\ t-2)) \qquad (1.5.24)$$

The inverse of the update rule is immediately constructed using the properties of addition modulo 2 (Eq. (1.5.8)) as:

$$s(i, j, k;\ t-2) = \text{mod}_2(R(N(i, j, k;\ t-1)) + s(i, j, k;\ t)) \qquad (1.5.25)$$

1.5.6 *Conserved quantities and Margolus dynamics*

Standard CA are not well suited to the description of systems with constraints or conservation laws. For example, if we want to conserve the number of ON cells we must establish a rule where turning OFF one cell (switching it from ON to OFF) is tied to turning ON another cell. The standard rule considers each cell separately when an update is performed. This makes it difficult to guarantee that when this particular cell is turned OFF then another one will be turned ON. There are many examples of physical systems where the conservation of quantities such as number of particles, energy and momentum are central to their behavior.

A systematic way to construct CA that describe systems with conserved quantities has been developed. Rules of this kind are known as partitioned CA or Margolus rules (Fig. 1.5.11). These rules separate the space into nonoverlapping partitions (also known as neighborhoods). The new value of each cell in a partition is given in terms of the previous values of the cells in the same partition. This is different from the conventional automaton, since the local rule has more than one output as well as more than one input. Such a rule is not sufficient in itself to describe the system update, since there is no communication in a single update between different partitions. The complete rule must specify how the partitions are shifted after each update with respect to the underlying space. This shifting is an essential part of the dynamical rule that restores the cellular symmetry of the space.

The convenience of this kind of CA is that specification of the rule gives us direct control of the dynamics within each partition, and therefore we can impose conservation rules within the partition. Once the conservation rule is imposed inside the partition, it will be maintained globally—throughout the whole space and through every time step. Fig. 1.5.12 illustrates a rule that conserves the number of ON cells inside a 2 × 2 neighborhood. The ON cells may be thought of as particles whose num-

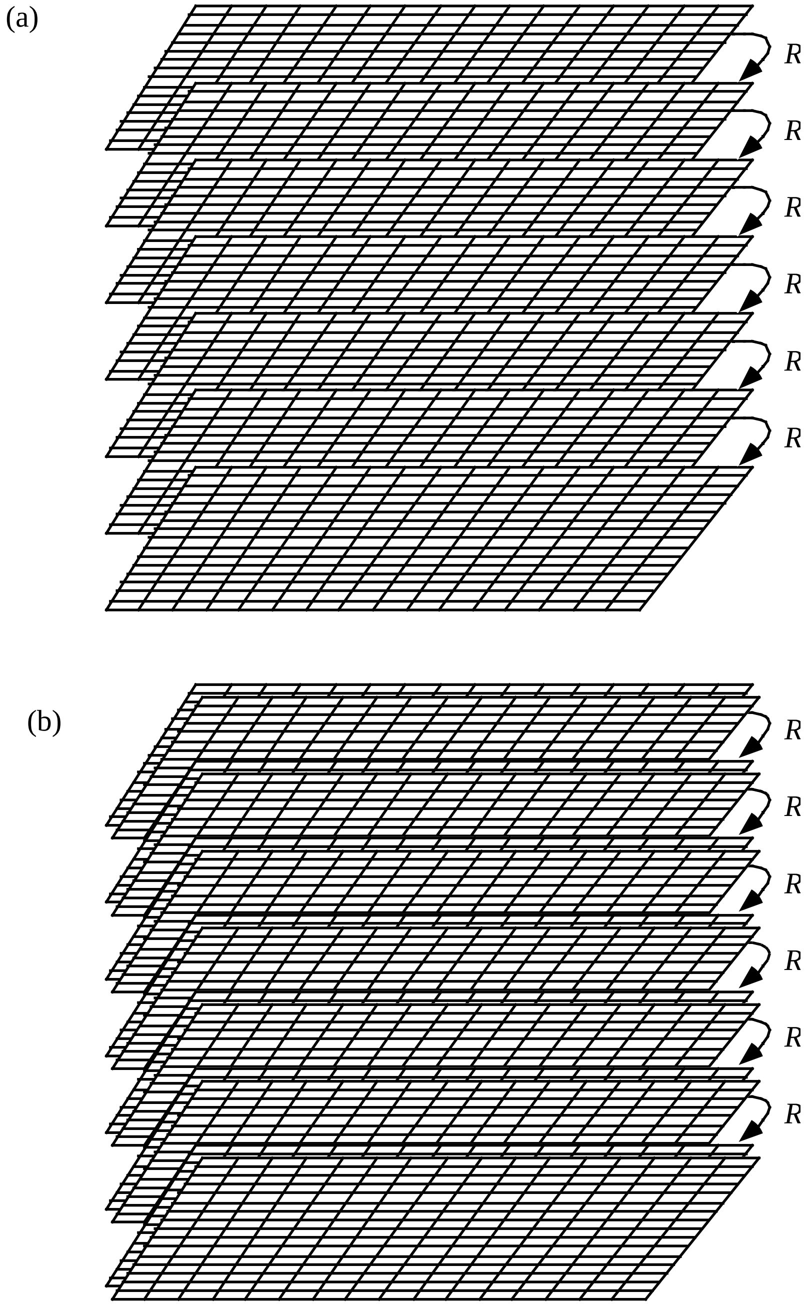
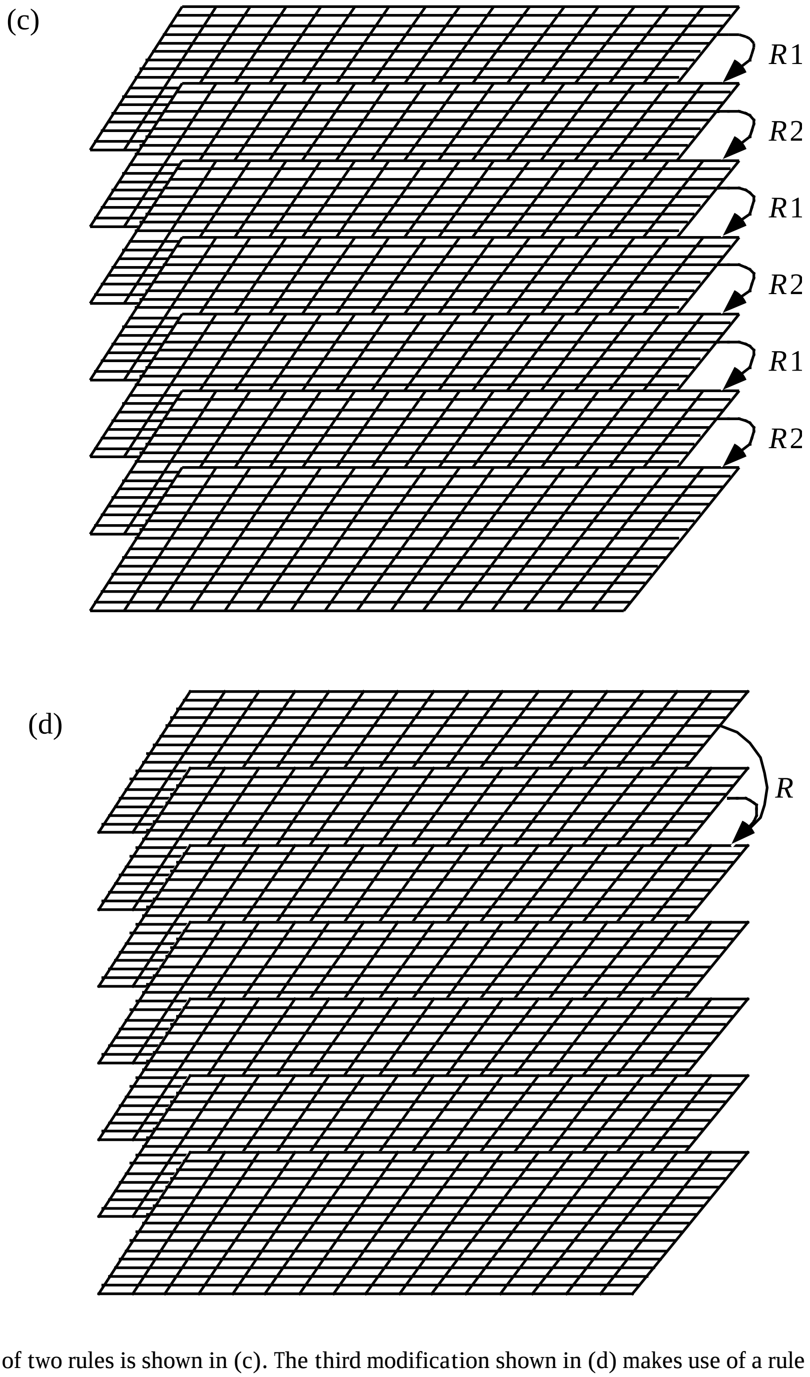
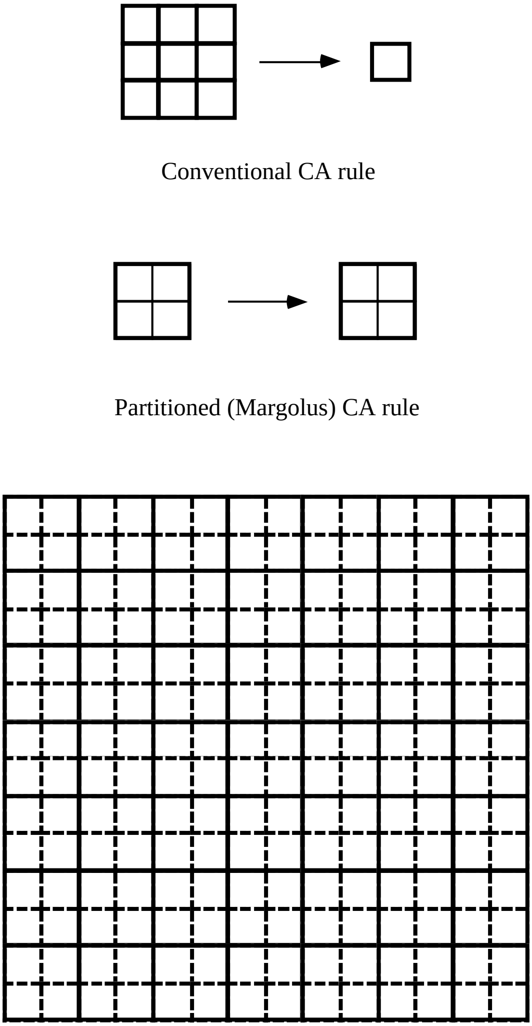


Figure 1.5.10 Schematic illustrations of several modifications of the simplest CA rule. The basic CA rule updates a set of spatially arrayed cell variables shown in (a). The first modification uses more than one variable in each cell. Conceptually this may be thought of as describing a set of coupled spaces, where the case of two spaces is shown in (b). The second modification makes use of a compound rule that combines several different rules, where the



case of two rules is shown in (c). The third modification shown in (d) makes use of a rule that depends on not just the most recent value of the cell variables but also the previous one. Both (c) and (d) may be described as special cases of (b) where two successive values of the cell variables are considered instead as occurring at the same time in different spaces. ∎



Conventional CA rule

Partitioned (Margolus) CA rule

Partition Alternation

Figure 1.5.11 Partitioned CA (Margolus rules) enable the imposition of conservation laws in a direct way. A conventional CA gives the value of an individual cell in terms of the previous values of cells in its neighborhood (top). A partitioned CA gives the value of several cells in a particular partition in terms of the previous values of the same cells (center). This enables conservation rules to be imposed directly within a particular partition. An example is given in Fig. 1.5.12. In addition to the rule for updating the partition, the dynamics must specify how the partitions are to be shifted from step to step. For example (bottom), the use of a 2 × 2 partition may be implemented by alternating the partitions from the solid lines to the dashed lines. Every even update the dashed lines are used and every odd update the solid lines are used to partition the space. This restores the cellular periodicity of the space and enables the cells to communicate with each other, which is not possible without the shifting of partitions. ❚

ber is conserved. The only requirement is that each of the possible arrangement of particles on the left results in an arrangement on the right with the same number of particles. This rule is augmented by specifying that the 2 × 2 partitions are shifted by a single cell to the right and down after every update. The motion of these particles is that of an unusual gas of particles.

The rule shown is only one of many possible that use this 2 × 2 neighborhood and conserve the number of particles. Some of these rules have additional properties or symmetries.A rule that is constructed to conserve particles may or may not be reversible. The one illustrated in Fig. 1.5.12 is not reversible. There exist more than one predecessor for particular values of the cell variables. This can be seen from the two mappings on the lower left that have the same output but different input.A rule that conserves particles also may or may not have a particular symmetry, such as a symmetry of reflection.A symmetry of reflection means that reflection of a configuration across a particular axis before application of the rule results in the same effect as reflection after application of the rule.

The existence of a well-defined set of rules that conserves the number of particles enables us to choose to study one of them for a specific reason. Alternatively, by randomly constructing a rule which conserves the number of particles, we can learn what particle conservation does in a dynamical system independent of other regularities of the system such as reversibility and reflection or rotation symmetries. More systematically, it is possible to consider the class of automata that conserve particle number and investigate their properties.

Question 1.5.6 Design a 2-d Margolus CA that represents a particle or chemical reaction: $A + B \leftrightarrow C$. Discuss some of the parameters that must be set and how you could use symmetries and conservation laws to set them.

Solution 1.5.6 We could use a 2 × 2 partition just like that in Fig. 1.5.12. On each of the four squares there can appear any one of the four possibilities (O, A, B, C). There are $4^4 = 256$ different initial conditions of the partition.Each of these must be paired with one final condition,if the rule is deterministic. If the rule is probabilistic, then probabilities must be assigned for each possible transition.

To represent a chemical reaction, we choose cases where A and B are adjacent (horizontally or vertically) and replace them with a C and a 0. If we prefer to be consistent, we can always place the C where A was before. To go the other direction, we take cases where C is next to a 0 and replace them with an A and a B. One question we might ask is, Do we want to have a reaction whenever it is possible, or do we want to assign some probability for the reaction? The latter case is more interesting and we would have to use a probabilistic CA to represent it. In addition to the reaction, the rule would include particle motion similar to that in Fig. 1.5.12.

To apply symmetries, we could assume that reflection along horizontal or vertical axes, or rotations of the partition by 90° before the update, will have the same effect as a reflection or rotation of the partition after the

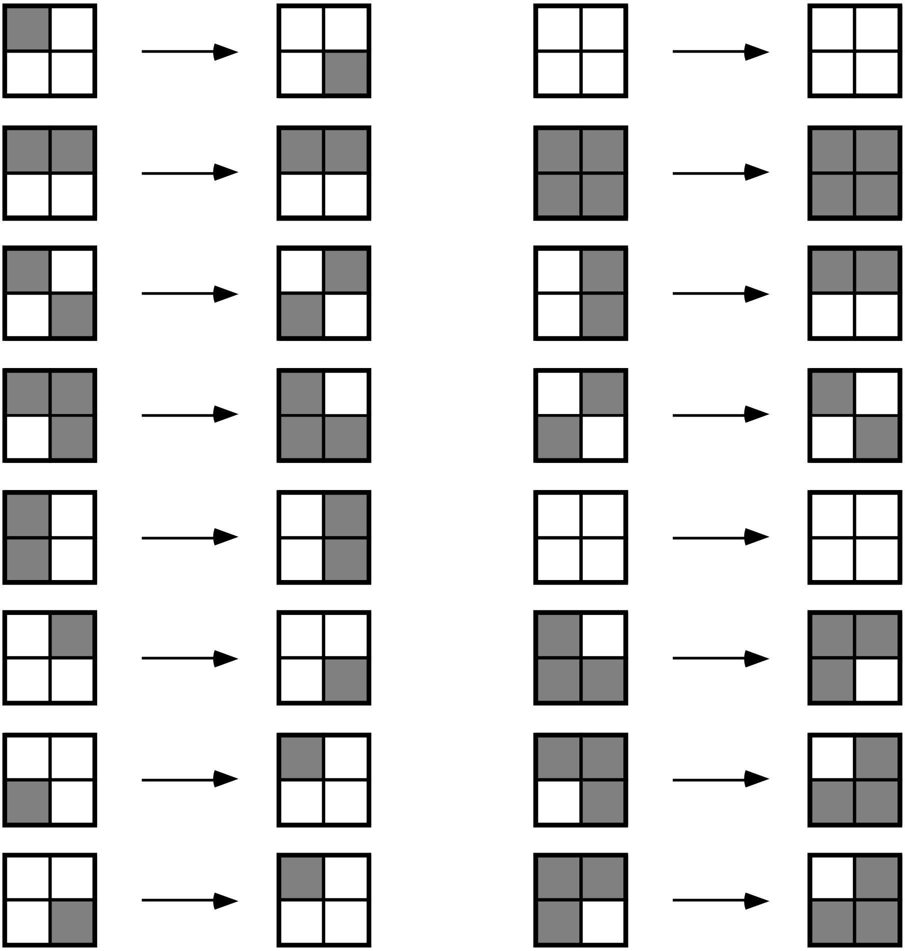


Figure 1.5.12 Illustration of a particular 2-d Margolus rule that preserves the number of ON cells which may be thought of as particles in a gas. The requirement for conservation of number of particles is that every initial configuration is matched with a final configuration having the same number of ON cells. This particular rule does not observe conventional symmetries such as reflection or rotation symmetries that might be expected in a typical gas. Many rules that conserve particles may be constructed in this framework by changing around the final states while preserving the number of particles in each case. ∎

update. We could also assume that *A*, *B* and *C* move in the same way when they are by themselves. Moreover, we might assume that the rule is symmetric under the transformation $A \leftrightarrow B$.

There is a simpler approach that requires enumerating many fewer states. We choose a 2 × 1 rectangular partition that has only two cells,and $4^2 = 16$ possible states. Of these, four do not change: [*A*,*A*], [*B*,*B*], [*C*,*C*] and [0,0].

Eight others are paired because the cell values can be switched to achieve particle motion (with a certain probability): [A,0] ↔ [0,A], [B,0] ↔ [0,B], [C,A] ↔ [A,C],and [C,B] ↔ [B,C].Finally, the last four, [C,0],[0,C], [A,B] and [B, A],can participate in reactions. If the rule is deterministic,they must be paired in a unique way for possible transitions. Otherwise,each possibility can be assigned a probability: [C,0] ↔ [A,B],[0,C] ↔ [B,A],[C,0] ↔ [B,A] and [0,C] ↔ [A,B]. The switching of the particles without undergoing reaction for these states may also be allowed with a certain probability. Thus,each of the four states can have a nonzero transition probability to each of the others. These probabilities may be related by the symmetries mentioned before. Once we have determined the update rule for the 2x1 partition, we can choose several ways to map the partitions onto the plane.The simplest are obtained by dividing each of the 2 × 2 partitions in Fig. 1.5.11 horizontally or vertically. This gives a total of four ways to partition the plane.These four can alternate when we simulate this CA. ▮

1.5.7 *Differential equations and CA*

Cellular automata are an alternative to differential equations for the modeling of physical systems. Differential equations when modeled numerically on a computer are often discretized in order to perform integrals. This discretization is an approximation that might be considered essentially equivalent to setting up a locally discrete dynamical system that in the macroscopic limit reduces to the differential equation. Why not then start from a discrete system and prove its relevance to the problem of interest? This a priori approach can provide distinct computational advantages. This argument might lead us to consider CA as an approximation to differential equations. However, it is possible to adopt an even more direct approach and say that differential equations are themselves an approximation to aspects of physical reality. CA are a different but equally valid approach to approximating this reality. In general, differential equations are more convenient for analytic solution while CA are more convenient for simulations. Since complex systems of differential equations are often solved numerically anyway, the alternative use of CA appears to be worth systematic consideration.

While both cellular automata and differential equations can be used to model macroscopic systems,this should not be taken to mean that the relationship between differential equations and CA is simple. Recognizing a CA analog to a standard differential equation may be a difficult problem.One of the most extensive efforts to use CA for simulation of a system more commonly known by its differential equation is the problem of hydrodynamics. Hydrodynamics is typically modeled by the Navier-Stokes equation. A type of CA called a lattice gas (Section 1.5.8) has been designed that on a length scale that is large compared to the cellular scale reproduces the behavior of the Navier-Stokes equation. The difficulties of solving the differential equation for specific boundary conditions make this CA a powerful tool for studying hydrodynamic flow.

A frequently occurring differential equation is the wave equation. The wave equation describes an elastic medium that is approximated as a continuum. The wave equation emerges as the continuum limit of a large variety of systems. It is to be expected that many CA will also display wavelike properties. Here we use a simple example to illustrate one way that wavelike properties may arise. We also show how the analogy may be quite different than intuition might suggest. The wave equation written in 1-d as

$$\frac{\partial^2 f}{\partial t^2} = c^2 \frac{\partial^2 f}{\partial x^2} \tag{1.5.26}$$

has two types of solutions that are waves traveling to the right and to the left with wave vectors k and frequencies of oscillation $\omega_k = ck$:

$$f = \sum_k \left(A_k e^{i(kx - \omega_k t)} + B_k e^{i(kx + \omega_k t)} \right) \tag{1.5.27}$$

A particular solution is obtained by choosing the coefficients A_k and B_k. These solutions may also be written in real space in the form:

$$f = \tilde{A}(x - ct) + \tilde{B}(x + ct) \tag{1.5.28}$$

where

$$\begin{aligned} \tilde{A}(x) &= \sum_k A_k e^{ikx} \\ \tilde{B}(x) &= \sum_k B_k e^{ikx} \end{aligned} \tag{1.5.29}$$

are two arbitrary functions that specify the initial conditions of the wave in an infinite space.

We can construct a CA analog of the wave equation as illustrated in Fig. 1.5.13. It should be understood that the wave equation will arise only as a continuum or long wave limit of the CA dynamics. However, we are not restricted to considering a model that mimics a vibrating elastic medium. The rule we construct consists of a 1-d partitioned space dynamics.Each update, adjacent cells are paired into partitions of two cells each. The pairing switches from update to update,analogous to the 2-d example in Fig. 1.5.11. The dynamics consists solely of switching the contents of the two adjacent cells in a single partition. Starting from a particular initial configuration, it can be seen that the contents of the odd cells moves systematically in one direction (right in the figure),while the contents of the even cells moves in the opposite direction (left in the figure). The movement proceeds at a constant velocity of $c = 1$ cell/update. Thus we identify the contents of the odd cells as the rightward traveling wave,and the even cells as the leftward traveling wave.

The dynamics of this CA is the same as the dynamics of the wave equation of Eq.(1.5.28) in an infinite space. The only requirement is to encode appropriately the initial conditions $\tilde{A}(x)$, $\tilde{B}(x)$ in the cells. If we use variables with values in the conven-

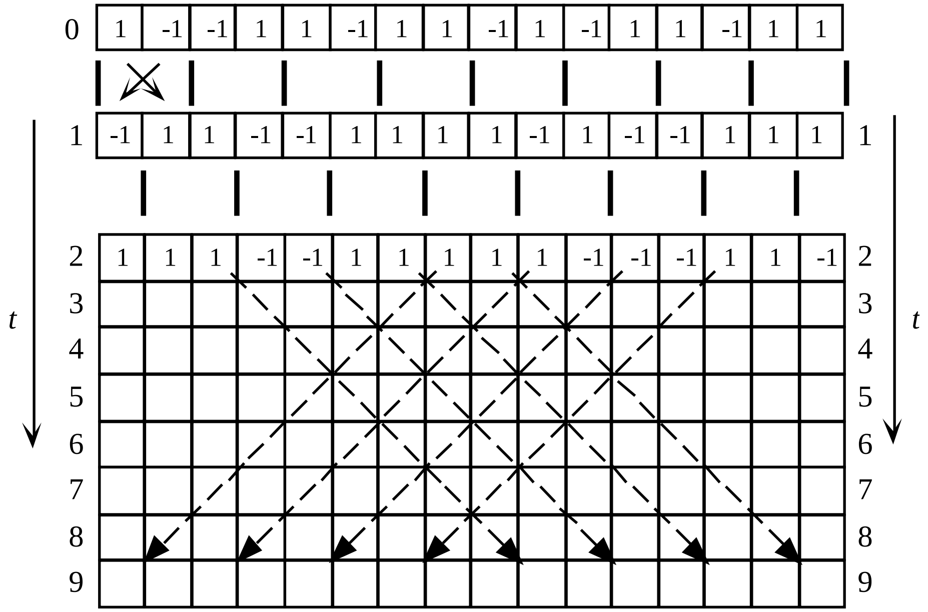


Figure 1.5.13 A simple 1-d CA using a Margolus rule, which switches the values of the two adjacent cells in the partition, can be used to model the wave equation. The partitions alternate between the two possible ways of partitioning the cells every time step. It can be seen that the initial state is propagated in time so that the odd (even) cells move at a fixed rate of one cell per update to the right (left). The solutions of the wave equation likewise consist of a right and left traveling wave. The initial conditions of the wave equation solution are the analog of the initial condition of the cells in the CA. ∎

tional real continuum s_i , then the (discretized) waves may be encoded directly. If a binary representation $s_i = \pm 1$ is used, the local average over odd cells represents the right traveling wave $\tilde{A}(x - ct)$,and the local average over even cells represents the left traveling wave $\tilde{B}(x + ct)$.

1.5.8 *Lattice gases*

A lattice gas is a type of CA designed to model gases or liquids of colliding particles. Lattice gases are formulated in a way that enables the collisions to conserve momentum as well as number of particles. Momentum is represented by setting the velocity of each particle to a discrete set of possibilities.A simple example, the HPP gas,is illustrated in Fig. 1.5.14.Each cell contains four binary variables that represent the presence (or absence) of particles with unit velocity in the four compass directions NESW. In the figure,the presence of a particle in a cell is indicated by an arrow. There can be up to four particles at each site.Each particle present in a single cell must have a distinct velocity.

The dynamics of the HPP gas is performed in two steps that alternate: propagation and collision. In the propagation step, particles move from the cell they are in to the neighboring cell in the direction of their motion. In the collision step, each cell acts independently, changing the particles from incoming to outgoing according to prespecified collision rules. The rule for the HPP gas is illustrated in Fig. 1.5.15. Because of momentum conservation in this rule, there are only two possibilities for changes in the particle velocity as a result of a collision.A similar lattice gas,the FHP gas, which is implemented on a hexagonal lattice of cells rather than a square lattice, has been proven to give rise to the Navier-Stokes hydrodynamic equations on a macroscopic scale. Due to properties of the square lattice in two dimensions, this behavior does not occur for the HPP gas. One way to understand the limitation of the square lattice is to realize that for the HPP gas (Fig. 1.5.14),momentum is conserved in any individual horizontal or vertical stripe of cells. This type of conservation law is not satisfied by hydrodynamics.

1.5.9 *Material growth*

One of the natural physical systems to model using CA is the problem of layer-by-layer material growth such as is achieved in molecular beam epitaxy. There are many areas of study of the growth of materials. For example,in cases where the material is formed of only a single type of atom,it is the surface structure during growth that is of interest. Here, we focus on an example of an alloy formed of several different atoms, where the growth of the atoms is precisely layer by layer. In this case the surface structure is simple, but the relative abundance and location of different atoms in the material is of interest. The simplest case is when the atoms are found on a lattice that is prespecified, it is only the type of atom that may vary.

The analogy with a CA is established by considering each layer of atoms, when it is deposited, as represented by a 2-d CA at a particular time. As shown in Fig. 1.5.16 the cell values of the automaton represent the type of atom at a particular site. The values of the cells at a particular time are preserved as the atoms of the layer deposited at that time. It is the time history of the CA that is to be interpreted as representing the structure of the alloy. This picture assumes that once an atom is incorporated in a complete layer it does not move.

In order to construct the CA, we assume that the probability of a particular atom being deposited at a particular location depends on the atoms residing in the layer immediately preceding it. The stochastic CA rule in the form of Eq.(1.5.20) specifies the probability of attaching each kind of atom to every possible atomic environment in the previous layer.

We can illustrate how this might work by describing a specific example. There exist alloys formed out of a mixture of gallium,arsenic and silicon.A material formed of equal proportions of gallium and arsenic forms a GaAs crystal, which is exactly like a silicon crystal, except the Ga and As atoms alternate in positions. When we put silicon together with GaAs then the silicon can substitute for either the Ga or the As atoms. If there is more Si than GaAs, then the crystal is essentially a Si crystal with small regions of GaAs,and isolated Ga and As. If there is more GaAs than Si,then the

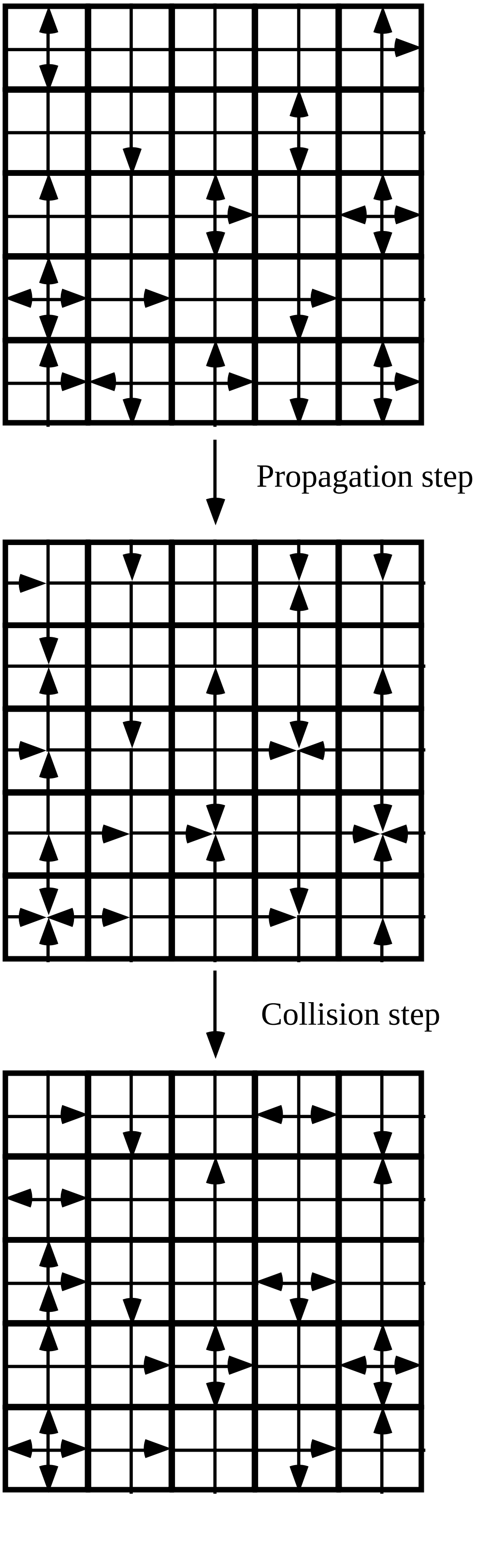


Figure 1.5.14 Illustration of the update of the HPP lattice gas. In a lattice gas, binary variables in each cell indicate the presence of particles with a particular velocity. Here there are four possible particles in each cell with unit velocities in the four compass directions, NESW. Pictorially the presence of a particle is indicated by an arrow in the direction of its velocity. Updating the lattice gas consists of two steps: propagating the particles according to their velocities, and allowing the particles to collide according to a collision rule. The propagation step consists of moving particles from each cell into the neighboring cells in the direction of their motion. The collision step consists of each cell independently changing the velocities of its particles. The HPP collision rule is shown in Fig. 1.5.15, and implemented here from the middle to the bottom panel. For convenience in viewing the different steps the arrows in this figure alternate between incoming and outcoming. Particles before propagation (top) are shown as outward arrows from the center of the cell. After the propagation step (middle) they are shown as incoming arrows. After collision (bottom) they are again shown as outgoing arrows. ∎

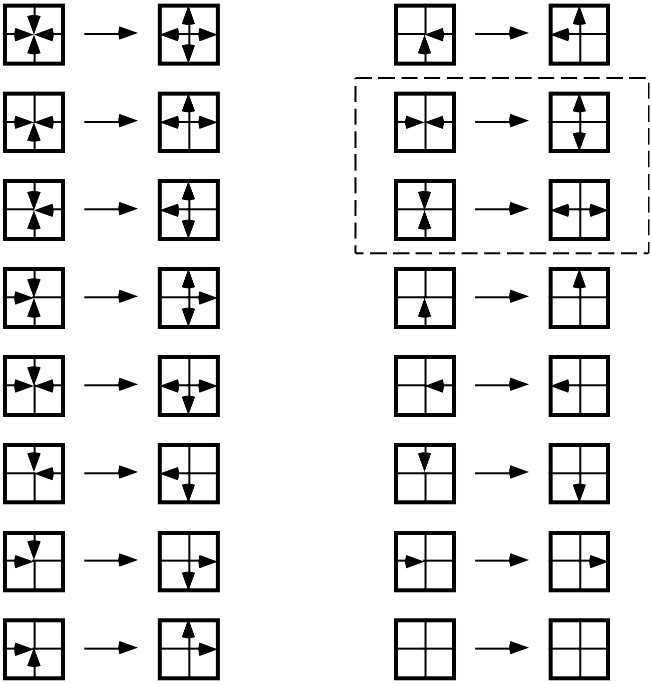


Figure 1.5.15 The collision rule for the HPP lattice gas. With the exception of the case of two particles coming in from N and S and leaving from E and W, or vice versa (dashed box), there are no changes in the particle velocities as a result of collisions in this rule. Momentum conservation does not allow any other changes. ∎

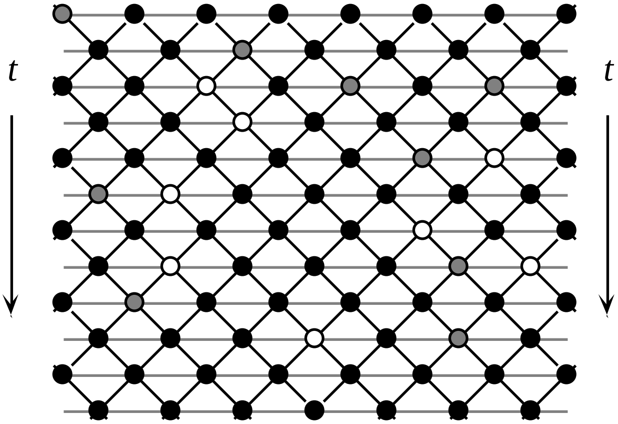


Figure 1.5.16 Illustration of the time history of a CA and its use to model the structure of a material (alloy) formed by a layer by layer growth. Each horizontal dashed line represents a layer of the material. The alloy has three types of atoms. The configuration of atoms in each layer depends only on the atoms in the layer preceding it. The type of atom, indicated in the figure by filled, empty and shaded dots, are determined by the values of the cell variables of the CA at a particular time, $s_i(t) = \pm 1{,}0$. The time history of the CA is the structure of the material. ∎

crystal will be essentially a GaAs crystal with isolated Si atoms. We can model the growth of the alloys formed by different relative proportions of GaAs and Si of the form $(\mathrm{GaAs})_{1\text{-}x}\mathrm{Si}_x$ using a CA. Each cell of the CA has a variable with three possible values $s_i = \pm 1{,}0$ that would represent the occupation of a crystal site by Ga, As and Si respectively. The CA rule (Eq. (1.5.20)) would then be constructed by assuming different probabilities for adding a Si, Ga and As atom at the surface. For example, the likelihood of finding a Ga next to a Ga atom or an As next to an As is small, so the probability of adding a Ga on top of a Ga can be set to be much smaller than other probabilities. The probability of an Si atom $s_i = 0$ could be varied to reflect different concentrations of Si in the growth. Then we would be able to observe how the structure of the material changes as the Si concentration changes.

This is one of many examples of physical, chemical and biological systems that have been modeled using CA to capture some of their dynamical properties. We will encounter others in later chapters.

1.6 Statistical Fields

In real systems as well as in kinetic models such as cellular automata (CA) discussed in the previous section, we are often interested in finding the state of a system—the time averaged (equilibrium) ensemble when cycles or randomness are present—that arises after the fast initial kinetic processes have occurred. Our objective in this section is to treat systems with many degrees of freedom using the tools of equilibrium statistical mechanics (Section 1.3).These tools describe the equilibrium ensemble directly rather than the time evolution. The simplest example is a collection of interacting binary variables, which is in many ways analogous to the simplest of the CA models. This model is known as the Ising model,and was introduced originally to describe the properties of magnets.Each of the individual variables corresponds to a microscopic magnetic region that arises due to the orbital motion of an electron or the internal degree of freedom known as the spin of the electron.

The Ising model is the simplest model of interacting degrees of freedom. Each of the variables is binary and the interactions between them are only specified by one parameter—the strength of the interaction. Remarkably, many complex systems we will be considering can be modeled by the Ising model as a first approximation. We will use several versions of the Ising model to discuss neural networks in Chapter 2 and proteins in Chapter 4. The reason for the usefulness of this model is the very existence of interactions between the elements. This interaction is not present in simpler models and results in various behaviors that can be used to understand some of the key aspects of complex systems. The concepts and tools that are used to study the Ising model also may be transferred to more complicated models. It should be understood, however, that the Ising model is a simplistic model of magnets as well as of other systems.

In Section 1.3 we considered the ideal gas with collisions. The collisions were a form of interaction. However, these interactions were incidental to the model because they were assumed to be so short that they were not present during observation. This is no longer true in the Ising model.

1.6.1 *The Ising model without interactions*

The Ising model describes the energy of a collection of elements (spins) represented by binary variables.It is so simple that there is no kinetics, only an energy $E[\{s_i\}]$.Later we will discuss how to reintroduce a dynamics for this model. The absence of a dynamics is not a problem for the study of the equilibrium properties of the system, since the Boltzmann probability (Eq.(1.3.29)) depends only upon the energy. The energy is specified as a function of the values of the binary variables $\{s_i = \pm 1\}$. Unless necessary, we will use one index for all of the spin variables regardless of dimensionality. The use of the term "spin" originates from the magnetic analogy. There is no other specific term,so we adopt this terminology. The term "spin" emphasizes that the binary variable represents the state of a physical entity such that the collection of spins is the system we are interested in.A spin can be illustrated as an arrow of fixed length (see Fig. 1.6.1). The value of the binary variable describes its orientation, where +1 indicates a spin oriented in the positive *z* direction (UP),and –1 indicates a spin oriented in the negative *z* direction (DOWN).

Before we consider the effects of interactions between the spins, we start by considering a system where there are no interactions. We can write the energy of such a system as:

$$E[\{s_i\}] = \sum_i e_i(s_i) \tag{1.6.1}$$

Where $e_i(s_i)$ is the energy of the *i*th spin that does not depend on the values of any of the other spins. Since s_i are binary we can write this as:

$$E[\{s_i\}] = \frac{1}{2}\sum_i (e_i(1) - e_i(-1))s_i + (e_i(1) + e_i(-1)) = E_0 - \sum_i h_i s_i \to -\sum_i h_i s_i \tag{1.6.2}$$

All of the terms that do not depend on the spin variables have been collected together into a constant. We set this constant to zero by redefining the energy scale. The quantities $\{h_i\}$ describe the energy due to the orientation of the spins. In the magnetic system they correspond to an external magnetic field that varies from location to location.Like small magnets, spins try to orient along the magnetic field. A spin oriented along the magnetic field (s_i and h_i have the same sign) has a lower energy than if it is antiparallel to the magnetic field. As in Eq.(1.6.2), the contribution of the magnetic field to the energy is $-|h_i|(|h_i|)$when the spin is parallel (antiparallel) to the field direction. When convenient we will simplify to the case of a uniform magnetic field, $h_i = h$.

When the spins are noninteracting, the Ising model reduces to a collection of two-state systems that we investigated in Section 1.4. Later, when we introduce interactions between the spins, there will be differences. For the noninteracting case we can write the probability for a particular configuration of the spins using the Boltzmann probability:

$$P[\{s_i\}] = \frac{e^{-\beta E[\{s_i\}]}}{Z} = \frac{e^{\beta \sum_i h_i s_i}}{Z} = \frac{\prod_i e^{\beta h_i s_i}}{Z} \tag{1.6.3}$$

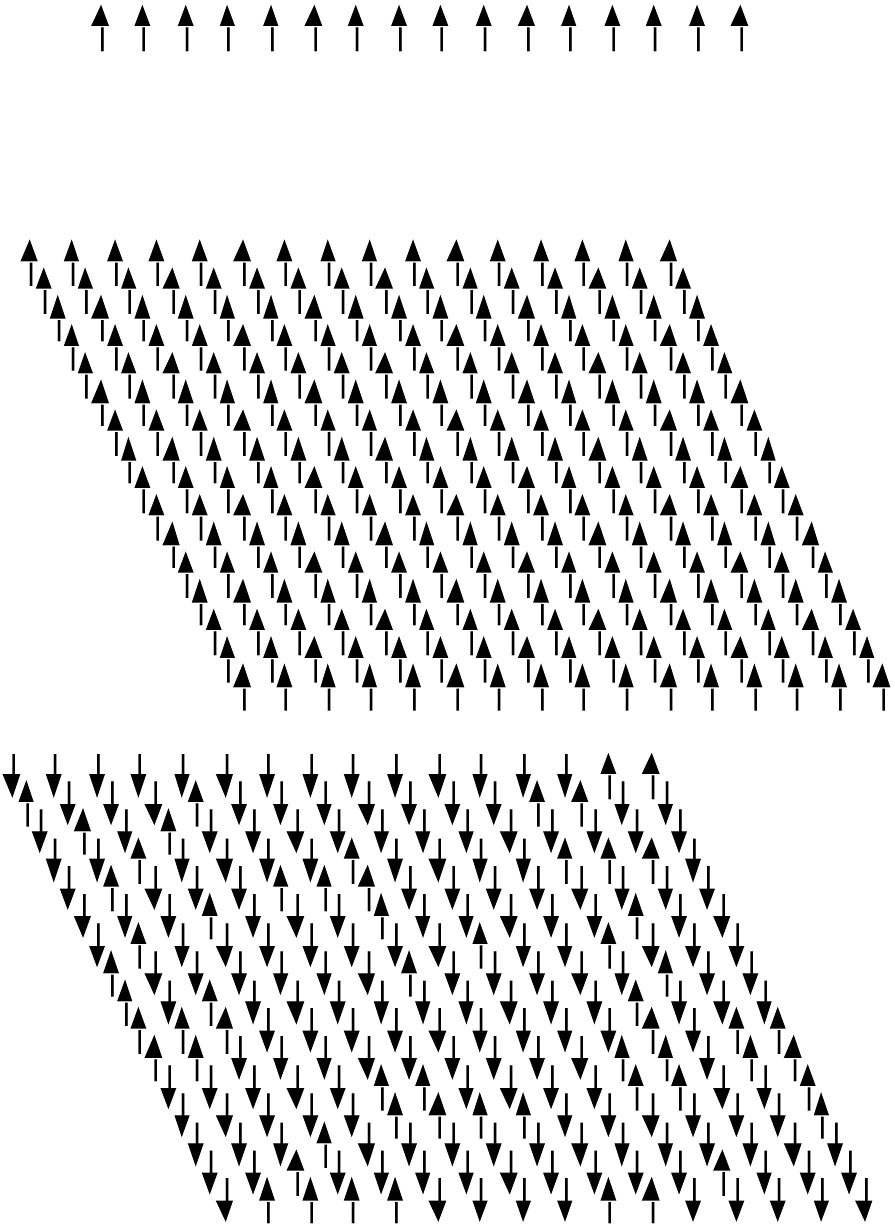


Figure 1.6.1 One way to visualize the Ising model is as a spatial array of binary variables called spins, represented as UP or DOWN arrows. A one-dimensional (1-d) example with all spins UP is shown on top. The middle and lower figures show two-dimensional (2-d) arrays which have all spins UP (middle) or have some spins UP and some spins DOWN (bottom). ▮

where $\beta = 1/kT$. The partition function Z is given by:

$$Z = \sum_{\{s_i\}} e^{-\beta E[\{s_i\}]} = \sum_{\{s_i\}} \prod_i e^{\beta h_i s_i} = \prod_i \sum_{s_i} e^{\beta h_i s_i} = \prod_i \left(e^{\beta h_i} + e^{-\beta h_i} \right) \tag{1.6.4}$$

where the second to last equality replaces the sum over all possible values of the spin variables with a sum over each spin variable $s_i = \pm 1$ within the product. Thus the probability factors as:

$$P[\{s_i\}] = \prod_i P(s_i) = \prod_i \left(\frac{e^{\beta h_i s_i}}{e^{\beta h_i} + e^{-\beta h_i}} \right) \tag{1.6.5}$$

This is a product over the result we found for probability of the two-state system (Eq. (1.4.14)) if we write the energy of a single spin using the notation $E_i(s_i) = -h_i s_i$.

Now that we have many spin variables, we can investigate the thermodynamics of this model by writing down the free energy and entropy of this model. This is discussed in Question 1.6.1.

Question 1.6.1 Evaluate the thermodynamic free energy, energy and entropy for the Ising model without interactions.

Solution 1.6.1 The free energy is given in terms of the partition function by Eq. (1.3.37):

$$F = -kT\ln(Z) = -kT\sum_i \ln\left(e^{\beta h_i} + e^{-\beta h_i}\right) = -kT\sum_i \ln\left(2\cosh\left(\beta h_i\right)\right) \tag{1.6.6}$$

The latter expression is a more common way of writing this result.

The thermodynamic energy of the system is found from Eq.(1.3.38) as

$$U = -\frac{\partial \ln(Z)}{\partial \beta} = -\sum_i \frac{h_i(e^{\beta h_i} - e^{-\beta h_i})}{(e^{\beta h_i} + e^{-\beta h_i})} = -\sum_i h_i \tanh(\beta h_i) \tag{1.6.7}$$

There is another way to obtain the same result. The thermodynamic energy is the average energy of the system (Eq.(1.3.30)), which can be evaluated directly:

$$\begin{aligned} U &= \langle E[\{s_i\}] \rangle = \left\langle -\sum_i h_i s_i \right\rangle = -\sum_i h_i \langle s_i \rangle = -\sum_i h_i \sum_{s_i} s_i P(s_i) \\ &= -\sum_i h_i \frac{(e^{\beta h_i} - e^{-\beta h_i})}{(e^{\beta h_i} + e^{-\beta h_i})} = -\sum_i h_i \tanh(\beta h_i) \end{aligned} \tag{1.6.8}$$

which is the same as before. We have used the possibility of writing the probability of a single spin variable independent of the others in order to perform this average. It is convenient to define the local magnetization m_i as the average value of a particular spin variable:

$$m_i = \langle s_i \rangle = \sum_{s_i = \pm 1} s_i P_{s_i}(s_i) = P_{s_i}(1) - P_{s_i}(-1) \tag{1.6.9}$$

Or using Eq. (1.6.5):

$$m_i = \langle s_i \rangle = \tanh(\beta h_i) \tag{1.6.10}$$

In Fig. 1.6.2, the magnetization at a particular site is plotted as a function of the magnetic field for several different temperatures ($\beta = 1/kT$). The magnetization increases with increasing magnetic field and with decreasing temperature until it saturates asymptotically to a value of +1 or –1. In terms of the magnetization the energy is:

$$U = -\sum_i h_i m_i \tag{1.6.11}$$

We can calculate the entropy of the Ising model using Eq. (1.3.36)

$$S = k\beta U + k \ln Z = -k\sum_i \beta h_i \tanh(\beta h_i) + k\sum_i \ln\left(2\cosh\left(\beta h_i\right)\right) \tag{1.6.12}$$

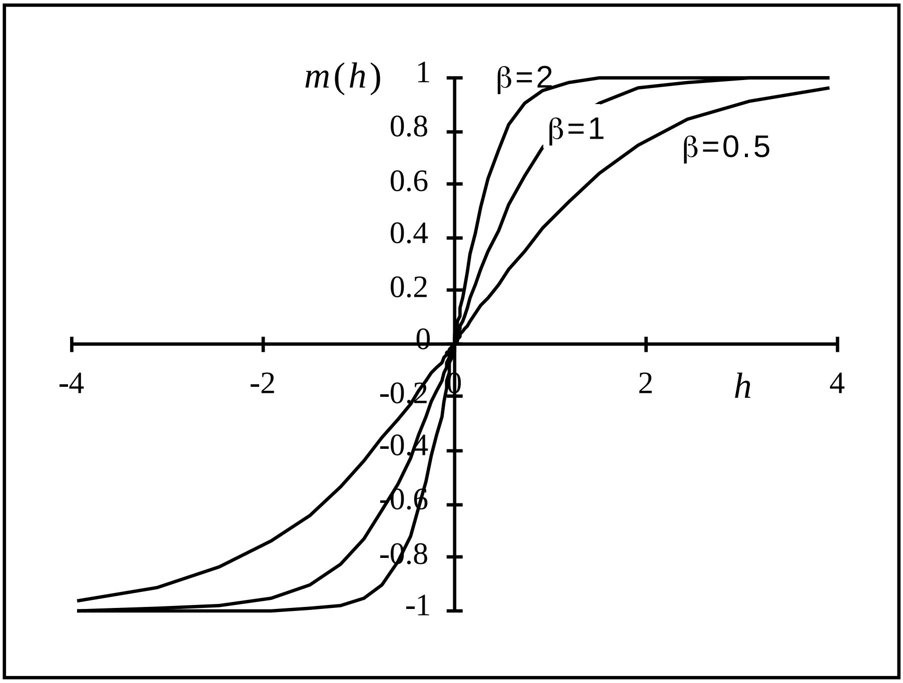


Figure 1.6.2 Plot of the magnetization at a particular site as a function of the magnetic field for independent spins in a magnetic field. The magnetization is the average of the spin value, so the magnetization shows the degree to which the spin is aligned to the magnetic field. The different curves are for several temperatures β = 0.5,1,2 ($\beta = 1/kT$). The magnetization has the same sign as the magnetic field. The magnitude of the spin increases with increasing magnetic field. Increasing temperature, however, decreases the alignment due to increased random motion of the spins. The maximum magnitude of the magnetization is 1, corresponding to a fully aligned spin. ∎

which is not particularly enlightening. However, we can rewrite this in terms of the magnetization using the identity:

$$\cosh(x) = \frac{1}{\sqrt{1-\tanh^2(x)}} \tag{1.6.13}$$

and the inverse of Eq. (1.6.10):

$$\beta h_i = \frac{1}{2}\ln\left(\frac{1+m_i}{1-m_i}\right) \tag{1.6.14}$$

Substituting into Eq. (1.6.12) gives

$$S = -k\sum_i m_i \frac{1}{2}\ln\left(\frac{1+m_i}{1-m_i}\right) + kN\ln(2) - k\frac{1}{2}\sum_i \ln\left(1-m_i{}^2\right) \tag{1.6.15}$$

Rearranging slightly, we have:

$$S = +k\left[N\ln(2) - \frac{1}{2}\sum_i \Big((1+m_i)\ln\big(1+m_i\big) + (1-m_i)\ln\big(1-m_i\big)\Big)\right] \tag{1.6.16}$$

The final expression can be derived, at least for the case when all m_i are the same, by counting the number of states directly. It is worth deriving the entropy twice, because it may be used more generally than this treatment indicates. We will assume that all $h_i = h$ are the same. The energy then depends only on the total magnetization:

$$\begin{aligned} E[\{s_i\}] &= -h\sum_i s_i \\ U &= -h\sum_i m_i = -hNm \end{aligned} \tag{1.6.17}$$

To obtain the entropy from the counting of states (Eq.(1.3.25)) we evaluate the number of states within a particular narrow energy range. Since the energy is the sum over the values of the spins, it may also be written as the difference between the number of UP spins $N(1)$ and DOWN spins $N(-1)$:

$$E[\{s_i\}] = -h(N(1) - N(-1)) \tag{1.6.18}$$

Thus, to find the entropy for a particular energy we must count how many states there are with a particular number of UP and DOWN spins. Moreover, flipping a spin from DOWN to UP causes a fixed increment in the energy. Thus there is no need to include in the counting the width of the energy interval in which we are counting states. The number of states with $N(1)$ UP spins and $N(-1)$ DOWN spins is:

$$\Omega(E,N) = \binom{N}{N(1)} = \frac{N!}{N(1)!N(-1)!} \tag{1.6.19}$$

The entropy can be written using Sterling's approximation (Eq. (1.2.27)), neglecting terms that are less than of order N, as:

$$\begin{aligned} S = k\ln(\Omega\,(E,N)) &= k[N(\ln N - 1) - N(1)(\ln N(1) - 1) - N(-1)(\ln N(-1) - 1] \\ &= k[N\ln N - N(1)\ln N(1) - N(-1)\ln N(-1)] \end{aligned} \quad (1.6.20)$$

the latter following from $N = N(1) + N(-1)$. To simplify this expression further, we write it in terms of the magnetization. Using $P_{s_i}(-1) + P_{s_i}(1) = 1$ and Eq. (1.6.9) for the magnetization we have the probability that a particular spin is UP and DOWN in terms of the magnetization as:

$$\begin{aligned} P_{s_i}(1) &= (1 + m) / 2 \\ P_{s_i}(-1) &= (1 - m) / 2 \end{aligned} \quad (1.6.21)$$

Since there are many spins in the system, we can obtain the number of UP spins using

$$\begin{aligned} N(1) &= NP_{s_i}(1) = N(1 + m) / 2 \\ N(-1) &= NP_{s_i}(1) = N(1 - m) / 2 \end{aligned} \quad (1.6.22)$$

Using these expressions, Eq.(1.6.20) becomes the same as Eq.(1.6.16), with $h_i = h$.

There is an important difference between the two derivations, in that the second assumed that all of the magnetic fields were the same. Thus, the first derivation appears more general. However, since the original system has no interactions, we could consider each of the spins with its own field h_i as a separate system. If we want to calculate the entropy of the individual spin, we would consider an ensemble of such spins. The ensemble consists of many spins with the same field $h = h_i$. The derivation of the entropy using the ensemble would be identical to the derivation we have just given, except that at the end we would divide by the number of different systems in the ensemble N. Adding together the entropies of different spins would then give exactly Eq. (1.6.16).

The entropy of a spin from Eq. (1.6.16) is maximal for a magnetization of zero when it has the value $k\ln(2)$. From the original definition of the entropy, this corresponds to the case when there are exactly two different possible states of the system. It thus corresponds to the case where the probability of each state $s = \pm 1$ is 1/2. The minimal entropy is for either $m = 1$ or $m = -1$—when there is only one possible state of the spin, so the entropy must be zero. ▮

1.6.2 *The Ising model*

We now add the essential aspect of the Ising model—interactions between the spins. The location of the spins in space was unimportant in the case of the noninteracting model. However, for the interacting model, we consider the spins to be located on a periodic lattice in space. Similar to the CA models of Section 1.5, we allow the spins to interact only with their nearest neighbors. It is conventional to interpret neighbors strictly as the spins with the shortest Euclidean distance from a particular site. This means that for a cubic lattice there are two, four and six neighbors in one, two and three dimensions respectively. We will assume that the interaction with each of the neighbors is the same and we write the energy as:

$$E[\{s_i\}] = -\sum_i h_i s_i - J \sum_{<ij>} s_i s_j \tag{1.6.23}$$

The notation <*ij*> under the summation indicates that the sum is to be performed over all *i* and *j* that are nearest neighbors. For example,in one dimension this could be written as:

$$E[\{s_i\}] = -\sum_i h_i s_i - J \sum_i s_i s_{i+1} \tag{1.6.24}$$

If we wanted to emphasize that each spin interacts with its two neighbors, we could write this as

$$E[\{s_i\}] = -\sum_i h_i s_i - J \frac{1}{2} \sum_i (s_i s_{i+1} + s_i s_{i-1}) \tag{1.6.25}$$

where the factor of 1/2 corrects for the double counting of the interaction between every two neighboring spins. In two and three dimensions (2-d and 3-d), there is need of additional indices to represent the spatial dependence. We could write the energy in 2-d as:

$$E[\{s_{i,j}\}] = -\sum_{i,j} h_{i,j} s_{i,j} - J \sum_{i,j} (s_{i,j} s_{i+1,j} + s_{i,j} s_{i,j+1}) \tag{1.6.26}$$

and in 3-d as:

$$E[\{s_{i,j,k}\}] = -\sum_{i,j,k} h_{i,j,k} s_{i,j,k} - J \sum_{i,j,k} (s_{i,j,k} s_{i+1,j,k} + s_{i,j,k} s_{i,j+1,k} + s_{i,j,k} s_{i,j,k+1}) \tag{1.6.27}$$

In these sums,each nearest neighbor pair appears only once. We will be able to hide the additional indices in 2-d and 3-d by using the nearest neighbor notation <*ij*> as in Eq. (1.6.23).

The interaction *J* between spins may arise from many different sources. Similar to the derivation of h_i in Eq.(1.6.2), this is the only form that an interaction between two spins can take (Question 1.6.2). There are two distinct possibilities for the behavior of the system depending on the sign of the interaction. Either the interaction tries to orient the spins in the same direction ($J > 0$) or in the opposite direction ($J < 0$). The former is called a ferromagnet and is the common form of a magnet. The other is called an antiferromagnet (Section 1.6.4) and has very different external properties but can be represented by the same model, with *J* having the opposite sign.

Question 1.6.2 Show that the form of the interaction given in Eq. (1.6.24) Jss' is the most general interaction between two spins.

Solution 1.6.2 We write as a general form of the energy of two spins:

$$\begin{aligned} e(s,s') = {} & e(1,1)\frac{(1+s)(1+s')}{4} + e(1,-1)\frac{(1+s)(1-s')}{4} \\ & + e(1,-1)\frac{(1-s)(1+s')}{4} + e(-1,-1)\frac{(1-s)(1-s')}{4} \end{aligned} \qquad (1.6.28)$$

If we expand this we will find a constant term, terms that are linear in s and s' and a term that is proportional to ss'. The linear terms give rise to the local field h_i, and the final term is the interaction. There are other possible interactions that could be written that would include three or more spins. ▮

In a magnetic system, each microscopic spin is itself the source of a small magnetic field. Magnets have the property that they can be the source of a macroscopic magnetic field. When a material is a source of a magnetic field, we say that it is magnetized. The magnetic field arises from constructive superposition of the microscopic sources of the magnetic field that we represent as spins. In effect,the small spins combine together to form a large spin. We have seen in Section 1.6.1 that when there is a magnetic field h_i, each spin will orient itself with the magnetic field. This means that in an external field—a field due to a source outside of the magnet—there will be a macroscopic orientation of the spins and they will in turn give rise to a magnetic field. Magnets,however, can be the source of a magnetic field even when there is no external field. This occurs only below a particular temperature known as the Curie temperature of the material. At higher temperatures,a magnetization exists only in an external magnetic field. The Ising model captures this behavior by showing that the interactions between the spins can cause a spontaneous orientation of the spins without any external field. The spontaneous magnetization is a collective phenomenon. It would not exist for an isolated spin or even for a small collection of interacting spins.

Ultimately, the reason that the spontaneous magnetization is a collective phenomenon has more to do with the kinetics than the thermodynamics of the system. The spontaneous magnetization must occur in a particular direction. Without an external field,there is no reason for any particular direction, but the system must choose one. In our case,it must choose between one of two possibilities—UP or DOWN. Once the magnetization occurs,it breaks a symmetry of the system, because we can now tell the difference between UP and DOWN on the macroscopic scale. At this point,the kinetics of the system must reenter. If the system were able to flip between UP and DOWN very rapidly, we would not be able to measure either case. However, we know that if all of the spins have to flip at once, the likelihood of this happening becomes vanishingly small as the number of spins grows. Thus for a large number of spins in a macroscopic material, this flipping becomes slower than our observation of the magnet.On the other hand,if we had only a few spins,they would still flip back and forth. It is this property of the system that makes the spontaneous magnetization a collective phenomenon.

Returning briefly to the discussion at the end of Section 1.3, we see that by choosing a direction for the magnetization,the magnet breaks the ergodic theorem. It is no longer possible to represent the system using an ensemble with all possible states of the system. We must exclude half of the states that have the opposite magnetization. The reason, as we described there, is because of the existence of a slow process, or a long time scale, that prevents the system from going from one choice of magnetization to the other.

The existence of a spontaneous magnetization arises because of the energy lowering of the system when neighboring spins align with each other. At sufficiently low temperatures, this causes the system to align collectively one way or another. Above the Curie temperature, T_c, the energy gain by alignment is destroyed by the temperature-induced random flipping of individual spins. We say that the higher temperature phase is a disordered phase, as compared to the ordered low temperature phase, where all spins are aligned. When we think about this thermodynamically, the disorder is an effect of optimizing the entropy, which promotes the disordered state and competes with the energy as the temperature is increased.

1.6.3 *Mean field theory*

Despite the simplicity of the Ising model, it has never been solved exactly except in one dimension, and in two dimensions for $h_i = 0$. The techniques that are useful in these cases do not generalize well. We will emphasize instead a powerful approximation technique for describing systems of many interacting parts known as the mean field approximation. The idea of this approximation is to treat a single element of the system under the average influence of the rest of the system. The key to doing this correctly is to recognize that this average must be performed self-consistently. The meaning of self-consistency will be described shortly. The mean field approximation cannot be applied to all interacting systems. However, when it can be, it enables the system to be understood in a direct way.

To use the mean field approximation we single out a particular spin s_i and find the effective field (or mean field) it experiences h_i'. This field is obtained by replacing all variables in the energy by their average values, except for s_i. This leads to an effective energy $E_{MF}(s_i)$ for s_i. To obtain it we can neglect all terms in the energy (Eq. (1.6.23)) that do not include s_i.

$$\begin{aligned} E_{MF}(s_i) &= -h_i s_i - J\sum_{jnn} s_i < s_j > = -h_i' s_i \\ h_i' &= h_i + J\sum_{jnn} < s_j > \end{aligned} \tag{1.6.29}$$

The sum is over all nearest neighbors of s_i. If we are able to find what the mean field h_i' is, then we can solve this interacting Ising model using the solution of the Ising model without interactions. The problem is that in order to find the field we have to know the average value of the spins, which in turn depends on the effective fields. This is the self-consistency. We will develop a single algebraic equation for the solution. It is interesting first to consider this problem when the external fields h_i are zero. Eq. (1.6.29) shows that a mean field might still exist. When the external field is zero, each of the spin variables has the same equation. We might guess that the average value of the spin in one location will be the same as that in any other location:

$$m = m_i = < s_i > \tag{1.6.30}$$

In this case our equations become

$$\begin{aligned} E_{MF}(s_i) &= -h_i' s_i \\ h_i' &= J \sum_{jnn} m = zJm \end{aligned} \tag{1.6.31}$$

where z is the number of nearest neighbors,known as the coordination number of the system. Eq.(1.6.10) gives us the value of the average magnetization when the spin is subject to a field. Using this same expression under the influence of the mean field we have

$$m = \tanh(\beta h_i') = \tanh(\beta zJm) \tag{1.6.32}$$

This is the self-consistent equation, which gives the value of the magnetization in terms of itself. The solution of this equation may be found graphically, as illustrated in Fig. 1.6.3, by plotting the functions $y = m$ and $y = \tanh(\beta zJm)$ and finding their intersections. There is always a solution $m = 0$. In addition, for values of $\beta zJ > 1$, there are two more solutions related by a change of sign $m = \pm m_0(\beta zJ)$, where we name the positive solution $m_0(\beta zJ)$. When $\beta zJ = 1$, the line $y = m$ is tangent to the plot of $y = \tanh(\beta zJm)$ at $m = 0$. For values $\beta zJ > 1$,the value of $y = \tanh(\beta zJm)$ must rise above the line $y = m$ for small positive m and then cross it. The crossing point is the solution $m_0(\beta zJ)$. $m_0(\beta zJ)$ approaches one asymptotically as $\beta zJ \rightarrow \infty$, e. g. as the temperature goes to zero. A plot of $m_0(\beta zJ)$ from a numerical solution of Eq. (1.6.32) is shown in Fig. 1.6.4.

We see that there are two different regimes for this model with a transition at a temperature T_c given by $\beta zJ = 1$ or

$$kT_c = zJ \tag{1.6.33}$$

To understand what is happening it is helpful to look at the energy $U(m)$ and the free energy $F(m)$ as a function of the magnetization,assuming that all spins have the same magnetization. We will treat the magnetization as a parameter that can be varied. The actual magnetization is determined by minimizing the free energy.

To determine the energy, we must average Eq.(1.6.23), which includes a product of spins on neighboring sites. The mean field approximation treats each spin as if it were independent of other spins except for their average field. This implies that we have neglected correlations between the value of one spin and the others around it. Assuming that the spins are uncorrelated means the average over the product over two spins may be approximated by the product over the averages:

$$<s_i s_j> \approx <s_i><s_j> = m^2 \tag{1.6.34}$$

The average over the energy without any external fields is then:

$$U(m) = < -J \sum_{<ij>} s_i s_j > = -\frac{1}{2} NJzm^2 \tag{1.6.35}$$

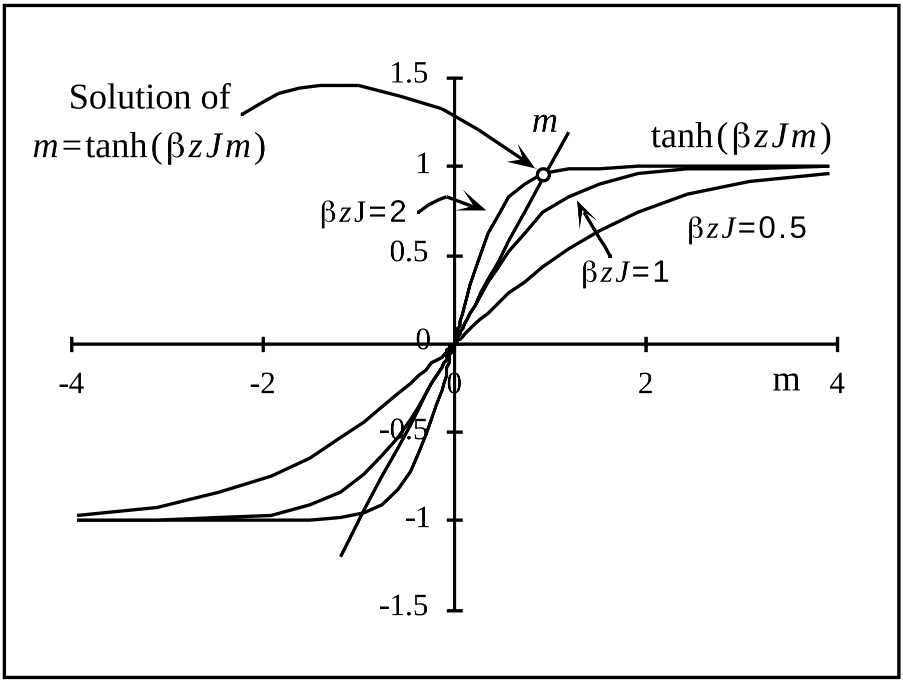


Figure 1.6.3 Graphical solution of Eq. (1.6.32) $m = \tanh(\beta zJm)$ by plotting both the left- and right-hand sides of the equation as a function of m and looking for the intersections. $m = 0$ is always a solution. To consider other possible solutions we note that both functions are antisymmetric in m so we need only consider positive values of m. For every positive solution there is a negative solution of equal magnitude. When $\beta zJ = 1$ the slope of both sides of the equation is the same at $m = 0$. For $\beta zJ > 1$ the slope of the right is greater than the left side. For large positive values of m the right side of the equation is always less than the left side. Thus for $\beta zJ > 1$, there must be an additional solution. The solution is plotted in Fig. 1.6.4. ▮

The factor of 1/2 arises because we count each interaction only once (see Eqs. (1.6.24)–(1.6.27)). A sum over the average of $E_{MF}(s_i)$ would give twice as much, due to counting each of the interactions twice.

Since we have fixed the magnetization of all spins to be the same, we can use the entropy we found in Question 1.6.1 to obtain the free energy as:

$$F(m) = -\frac{1}{2}NJzm^2 - NkT\left[\ln(2) - \frac{1}{2}\left((1+m)\ln\left(1+m\right) + (1-m)\ln\left(1-m\right)\right)\right] \quad (1.6.36)$$

This free energy is plotted in Fig. 1.6.5 as a function of m/Jz for various values of kT/Jz. We see that the behavior of this system is precisely the behavior of a second-order phase transition described in Section 1.3. Above the transition temperature T_c there is only one possible phase and below T_c there are two phases of equal en-

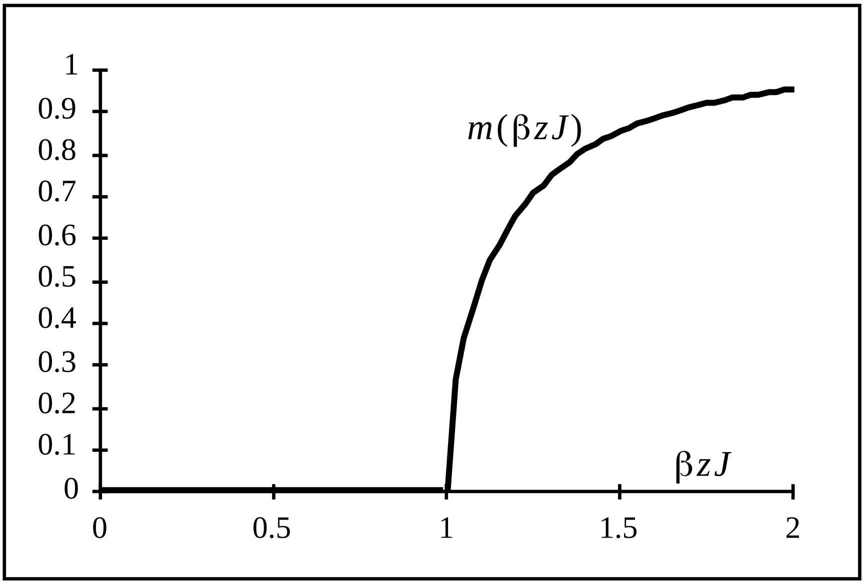


Figure 1.6.4 The mean field approximation solution of the Ising model gives the magnetization (average value of the spin) as a solution of Eq. (1.6.32). The solution is shown as a function of βzJ. As discussed in Fig. 1.6.3 and the text for $\beta zJ > 1$ there are three solutions. Only the positive one is shown. The solution $m = 0$ is unstable, as can be seen by analysis of the free energy shown in Fig. 1.6.5. The other solution is the negative of that shown. ❚

ergy. Question 1.6.3 clarifies a technical point in this derivation, and Question 1.6.4 generalizes the solution to include nonzero magnetic fields $h_i \neq 0$.

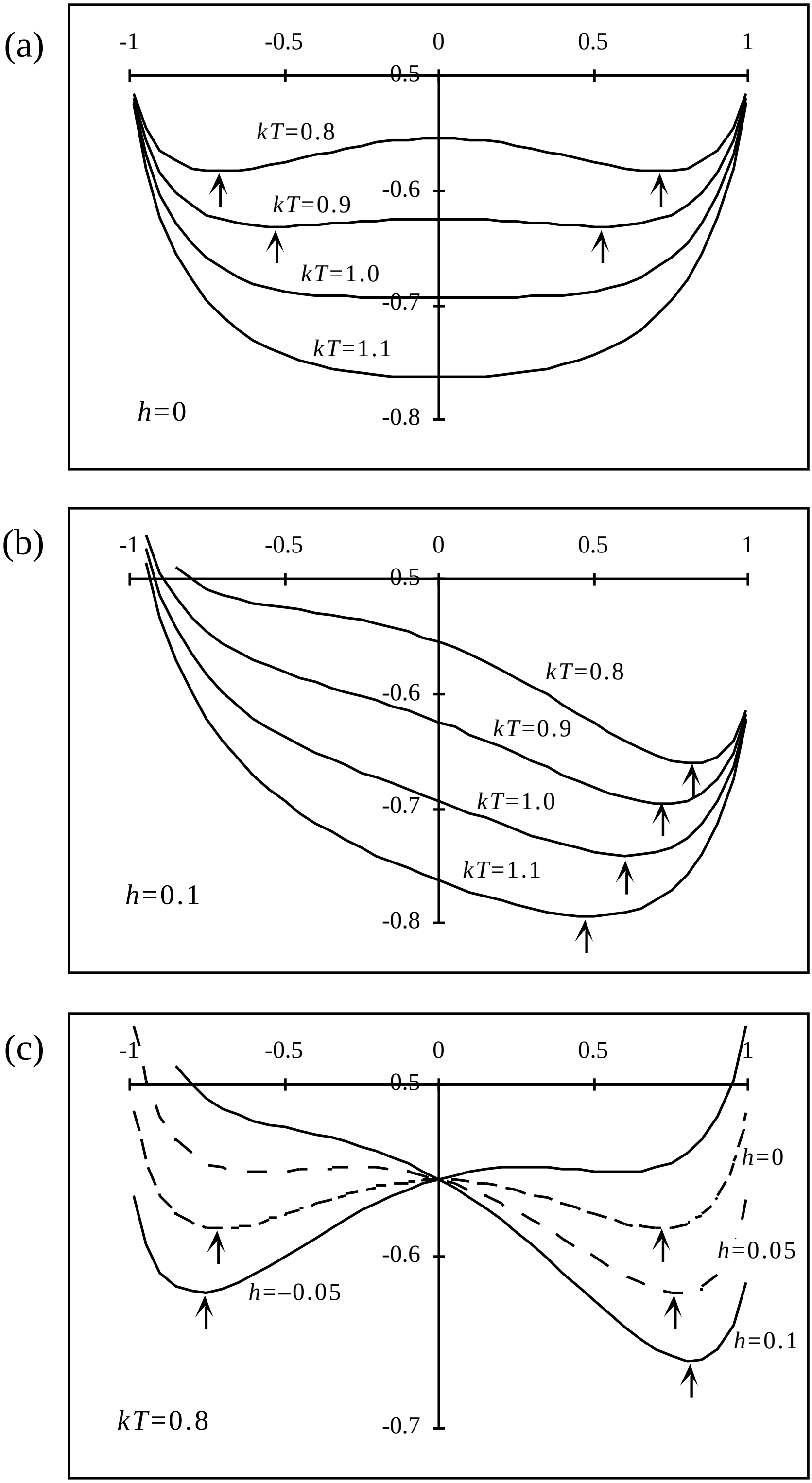
Question 1.6.3 Show that the minima of the free energy are the solutions of Eq.(1.6.32). This shows that our derivation is internally consistent. Specifically, that our two ways of defining the mean field approximation, first using Eq. (1.6.29) and then using Eq. (1.6.34), are compatible.

Solution 1.6.3 Taking the derivative of Eq. (1.6.35) with respect to m and setting it to zero gives:

$$0 = -Jzm - kT\left[-\frac{1}{2}\left(\ln\left(1+m\right) - \ln\left(1-m\right)\right)\right] \tag{1.6.37}$$

Recognizing the inverse of tanh,as in Eq.(1.6.14), gives back Eq.(1.6.32) as desired. ❚

Question 1.6.4 Find the replacements for Eq. (1.6.31)–(1.6.36) for the case where there is a uniform external magnetic field $h_i = h$. Plot the free energy for a few cases.


(a)
-1
-0.5
0
0.5
1
-0.5
-0.6
-0.7
-0.8
kT=0.8
kT=0.9
kT=1.0
kT=1.1
h=0
(b)
-1
-0.5
0
0.5
1
-0.5
-0.6
-0.7
-0.8
kT=0.8
kT=0.9
kT=1.0
kT=1.1
h=0.1
(c)
-1
-0.5
0
0.5
1
-0.5
-0.6
-0.7
h=0
h=0.05
h=0.1
h=−0.05
kT=0.8

Solution 1.6.4 Applying an external magnetic field breaks the symmetry between the two different minima in the energy that we have found. In this case we have instead of Eq. (1.6.29)

$$\begin{aligned} E_{MF}(s_i) &= -h_i' s_i \\ h_i' &= h + zJm \end{aligned} \tag{1.6.38}$$

The self-consistent equation instead of Eq. (1.6.32) is:

$$m = \tanh(\beta h + \beta zJm) \tag{1.6.39}$$

Averaging over the energy gives:

$$U(m) = < -h\sum_i s_i - J\sum_{<ij>} s_i s_j > = -Nhm - \frac{1}{2}NJzm^2 \tag{1.6.40}$$

The entropy is unchanged, so the free energy becomes:

$$F(m) = -Nhm - \frac{1}{2}NJzm^2 - NkT\left[\ln(2) - \frac{1}{2}\left((1+m)\ln\left(1+m\right) + (1-m)\ln\left(1-m\right)\right)\right] \tag{1.6.41}$$

Several plots are shown in Fig. 1.6.5. Above kT_c of Eq. (1.6.33) the application of an external magnetic field gives rise to a magnetization by shifting the location of the single minimum. Below this temperature there is a tilting of the two minima. Thus, going from a positive to a negative value of h would give an abrupt transition—a first-order transition which occurs at exactly $h = 0$. ▮

In discussing the mean field equations, we have assumed that we could specify the magnetization as a parameter to be optimized. However, the prescription we have from thermodynamics is that we should take all possible states of the system with a Boltzmann probability. What is the justification for limiting ourselves to only one value of the magnetization? We can argue that in a macroscopic system, the optimal

Figure 1.6.5 Plots of the mean field approximation to the free energy. (a) shows the free energy for $h = 0$ as a function of m for various values of kT. The free energy m and kT are measured in units of Jz. As the temperature is lowered below $kT/zJ = 1$ there are two minima instead of one (shown by arrows). These minima are the solutions of Eq. (1.6.32) (see Question 1.6.3). The solutions are illustrated in Fig. 1.6.4. (b) Shows the same curves as (a) but with a magnetic field $h/zJ = 0.1$. The location of the minimum gives the value of the magnetization. The magnetic field causes a magnetization to exist at all temperatures, but it is larger at lower temperatures. At the lowest temperature shown $kT/zJ = 0.8$ the effect of the phase transition can be seen in the beginnings of a second (metastable) minimum at negative values of the magnetization. (c) shows plots at a fixed temperature of $kT/zJ = 0.8$ for different values of the magnetic field. As the value of the field goes from positive to negative, the minimum of the free energy switches from positive to negative values discontinuously. At exactly $h = 0$ there is a discontinuous jump from positive to negative magnetization—a first-order phase transition. ▮

value of the magnetization will so dominate other magnetizations that any other possibility is negligible. This is reasonable except for the case when the magnetic field is close to zero, below T_c, and we have two equally likely magnetizations. In this case, the usual justification does not hold, though it is often implicitly applied. A more complete justification requires a discussion of kinetics given in Section 1.6.6.

Using the results of Question 1.6.4, we can draw a phase diagram like that illustrated in Section 1.3 for water (Fig. 1.3.7). The phase diagram of the Ising model (Fig. 1.6.6) describes the transitions as a function of temperature (or β) and magnetic field h. It is very simple for the case of the magnetic system, since the first-order phase transition line lies along the $h = 0$ axis and ends at the second-order transition point given by Eq. (1.6.33).

1.6.4 *Antiferromagnets*

We found the existence of a phase transition in the last section from the self-consistent mean field result (Eq. (1.6.32)), which showed that there was a nonzero magnetization for $\beta zJ > 1$. This condition is satisfied for small enough temperature as long as $J > 0$. What about the case of $J < 0$? There are no additional solutions of Eq. (1.6.32) for this case. Does this mean there is no phase transition? Actually, it means that one of our assumptions is not a good one. When $J < 0$, each spin would like (has a lower energy if...) its neighbors to antialign rather than align their spins. However, we have assumed that all spins have the same magnetization, Eq. (1.6.30). The self-consistent equation assumes and does not guarantee that all spins have the same magnetization. This assumption is not a good one when the spins are trying to antialign.

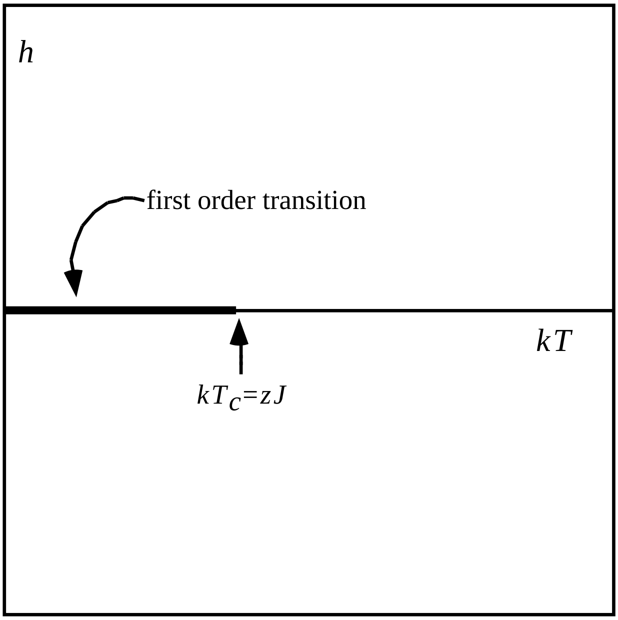


Figure 1.6.6 The phase diagram of the Ising model found from the mean field approximation. The line of first-order phase transitions at $h = 0$ ends at the second-order phase transition point given by Eq. (1.6.32). For positive values of h there is a net positive magnetization and for negative values there is a negative magnetization. The change through $h = 0$ is continuous above the second-order transition point, and discontinuous below it. ∎

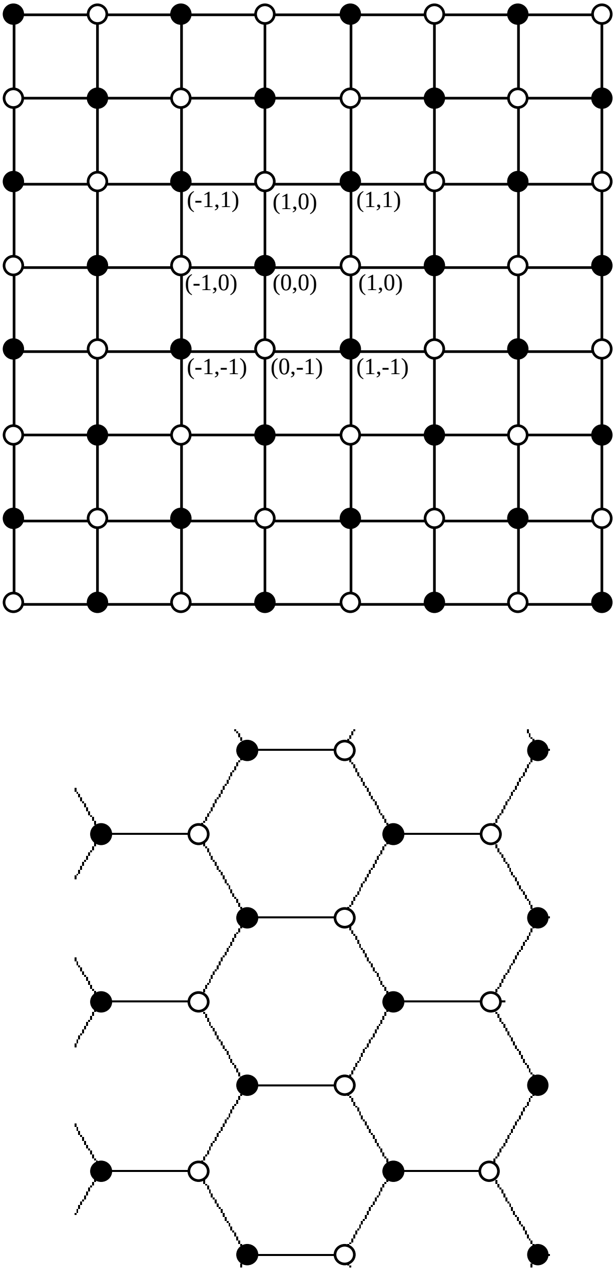


Figure 1.6.7 In order to obtain mean field equations for the antiferromagnetic case $J < 0$ we consider a square lattice (top) and label every site according to the sum of its rectilinear indices as odd (open circles) or even (filled circles). A few sites are shown with indices. Each site is understood to be the location of a spin. We then invert the spins (redefine them by $s \to -s$) that are on odd sites and find that the new system satisfies the same equations as the ferromagnet. The same trick works for any bipartite lattice; for example the hexagonal lattice shown (bottom). By using this trick we learn that at low temperatures the system will have a spontaneous magnetism that is positive on odd sites and negative on even sites or the opposite. ▮

We can solve the case of a system with $J < 0$ on a square or cubic lattice directly using a trick. We label every spin by indices (i,j) in 2-d, as indicated in Fig. 1.6.7, or (i,j,k) in 3-d. Then we consider separately the spins whose indices sum to an odd number ("odd spins") and those whose indices sum to an even number ("even spins"). Note that all the neighbors of an odd spin are even and all neighbors of an even spin are odd. Now we invert all of the odd spins. Explicitly we define new spin variables in 3-d as

$$s'_{ijk} = (-1)^{i+j+k} s_{ijk} \tag{1.6.42}$$

In terms of these new spins,the energy without an external magnetic field is the same as before, except that each term in the sum has a single additional factor of (–1). There is only one factor of (−1) because every nearest neighbor pair has one odd and one even spin. Thus:

$$E[\{s'_i\}] = -J \sum_{<ij>} s_i s_j = -(-J) \sum_{<ij>} s'_i s'_j = -J' \sum_{<ij>} s'_i s'_j \tag{1.6.43}$$

We have completed the transformation by defining a new interaction $J' = -J > 0$. In terms of the new variables, we are back to the ferromagnet. The solution is the same, and below the temperature given by $kT_c = zJ'$ there will be a spontaneous magnetization of the new spin variables. What happens in terms of the original variables? They become antialigned. All of the even spins have magnetization in one direction, UP, and the odd spins have magnetization in the opposite direction, DOWN, or vice versa. This lowers the energy of the system, because the negative interaction $J < 0$ means that all of the neighboring spins want to antialign. This is called an antiferromagnet.

The trick we have used to solve the antiferromagnet works for certain kinds of periodic arrangements of spins called bipartite lattices. A bipartite lattice can be divided into two lattices so that all the nearest neighbors of a member of one lattice are members of the other lattice. This is exactly what we need in order for our redefinition of the spin variables to work. Many lattices are bipartite,including the cubic lattice and the hexagonal honeycomb lattice illustrated in Fig. 1.6.7. However, the triangular lattice, illustrated in Fig. 1.6.8, is not.

The triangular lattice exemplifies an important concept in interacting systems known as frustration. Consider what happens when we try to assign magnetizations to each of the spins on a triangular lattice in an effort to create a configuration with a lower energy than a disordered system. We start at a position marked (1) on Fig. 1.6.8 and assign it a magnetization of m. Then, since it wants its neighbors to be antialigned, we assign position (2) a magnetization of $-m$. What do we do with the spin at (3)? It has interactions both with the spin at (1) and with the spin at (2). These interactions would have it be antiparallel with both—an impossible task. We say that the spin at (3) is frustrated,since it cannot simultaneously satisfy the conflicting demands upon it. It should not come as a surprise that the phenomenon of frustration becomes a commonplace occurrence in more complex systems. We might even say that frustration is a source of complexity.

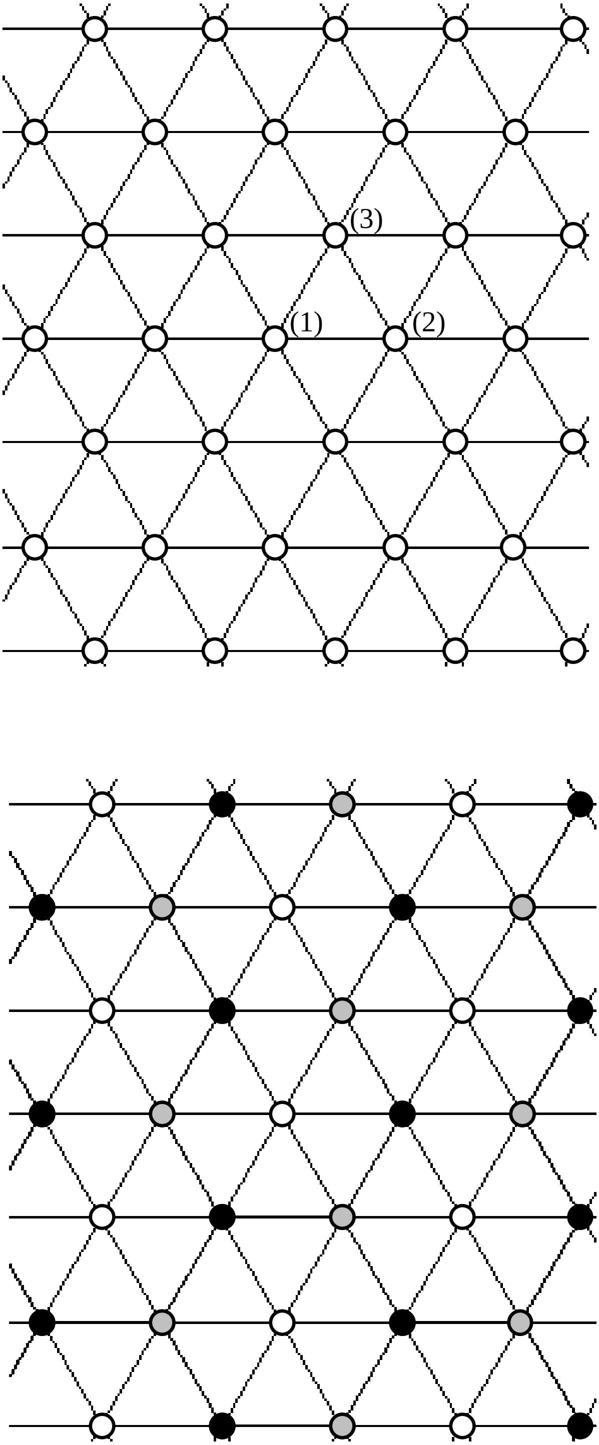


Figure 1.6.8 A triangular lattice (top) is not a bipartite lattice. In this case we cannot solve the antiferromagnet J < 0 by the same method as used for the square lattice (see Fig. 1.6.7). If we try to assign magnetizations to different sites we find that assigning a magnetization to site (1) would lead site (2) to be antialigned. This combination would, however require site (3) to be antialigned to both sites (1) and (2), which is impossible. We say that site (3) is “frustrated.” The bottom illustration shows what happens when we take the hexagonal lattice from Fig. 1.6.7 and superpose the magnetizations on the triangular lattice leaving the additional sites (shaded) as unmagnetized (see Questions 1.6.5–1.6.7). ▮

Question 1.6.5 Despite the existence of frustration, it is possible to construct a state with lower energy than a completely disordered state on the triangular lattice. Construct one of them and evaluate its free energy.

Solution 1.6.5 We construct the state by extending the process discussed in the text for assigning magnetizations to individual sites. We start by assigning a magnetization m to site (1) in Fig. 1.6.8 and $-m$ to site (2). Because site (3) is frustrated, we assign it no magnetization. We continue by assigning magnetizations to any site that already has two neighbors that are assigned magnetizations. We assign a magnetization of m when the neighbors are $-m$ and 0, a magnetization of $-m$ when the neighbors are m and 0 and a magnetization of 0 when the neighbors are m and $-m$. This gives the illustration at the bottom of Fig. 1.6.8. Comparing with Fig. 1.6.7, we see that the magnetized sites correspond to the honeycomb lattice. One-third of the triangular lattice sites have a magnetization of $+m$, $-m$ and 0. Each magnetized site has three neighbors of the opposite magnetization and three unmagnetized sites. The free energy of this state is given by:

$$\begin{aligned} F(m) = NJm^2 - \frac{1}{3}NkT\ln(2) \\ -\frac{2}{3}NkT\left[\ln(2) - \frac{1}{2}\left((1+m)\ln\left(1+m\right) + (1-m)\ln\left(1-m\right)\right)\right] \end{aligned} \qquad (1.6.44)$$

The first term is the energy. Each nearest neighbor pair of spins that are antialigned provides an energy Jm^2. Let us call this a bond between two spins. There are a total of three interactions for every spin (each spin interacts with six other spins but we can count each interaction only once). However, on average there is only one out of three interactions that is a bond in this system. To count the bonds, note that one out of three spins (with $m_i = 0$) has no bonds, while the other two out of three spins each have three bonds. This gives a total of six bonds for three sites, but each bond must be counted only once for a pair of interacting spins. We divide by two to get three bonds for three spins, or an average of one bond per site. The second term in Eq. (1.6.44) is the entropy of the $N / 3$ unmagnetized sites,and the third term is the entropy of the $2N / 3$ magnetized sites.

There is another way to systematically construct a state with an energy lower than a completely disordered state. Assign magnetizations $+m$ and $-m$ alternately along one straight line—a one-dimensional antiferromagnet. Then skip both neighboring lines by setting all of their magnetizations to zero. Then repeat the antiferromagnetic line on the next parallel line. This configuration of alternating antiferromagnetic lines is also lower in energy than the disordered state, but it is higher in energy than the configuration shown in Fig. 1.6.8 at low enough temperatures, as discussed in the next question. ∎

Question 1.6.6 Show that the state illustrated on the bottom of Fig. 1.6.8 has the lowest possible free energy as the temperature goes to zero, at least in the mean field approximation.

Solution 1.6.6 As the temperature goes to zero, the entropic contribution to the free energy is irrelevant. The energy of the Ising model is minimized in the mean field approximation when the magnetization is +1 if the local effective field is positive, or –1 ifit is negative. The magnetization is arbitrary if the effective field is zero. If we consider three spins arranged in a triangle, the lowest possible energy of the three interactions between them is given by having one with m = +1, one with m = –1 and the other arbitrary. This is forced, because we must have at least one +1 and one –1 and then the other is arbitrary. This is the optimal energy for any triangle of interactions. The configuration of Fig. 1.6.8 achieves this optimal arrangement for all triangles and therefore must give the lowest possible energy of any state. ∎

Question 1.6.7 In the case of the ferromagnet and the antiferromagnet, we found that there were two different states of the system with the same energy at low temperatures. How many states are there of the kind shown in Fig. 1.6.8 and described in Questions 1.6.5 and 1.6.6?

Solution 1.6.7 There are two ways to count the states. The first is to count the number of distinct magnetization structures. This counting is as follows. Once we assign the values of the magnetization on a single triangle, we have determined them everywhere in the system. This follows by inspection or by induction on the size of the assigned triangle. Since we can assign arbitrarily the three different magnetizations $(m, -m, 0)$ within a triangle, there are a total of six such distinct magnetization structures.

We can also count how many distinct arrangements of spins there are. This is relevant at low temperatures when we want to know the possible states at the lowest energy. We see that there are $2^{N/3}$ arrangements of the arbitrary spins for each of the magnetizations. If we want to count all of the states, we can almost multiply this number by 6. We have to correct this slightly because of states where the arbitrary spins are all aligned UP or DOWN. There are two of these for each arrangement of the magnetizations, and these will be counted twice. Making this correction gives $6(2^{N/3} - 1)$ states. We see that frustration gives rise to a large number of lowest energy states.

We have not yet proven that these are the only states with the lowest energy. This follows from the requirement that every triangle must have its lowest possible energy, and the observation that setting the value of the magnetizations of one triangle then forces the values of all other magnetizations uniquely. ∎

Question 1.6.8 We discovered that our assumption that all spins should have the same magnetization does not always apply. How do we know that we found the lowest energy in the case of the ferromagnet? Answer this for the case of $h = 0$ and $T = 0$.

Solution 1.6.8 To minimize the energy, we can consider each term of the energy, which is just the product of spins on adjacent sites. The minimum possible value for each term of a ferromagnet occurs for aligned spins. The two states we found at $T = 0$ with $m_i = 1$ and $m_i = –1$ are the only possible states with all spins aligned. Since they give the minimum possible energy, they must be the correct states. ❚

1.6.5 *Beyond mean field theory (correlations)*

Mean field theory treats only the average orientation of each spin and assumes that spins are uncorrelated. This implies that when one spin changes its sign, the other spins do not respond. Since the spins are interacting, this must not be true in a more complete treatment. We expect that even above T_c, nearby spins align to each other. Below T_c, nearby spins should be more aligned than would be suggested by the average magnetization. Alignment of spins implies their values are correlated. How do we quantify the concept of correlation? When two spins are correlated they are more likely to have the same value. So we might define the correlation of two spins as the average of the product of spins:

$$< s_i s_j > = \sum_{s_i, s_j} s_i s_j P(s_i, s_j) = P_{s_i s_j}(1,1) + P_{s_i s_j}(-1,-1) - P_{s_i s_j}(-1,1) - P_{s_i s_j}(1,-1) \qquad (1.6.45)$$

According to this definition, they are correlated if they are both always +1, so that $P_{s_i s_j}(1,1) = 1$. Then $< s_i s_j >$ achieves its maximum possible value +1. The problem with this definition is that when s_i and s_j are both always +1 they are completely independent of each other, because each one is +1 independently of the other. Our concept of correlation is the opposite of independence. We know that if spins are independent, then their joint probability distribution factors (see Section 1.2)

$$P(s_i, s_j) = P(s_i)P(s_j) \qquad (1.6.46)$$

Thus we define the correlation as a measure of the departure of the joint probability from the product of the individual probabilities.

$$\sum_{s_i, s_j} s_i s_j (P(s_i, s_j) - P(s_i)P(s_j)) = < s_i s_j > - < s_i >< s_j > \qquad (1.6.47)$$

This definition means that when the correlation is zero, we can say that s_i and s_j are independent. However, we must be careful not to assume that they are not aligned with each other. Eq. (1.6.45) measures the spin alignment.

Question 1.6.9 One way to think about the difference between Eq. (1.6.45) and Eq. (1.6.47) is by considering a hierarchy of correlations. The first kind of correlation is of individual spins with themselves and is just the average of the spin. The second kind are correlations between pairs of spins that are not contained in the first kind. Define the next kind of correlation in the hierarchy that would describe correlations between three spins but exclude the correlations that appear in the first two.

Solution 1.6.9 The first three elements in the hierarchy of correlations are:

$$\begin{gathered}< s_i > \\ < s_i s_j > - < s_i > < s_j > \qquad (1.6.48) \\ < s_i s_j s_k > - < s_i s_j > < s_k > - < s_i s_k > < s_j > - < s_j s_k > < s_i > +2 < s_i > < s_j > < s_k >\end{gathered}$$

The expression for the correlation of three spins can be checked by seeing what happens if the variables are independent. When variables are independent, the average of their product is the same as the product of their averages. Then all averages become products of averages of single variables and everything cancels. Similarly, if the first two variables s_i and s_j are correlated and the last one s_k is independent of them,then the first two terms cancel and the last three terms also cancel. Thus, this expression measures the correlations of three variables that are not present in any two of them. ▮

Question 1.6.10 To see the difference between Eqs. (1.6.45) and (1.6.47), evaluate them for two cases: (*a*) s_i is always equal to 1 and s_j is always equal to –1,and (*b*) s_i is always the opposite of s_j but each of them averages to zero (i.e., is equally likely to be +1 or –1).

Solution 1.6.10

a. $P_{s_i s_j}(1,-1) = 1$, so $< s_i s_j > = -1$, but $< s_i s_j > - < s_i > < s_j > = 0$.

b. $< s_i s_j > = -1$, and $< s_i s_j > - < s_i > < s_j > = -1$. ▮

Comparing Eq. (1.6.34) with Eq. (1.6.47), we see that correlations measure the departure of the system from mean field theory. When there is an average magnetization, such as there is below T_c in a ferromagnet, the effect of the average magnetization is removed by our definition of the correlation. This can also be seen from rewriting the expression for correlations as:

$$< s_i s_j > - < s_i > < s_j > \; = < (s_i - < s_i >) \, (s_j - < s_j >) > \qquad (1.6.49)$$

Correlations measure the behavior of the difference between the spin and its average value. In the rest of this section we discuss qualitatively the correlations that are found in a ferromagnet and the breakdown of the mean field approximation.

The energy of a ferromagnet is determined by the alignment of neighboring spins. Positive correlations between neighboring spins reduce its energy. Positive or negative correlations diminish the possible configurations of spins and therefore reduce the entropy. At very high temperatures, the competition between the energy and the entropy is dominated by the entropy, so there should be no correlations and each spin is independent. At low temperatures, well below the transition temperature, the average value of the spins is close to one. For example, for $\beta zJ = 2$, which corresponds to $T = T_c / 2$, the value of $m_0(\beta zJ)$ is 0.96 (see Fig. 1.6.4). So the correlations given by Eq. (1.6.47) play almost no role. Correlations are most significant near T_c, so it is near the transition that the mean field approximation is least valid.

For all $T > T_c$ and for $h = 0$, the magnetization is zero. However, starting from high temperature, the correlation between neighboring spins increases as the temperature is lowered. Moreover, the correlation of one spin with its neighbors,and their correlation with their neighbors,induces a correlation of each spin with spins farther away. The distance over which spins are correlated increases as the temperature decreases. The correlation decays exponentially, so a correlation length $\xi(T)$ may be defined as the decay constant of the correlation:

$$< s_i s_j > - < s_i > < s_j > \quad e^{-r_{ij}/\xi(T)} \tag{1.6.50}$$

where r_{ij} is the Euclidean distance between s_i and s_j. At T_c the correlation length diverges. This is one way to think about how the phase transition occurs. The divergence of the correlation length implies that two spins anywhere in the system become correlated. As mentioned previously, in order for the instantaneous magnetization to be measured, there must also be a divergence of the relaxation time between opposite values of the magnetization. This will be discussed in Sections 1.6.6 and 1.6.7.

For temperatures just below T_c, the average magnetization is small. The correlation length of the spins is large. The average alignment (Eq. (1.6.45)) is essentially the same as the correlation (Eq. (1.6.47)). However, as T is further reduced below T_c, the average magnetization grows precipitously and the correlation measures the difference between the spin-spin alignment and the average spin value. Both the correlation and the correlation length decrease away from T_c. As the temperature goes to zero, the correlation length also goes to zero, even as the correlation itself vanishes.

At $T = T_c$ there is a special circumstance where the correlation length is infinite. This does not mean that the correlation is unchanged as a function of the distance between spins, r_{ij}. Since the magnetization is zero, the correlation is the same as the spin alignment. If the alignment did not decay with distance, the magnetization would be unity, which is not correct. The infinite correlation length corresponds to power law rather than exponential decay of the correlations. A power law decay of the correlations is more gradual than exponential and implies that there is no characteristic size for the correlations: we can find correlated regions of spins that are of any size. Since the correlated regions fluctuate, we say that there are fluctuations on every length scale.

The existence of correlations on every length scale near the phase transition and the breakdown of the mean field approximation that neglects these correlations played an important role in the development of the theory of phase transitions. The discrepancy between mean field predictions and experiment was one of the great unsolved problems of statistical physics. The development of renormalization techniques that directly consider the behavior of the system on different length scales solved this problem. This will be discussed in greater detail in Section 1.10.

In Section 1.3 we discussed the nature of ensemble averages and indicated that one of the central issues was determining the size of an independent system. For the Ising model and other systems that are spatially uniform, it is the correlation length that determines the size of an independent system. If a physical system is much larger than a correlation length then the system is self-averaging, in that experimental measurements average over many independent samples. We see that far from a phase transition, uniform systems are generally self-averaging; near a phase transition, the physical size of a system may enter in a more essential way.

The mean field approximation is sufficient to capture the collective behavior of the Ising model. However, even T_c is not given correctly by mean field theory, and indeed it is difficult to calculate. The actual transition temperature differs from the mean field value by a factor that depends on the dimensionality and structure of the lattice. In 1-d, the failure of mean field theory is most severe, since there is actually no real transition. Magnetization does not occur, except in the limit of $T \rightarrow 0$. The reason that there is no magnetization in 1-d, is that there is always a finite probability that at some point along the chain there will be a switch from having spins DOWN to having spins UP. This is true no matter how low the temperature is. The probability of such a boundary between UP and DOWN spins decreases exponentially with the temperature. It is given by $1/(1 + e^{2J/kT}) \approx e^{-2J/kT}$ at low temperature. Even one such boundary destroys the average magnetization for an arbitrarily large system. While formally there is no phase transition in one dimension, under some circumstances the exponentially growing distance between boundaries may have consequences like a phase transition. The effect is, however, much more gradual than the actual phase transitions in 2-d and 3-d.

The mean field approximation improves as the dimensionality increases. This is a consequence of the increase in the number of neighbors. As the number of neighbors increases, the averaging used for determining the mean field becomes more reliable as a measure of the environment of the spin. This is an important point that deserves some thought. As the number of different influences on a particular variable increases, they become better represented as an average influence. Thus in 3-d, the mean field approximation is better than in 2-d. Moreover, it turns out that rather than just gradually improving as the number of dimensions increases, for 4-d the mean field approximation becomes essentially exact for many of the properties of importance in phase transitions. This happens because correlations become irrelevant on long length scales in more than 4-d. The number of effective neighbors of a spin also increases if we increase the range of the interactions. Several different models with long-range interactions are discussed in the following section.

The Ising model has no built-in dynamics; however, we often discuss fluctuations in this model. The simplest fluctuation would be a single spin flipping in time. Unless the average value of a spin is +1 or –1, a spin must spend some time in each state. We can see that the presence of correlations implies that there must be fluctuations in time that affect more than one spin. This is easiest to see if we consider a system above the transition, where the average magnetization is zero. When one spin has the value +1, then the average magnetization of spins around it will be positive. On average, a region of spins will tend to flip together from one sign to the other. The amount of time that the region takes to flip depends on the length of the correlations. We have defined correlations in space between two spins. We could generalize the definition in Eq. (1.6.47) to allow the indices i and j to refer to different times as well as spatial positions. This would tell us about the fluctuations over time in the system. The analog of the correlation length Eq.(1.6.50) would be the relaxation time (Eq.(1.6.69) below).

The Ising model is useful for describing a large variety of systems;however, there are many other statistical models using more complex variables and interactions that have been used to represent various physical systems. In general, these models are treated first using the mean field approximation. For each model,there is a lower dimension (the lower critical dimension) below which the mean field results are completely invalid. There is also an upper critical dimension, where mean field is exact. These dimensions are not necessarily the same as for the Ising model.

1.6.6 *Long-range interactions and the spin glass*

Long-range interactions enable the Ising model to serve as a model of systems that are much more complex than might be expected from the magnetic analog that motivated its original introduction. If we just consider ferromagnetic interactions separately, the model with long-range interactions actually behaves more simply. If we just consider antiferromagnetic interactions, larger scale patterns of UP and DOWN spins arise. When we include both negative and positive interactions together, there will be additional features that enable a richer behavior. We will start by considering the case of ferromagnetic long-range interactions.

The primary effect of the increase in the range of ferromagnetic interactions is improvement of the mean field approximation. There are several ways to model interactions that extend beyond nearest neighbors in the Ising model. We could set a sphere of a particular radius r_0 around each spin and consider all of the spins within the sphere to be neighbors of the spin at the center.

$$E[\{s_i\}] = -\sum_i h_i s_i - \frac{1}{2} J \sum_{r_{ij} < r_0} s_i s_j \tag{1.6.51}$$

Here we do not restrict the summations over i and j in the second term,so we explicitly include a factor of 1/2 to avoid counting interactions twice. Alternatively, we could use an interaction $J(r_{ij})$ that decays either exponentially or as a power law with distance from each spin:

$$E[\{s_i\}] = -\sum_i h_i s_i - \frac{1}{2} \sum_{i,j} J(r_{ij}) s_i s_j \tag{1.6.52}$$

In both Eqs. (1.6.51) and (1.6.52) the self-interaction terms $i = j$ are generally to be excluded. Since $s_i^2 = 1$ they only add a constant to the energy.

Quite generally and independent of the range or even the variability of interactions, when all interactions are ferromagnetic, $J > 0$, then all the spins will align at low temperatures. The mean field approximation may be used to estimate the behavior. All cases then reduce to the same free energy (Eq. (1.6.36) or Eq. (1.6.41)) with a measure of the strength of the interactions replacing zJ. The only difference from the nearest neighbor model then relates to the accuracy of the mean field approximation. It is simplest to consider the model of a fixed interaction strength with a cutoff length. The mean field is accurate when the correlation length is shorter than the interaction distance. When this occurs, a spin is interacting with other spins that are uncorrelated with it. The averaging used to obtain the mean field is then correct. Thus the approximation improves if the interaction between spins becomes longer ranged. However, the correlation length becomes arbitrarily long near the phase transition. Thus, for longer interaction lengths, the mean field approximation holds closer to T_c but eventually becomes inaccurate in a narrow temperature range around T_c. There is one model for which the mean field approximation is exact independent of temperature or dimension. This is a model of infinite range interactions discussed in Question 1.6.11. The distance-dependent interaction model of Eq. (1.6.52) can be shown to behave like a finite range interaction model for interactions that decay more rapidly than $1/r$ in 3-d. For weaker decay than $1/r$ this model is essentially the same as the long-range interaction model of Question 1.6.11. Interactions that decay as $1/r$ are a borderline case.

Question 1.6.11 Solve the Ising model with infinite ranged interactions in a uniform magnetic field. The infinite range means that all spins interact with the same interaction strength. In order to keep the energy extrinsic (proportional to the volume) we must make the interactions between pairs of spins weaker as the system becomes larger, so replace $J \rightarrow J/N$. The energy is given by:

$$E[\{s_i\}] = -h\sum_i s_i - \frac{1}{2N} J \sum_{i,j} s_i s_j \tag{1.6.53}$$

For simplicity, keep the $i = j$ terms in the second sum even though they add only a constant.

Solution 1.6.11 We can solve this problem exactly by rewriting the energy in terms of a collective coordinate which is the average over the spin variables

$$m = \frac{1}{N}\sum_i s_i \tag{1.6.54}$$

in terms of which the energy becomes:

$$E(\{s_i\}) = hNm - \frac{1}{2} JNm^2 \tag{1.6.55}$$

This is the same as the mean field Eq. (1.6.39) with the substitution $Jz \rightarrow J$. Here the equation is exact. The result for the entropy is the same as before, since we have fixed the average value of the spin by Eq.(1.6.54). The solution for the value of m for $h = 0$ is given by Eq.(1.6.32) and Fig. 1.6.4. For $h \neq 0$ the discussion in Question 1.6.4 applies. ▮

The case of antiferromagnetic interactions will be considered in greater detail in Chapter 7. If all interactions are antiferromagnetic $J < 0$,then extending the range of the interactions tends to reduce their effect, because it is impossible for neighboring spins to be antialigned and lower the energy. To be antialigned with a neighbor is to be aligned with a second neighbor. However, by forming patches of UP and DOWN spins it is possible to lower the energy. In an infinite-ranged antiferromagnetic system,all possible states with zero magnetization have the same lowest energy at $h = 0$.

This can be seen from the energy expression in Eq.(1.6.55). In this sense,frustration from many sources is almost the same as no interaction.

In addition to the ferromagnet and antiferromagnet, there is a third possibility where there are both positive and negative interactions. The physical systems that have motivated the study of such models are known as spin glasses. These are materials where magnetic atoms are found or placed in a nonmagnetic host.The randomly placed magnetic sites interact via long-range interactions that oscillate in sign with distance. Because of the randomness in the location of the spins, there is a randomness in the interactions between them. Experimentally, it is found that such systems also undergo a transition that has been compared to a glass transition, and therefore these systems have become known as spin glasses.

A model for these materials, known as the Sherrington-Kirkpatrick spin glass, makes use of the Ising model with infinite-range random interactions:

$$E[\{s_i\}] = -\frac{1}{2N}\sum_{ij} J_{ij}s_is_j \qquad (1.6.56)$$
$$J_{ij} = \pm J$$

The interactions J_{ij} are fixed uncorrelated random variables—quenched variables. The properties of this system are to be averaged over the random variables J_{ij} but only after it is solved.

Similar to the ferromagnetic or antiferromagnetic Ising model,at high temperatures $kT >> J$ the spin glass model has a disordered phase where spins do not feel the effect of the interactions beyond the existence of correlations. As the temperature is lowered,the system undergoes a transition that is easiest to describe as a breaking of ergodicity. Because of the random interactions,some arrangements of spins are much lower in energy than others. As with the case of the antiferromagnet on a triangular lattice,there are many of these low-energy states. The difference between any two of these states is large,so that changing from one state to the other would involve the flipping of a finite fraction of the spins of the system. Such a flipping would have to be cooperative, so that overcoming the barrier between low-energy states becomes impossible below the transition temperature during any reasonable time. The low-energy states have been shown to be organized into a hierarchy determined by the size of the overlaps between them.

Question 1.6.12 Solve a model that includes a special set of correlated random interactions of the type of the Sherrington-Kirkpatrick model, where the interactions can be written in the *separable* form

$$J_{ij} = \xi_i\xi_j \qquad (1.6.57)$$
$$\xi_i = \pm 1$$

This is the Mattis model. For simplicity, keep the terms where $i = j$.

Solution 1.6.12 We can solve this problem by defining a new set of variables

$$s'_i = \xi_i s_i \qquad (1.6.58)$$

In terms of these variables the energy becomes:

$$E[\{s_i\}] = -\frac{1}{2N}\sum_{ij}\xi_i\xi_j s_i s_j = -\frac{1}{2N}\sum_{ij} s_i' s_j' \qquad (1.6.59)$$

which is the same as the ferromagnetic Ising model. The phase transition of this model would lead to a spontaneous magnetization of the new variables. This corresponds to a net orientation of the spins toward (or opposite) the state $s_i = \xi_i$. This can be seen from

$$m = < s_i' > = \xi_i < s_i > \qquad (1.6.60)$$

This model shows that a set of mixed interactions can cause the system to choose a particular low-energy state that behaves like the ordered state found in the ferromagnet. By extension, this makes it plausible that fully random interactions lead to a variety of low-energy states. ∎

The existence of a large number of randomly located energy minima in the spin glass might suggest that by engineering such a system we could control where the minima occur. Then we might use the spin glass as a memory. The Mattis model provides a clue to how this might be accomplished. The use of an outer product representation for the matrix of interactions turns out to be closely related to the model developed by Hebb for biological imprinting of memories on the brain. The engineering of minima in a long-range-interaction Ising model is precisely the model developed by Hopfield for the behavior of neural networks that we will discuss in Chapter 2.

In the ferromagnet and antiferromagnet, there were intuitive ways to deal with the breaking of ergodicity, because we could easily define a macroscopic parameter (the magnetization) that differentiated between different macroscopic states of the system. More general ways to do this have been developed for the spin glass and applied to the study of neural networks.

1.6.7 *Kinetics of the Ising model*

We have introduced the Ising model without the benefit of a dynamics. There are many choices of dynamics that would lead to the equilibrium ensemble given by the Ising model. One of the most natural would arise from considering each spin to have the two-state system dynamics of Section 1.4. In this dynamics, transitions between UP and DOWN occur across an intermediate barrier that sets the transition rate. We call this the activated dynamics and will use it to discuss protein folding in Chapter 4 because it can be motivated microscopically. The activated dynamics describes a continuous rate of transition for each of the spins. It is often convenient to consider transitions as occurring at discrete times. A particularly simple dynamics of this kind was introduced by Glauber for the Ising model. It also corresponds to the dynamics popular in studies of neural networks that we will discuss in Chapter 2. In this section we will show that the two different dynamics are quite closely related. In Section 1.7 we will consider several other forms of dynamics when we discuss Monte Carlo simulations.

If there are many different possible ways to assign a dynamics to the Ising model, how do we know which one is correct? As for the model itself, it is necessary to consider the system that is being modeled in order to determine which kinetics is appropriate. However, we expect that there are many different choices for the kinetics that will provide essentially the same results as long as we consider its long time behavior. The central limit theorem in Section 1.2 shows that in a stochastic process, many independent steps lead to the same Gaussian distribution of probabilities,independent of the specific steps that are taken. Similarly, if we choose a dynamics for the Ising model that allows individual spin flips, the behavior of processes that involve many spin flips should not depend on the specific dynamics chosen. Having said this, we emphasize that the conditions under which different dynamic rules provide the same long time behavior are not fully established. This problem is essentially the same as the problem of classifying dynamic systems in general. We will discuss it in more detail in Section 1.7.

Both the activated dynamics and the Glauber dynamics assume that each spin relaxes from its present state toward its equilibrium distribution. Relaxation of each spin is independent of other spins. The equilibrium distribution is determined by the relative energy of its UP and DOWN state at a particular time. The energy difference between having the i th spin s_i UP and DOWN is:

$$E_{+i}(\{s_j\}_{j\neq i}) = E(s_i = +1,\{s_j\}_{j\neq i}) - E(s_i = -1,\{s_j\}_{j\neq i}) \tag{1.6.61}$$

The probability of the spin being UP or DOWN is given by Eq. (1.4.14) as:

$$P_{s_i}(1) = \frac{1}{1+e^{E_{+i}/kT}} = f(E_{+i}) \tag{1.6.62}$$

$$P_{s_i}(-1) = 1 - f(E_{+i}) = f(-E_{+i}) \tag{1.6.63}$$

In the activated dynamics, all spins perform transitions at all times with rates $R(1|-1)$ and $R(-1|1)$ given by Eqs.(1.4.38) and (1.4.39) with a site-dependent energy barrier E_{Bi} that sets the relaxation time for the dynamics τ_i. As with the two-state system, it is assumed that each transition occurs essentially instantaneously. The choice of the barrier E_{Bi} is quite important for the kinetics, particularly since it may also depend on the state of other spins with which the i th spin interacts. As soon as one of the spins makes a transition,all of the spins with which it interacts must change their rate of relaxation accordingly. Instead of considering directly the rate of transition, we can consider the evolution of the probability using the Master equation, Eq. (1.4.40) or (1.4.43). This would be convenient for Master equation treatments of the whole system. However, the necessity of keeping track of all of the probabilities makes this impractical for all but simple considerations.

Glauber dynamics is simpler in that it considers only one spin at a time. The system is updated in equal time intervals.Each time interval is divided into N small time increments. During each time increment, we select a particular spin and only consider its dynamics. The selected spin then relaxes completely in the sense that its state is set to be UP or DOWN according to its equilibrium probability, Eq. (1.6.62). The transitions of different spins occur sequentially and are not otherwise coupled. The way we select which spin to update is an essential part of the Glauber dynamics. The simplest and most commonly used approach is to select a spin at random in each time increment. This means that we do not guarantee that every spin is selected during a time interval consisting of N spin updates.Likewise,some spins will be updated more than once in a time interval.On average,however, every spin is updated once per time interval.

In order to show that the Glauber dynamics are intimately related to the activated dynamics, we begin by considering how we would implement the activated dynamics on an ensemble of independent two-state systems whose dynamics are completely determined by the relaxation time $\tau = (R(1|-1) + R(1|-1))^{-1}$ (Eq.(1.4.44)). We can think about this ensemble as representing the dynamics of a single two-state system, or, in a sense that will become clear, as representing a noninteracting Ising model. The total number of spins in our ensemble is N. At time t the ensemble is described by the number of UP spins given by $NP(1;t)$ and the number of DOWN spins $NP(-1;t)$.

We describe the activated dynamics of the ensemble using a small time interval Δt, which eventually we would like to make as small as possible. During the interval of time Δt, which is much smaller than the relaxation time τ, a certain number of spins make transitions. The probability that a particular spin will make a transition from UP to DOWN is given by $R(-1|1)\Delta t$. The total number of spins making a transition from DOWN to UP, and from UP to DOWN, is:

$$\begin{gathered} NP(-1;t)R(1|-1)\Delta t \\ NP(1;t)R(-1|1)\Delta t \end{gathered} \tag{1.6.64}$$

respectively. To implement the dynamics, we must randomly pick out of the whole ensemble this number of UP spins and DOWN spins and flip them. The result would be a new number of UP and DOWN spins $NP(1;t + \Delta t)$ and $NP(-1;t + \Delta t)$. The process would then be repeated.

It might seem that there is no reason to randomly pick the ensemble elements to flip, because the result is the same if we rearrange the spins arbitrarily. However, if each spin represents an identifiable physical system (e.g., one spin out of a noninteracting Ising model) that is performing an internal dynamics we are representing, then we must randomly pick the spins to flip.

It is somewhat inconvenient to have to worry about selecting a particular number of UP and DOWN spins separately. We can modify our prescription so that we select a subset of the spins regardless of orientation. To achieve this, we must allow that some of the selected spins will be flipped and some will not. We select a fraction η of the spins of the ensemble. The number of these that are DOWN is $\eta NP(-1;t)$. In order to flip the same number of spins from DOWN to UP, as in Eq.(1.6.64), we must flip UP a fraction $R(1|-1)\Delta t/\eta$ of the $\eta NP(-1;t)$ spins. Consequently, the fraction of spins we do not flip is $(1 - R(1|-1)\Delta t/\eta)$. Similarly, the number of selected UP spins is $\eta \mathrm{NP}(1;\mathrm{t})$ the fraction of these to be flipped is $R(-1|1)\Delta t/\eta$, and the fraction we do not flip is $(1 - R(-1|1)\,\Delta t/\eta)$. In order for these expressions to make sense (to be positive) η must be large enough so that at least one spin will be flipped. This implies $\eta >$ max $(R(1|-1)\Delta t,\ R(-1|1)\Delta t)$. Moreover, we do not want η to be larger than it must be because this will just force us to select additional spins we will not be flipping. A convenient choice would be to take

$$\eta = (R(1| - 1) + R(-1|1))\Delta t = \Delta t/\tau \tag{1.6.65}$$

The consequences of this choice are quite interesting, since we find that the fraction of selected DOWN spins to be flipped UP is $R(1|-1)$ / $(R(1|-1) + R(-1|1)) = P(1)$, the equilibrium fraction of UP spins. The fraction not to be flipped is the equilibrium fraction of DOWN spins. Similarly, the fraction of selected UP spins that are to be flipped DOWN is the equilibrium fraction of DOWN spins, and the fraction to be left UP is the equilibrium fraction of UP spins. Consequently, the outcome of the dynamics of the selected spin does not depend at all on the initial state of the spin. The revised prescription for the dynamics is to select a fraction η of spins from the ensemble and set them according to their equilibrium probability.

We still must choose the time interval Δt. The smallest time interval that makes sense is the interval for which the number of selected spins would be just one. A smaller number would mean that sometimes we would not choose any spins. Setting the number of selected spins $\eta N = 1$ using Eq. (1.6.65) gives:

$$\Delta t = \frac{1}{N(R(1|-1) + R(-1|1))} = \frac{\tau}{N} \tag{1.6.66}$$

which also implies the condition $\Delta t << \tau$, and means that the approximation of a finite time increment Δt is directly coupled to the size of the ensemble. Our new prescription is that we select a single spin and set it UP or DOWN according to its equilibrium probability. This would be the prescription of Glauber dynamics if the ensemble were considered to be the Ising model without interactions. Thus for a noninteracting Ising model, the Glauber dynamics and the activated dynamics are the same. So far we have made no approximation except the finite size of the ensemble. We still have one more step to go to apply this to the interacting Ising model.

The activated dynamics is a stochastic dynamics, so it does not make sense to discuss only the dynamics of a particular system but the dynamics of an ensemble of Ising models. At any moment, the activated dynamics treats the Ising model as a collection of several kinds of spins. Each kind of spin is identified by a particular value of E_+ and E_B. These parameters are controlled by the local environment of the spin. The dynamics is not concerned with the source of these quantities, only their values. The dynamics are that of an ensemble consisting of several kinds of spins with a different number N_k of each kind of spin, where k indexes the kind of spin. According to the result of the previous paragraph, and specifically Eq. (1.6.65), we can perform this dynamics over a time interval Δt by selecting $N_k \Delta t/\tau_k$ spins of each kind and updating them according to the Glauber method. This is strictly applicable only for an ensemble of Ising systems. If the Ising system that we are considering contains many correlation lengths, Eq. (1.6.50), then it represents the ensemble by itself. Thus for a large enough Ising model, we can apply this to a single system.

If we want to select spins arbitrarily, rather than of a particular kind, we must make the assumption that all of the relaxation times are the same, $\tau_k \rightarrow \tau$. This assumption means that we would select a total number of spins:

$$\sum_k \frac{N_k \Delta t}{\tau_k} \rightarrow N \frac{\Delta t}{\tau} \tag{1.6.67}$$

As before, Δt may also be chosen so that in each time interval only one spin is selected.

Using two assumptions, we have been able to derive the Glauber dynamics directly from the activated dynamics.One of the assumptions is that the dynamics must be considered to apply only as the dynamics of an ensemble. Even though both dynamics are stochastic dynamics, applying the Glauber dynamics directly to a single system is only the same as the activated dynamics for a large enough system. The second assumption is the equivalence of the relaxation times τ_k. When is this assumption valid? The expression for the relaxation time in terms of the two-state system is given by Eq. (1.4.44) as

$$1/\tau = (R(1|–1) + R(-1|1)) = \nu(e^{-(E_B - E_1)/kT} + e^{-(E_B - E_{-1})/kT}) \tag{1.6.68}$$

When the relative energy of the two states E_1 and E_{-1} varies between different spins, this will in general vary. The size of the relaxation time is largely controlled by the smaller of the two energy differences $E_B - E_1$ and $E_B - E_{-1}$. Thus,maintaining the same relaxation time would require that the smaller energy difference is nearly constant. This is essential, because the relaxation time changes exponentially with the energy difference.

We have shown that the Glauber dynamics and the activated dynamics are closely related despite appearing to be quite different. We have also found how to generalize the Glauber dynamics if we must allow different relaxation times for different spins. Finally, we have found that the time increment for a single spin update corresponds to τ / N. This means that a single Glauber time step consisting of N spin updates corresponds to a physical time τ—the microscopic relaxation time of the individual spins.

At this point we have introduced a dynamics for the Ising model, and it should be possible for us to investigate questions about its kinetics.Often questions about the kinetics may be described in terms of time correlations. Like the correlation length, we can introduce a correlation time τ_s that is given by the decay of the spin-spin correlation

$$< s_i(t')s_i(t) > - < s_i >^2 \quad e^{-|t-t'|/\tau_s} \tag{1.6.69}$$

For the case of a relaxing two-state system,the correlation time is the relaxation time τ. This follows from Eq. (1.4.45), with some attention to notation as described in Question 1.6.13.

Question 1.6.13 Show that for a two-state system, the correlation time is the relaxation time τ.

Solution 1.6.13 The difficulty in this question is restoring some of the notational details that we have been leaving out for convenience. From Eq. (1.6.45) we have for the average:

$$\begin{aligned} < s_i(t')s_i(t) > = & P_{s_i(t'),s_i(t)}(1,1) + P_{s_i(t'),s_i(t)}(-1,-1) \\ & - P_{s_i(t'),s_i(t)}(1,-1) - P_{s_i(t'),s_i(t)}(-1,1) \end{aligned} \tag{1.6.70}$$

Let's assume that $t' > t$, then each of these joint probabilities of the form $\mathrm{P}_{s_i(t'),s_i(t)}(s_2,s_1)$ is given by the probability that the two-state system starts in the state s_1 at time t, multiplied by the probability that it will evolve from s_1 into s_2 at time t'.

$$P_{s_i(t'),s_i(t)}(s_2,s_1) = P_{s_i(t'),s_i(t)}(s_2 | s_1) P_{s_i(t)}(s_1) \tag{1.6.71}$$

The first factor on the right is called the conditional probability. The probability for a particular state of the spin is the equilibrium probability that we wrote as $P(1)$ and $P(-1)$. The conditional probabilities satisfy $P_{s_i(t'),s_i(t)}$(1çs_1) + $P_{s_i(t'),s_i(t)}$(−1çs_1) = 1 , so we can simplify Eq. (1.6.70) to:

$$< s_i(t')s_i(t) > = (2P_{s_i(t'),s_i(t)}(1|1) - 1)P(1) + (2P_{s_i(t'),s_i(t)}(-1|-1) - 1)P(-1) \tag{1.6.72}$$

The evolution of the probabilities are described by Eq.(1.4.45),repeated here:

$$P(1;t) = (P(1;0) - P(1;\infty))e^{-t/\tau} + P(1;\infty) \tag{1.6.73}$$

Since the conditional probability assumes a definite value for the initial state (e.g., $P(1;0) = 1$ for $P_{s(t'),s(t)}(1|1)$), we have:

$$\begin{aligned} P_{s(t'),s(t)}(1|1) &= (1 - P(1))e^{-(t'-t)/\tau} + P(1) \\ P_{s(t'),s(t)}(-1|-1) &= (1 - P(-1))e^{-(t'-t)/\tau} + P(-1) \end{aligned} \tag{1.6.74}$$

Inserting these into Eq. (1.6.72) gives:

$$\begin{aligned} < s_i(t')s_i(t) > = & (2\Big[(1 - P(1))e^{-(t'-t)/\tau} + P(1)\Big] - 1)P(1) \\ & + (2\Big[(1 - P(-1))e^{-(t'-t)/\tau} + P(-1)\Big] - 1)P(-1) \\ = & 4P(1)P(-1)e^{-(t'-t)/\tau} + (P(1) - P(-1))^2 \end{aligned} \tag{1.6.75}$$

The constant term on the right is the same as the square of the average of the spin:

$$< s_i(t) >^2 = (P(1) - P(-1))^2 \tag{1.6.76}$$

Inserting into Eq.(1.6.69) leads to the desired result (we have assumed that $t' > t$):

$$< s_i(t')s_i(t) > - < s_i(t) >^2 = 4P(1)P(-1)e^{-(t'-t)/\tau} \quad e^{-(t'-t)/\tau} \tag{1.6.77}$$ ∎

From the beginning of our discussion of the Ising model,a central issue has been the breaking of the ergodic theorem associated with the spontaneous magnetization. Now that we have introduced a kinetic model, we will tackle this problem directly. First we describe the problem fully. The ergodic theorem states that a time average may be replaced by an ensemble average. In the ensemble,all possible states of the system are included with their Boltzmann probability. Without formal justification, we have treated the spontaneous magnetization of the Ising model at $h = 0$ as a macroscopically observable quantity. According to our prescription,this is not the case. Let us perform the average $< s_i >$ over the ensemble at $T = 0$ and $h = 0$. There are two possible states of the system with the same energy, one with $\{s_i = 1\}$ and one with $\{s_i = –1\}$. Since they must occur with equal probability by our assumption, we have that the average $< s_i >$ is zero.

This argument breaks down because of the kinetics of the system that prevents a transition from one state to the other during the course of a measurement. Thus we measure only one of the two possible states and find a magnetization of 1 or –1. How can we prove that this system breaks the ergodic theorem? The most direct test is to start from a system with a slightly positive magnetic field near $T = 0$ where the magnetization is +1, and reverse the sign of the magnetic field. In this case the equilibrium state of the system should have a magnetization of –1. Instead the system will maintain its magnetization as +1 for a long time before eventually switching from one to the other. The process of switching corresponds to the kinetics of a first-order transition.

1.6.8 *Kinetics of a first-order phase transition*

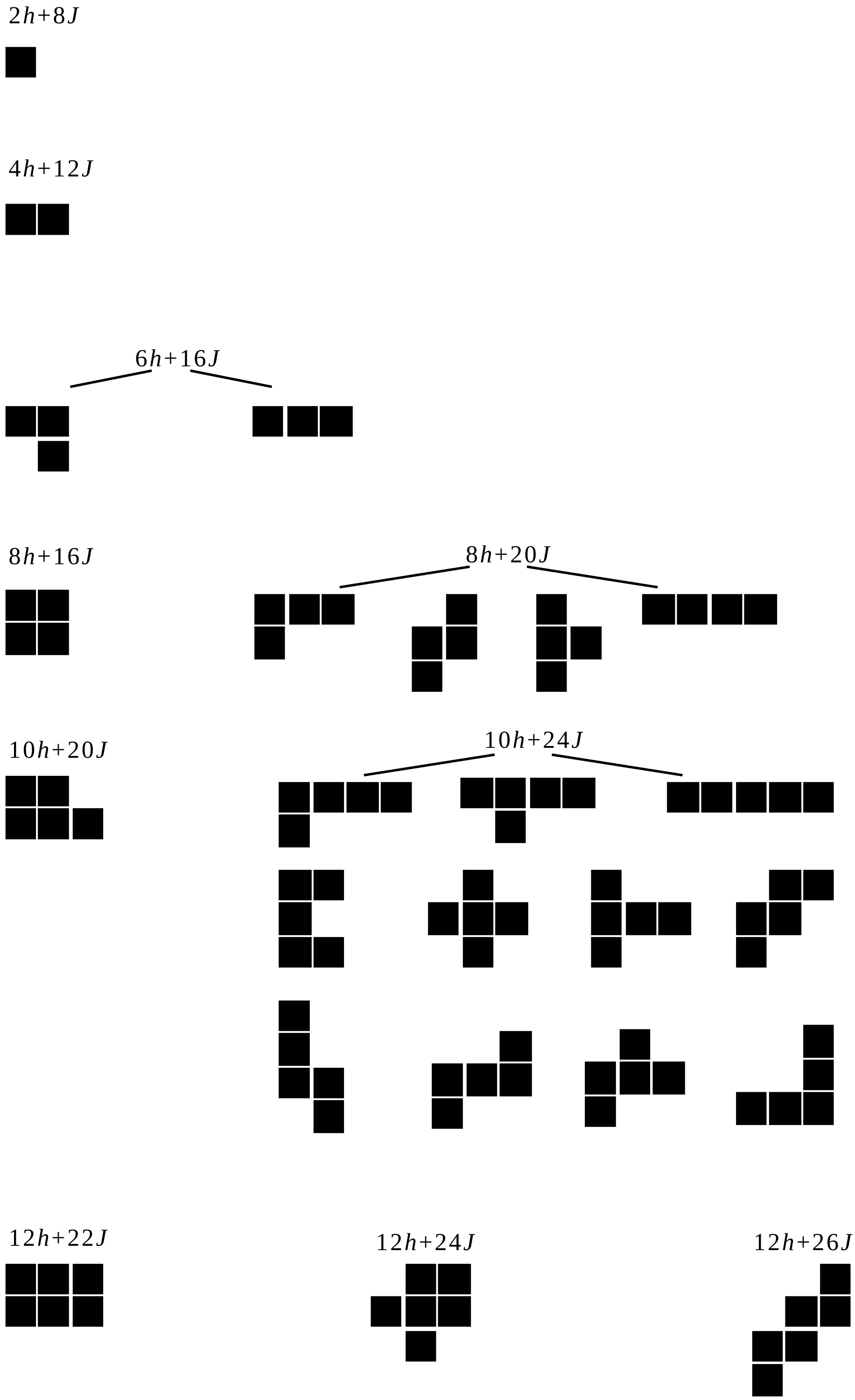
In this section we discuss the first-order transition kinetics in the Ising model. Similar arguments apply to other first-order transitions like the freezing or boiling of water. If we start with an Ising model in equilibrium at a temperature $T < T_c$ and a small positive magnetic field $h << zJ$, the magnetization of the system is essentially $m_0(\beta zJ)$. If we change the magnetic field suddenly to a small negative value, the equilibrium state of the system is $-m_0(\beta zJ)$;however, the system will require some time to change its magnetization. The change in the magnetic field has very little effect on the energy of an individual spin s_i. This energy is mostly due to the interaction with its neighbors, with a relatively small contribution due to the external field. Most of the time the neighbors are oriented UP, and this makes the spin have a lower energy when it is UP. This gives rise to the magnetization $m_0(\beta zJ)$. Until s_i's neighbors change their average magnetization, s_i has no reason to change its magnetization. But then neither do the neighbors. Thus, because each spin is in its own local equilibrium,the process that eventually equilibrates the system requires a cooperative effect including more than one spin. The process by which such a first-order transition occurs is not the simultaneous switching of all of the spins from one value to the other. This would require an impossibly long time. Instead the transition occurs by nucleation and growth of the equilibrium phase.

It is easiest to describe the nucleation process when T is sufficiently less than T_c, so that the spins are almost always +1. In mean field, already for $T < 0.737T_c$ the probability of a spin being UP is greater than 90% ($P(1) = (1 + m)/2 > 0.9$),and for $T < 0.61T_c$ the probability of a spin being UP is greater than 95%. As long as T is greater than zero, individual spins will flip from time to time. However, even though the magnetic field would like them to be DOWN, their local environment consisting of UP spins does not. Since the interaction with their neighbors is stronger than the interaction with the external field,the spin will generally flip back UP after a short time. There is a smaller probability that a second spin,a neighbor of the first spin, will also flip DOWN. Because one of the neighbors of the second spin is already DOWN, there is a lower energy cost than for the first one. However, the energy of the second spin is still higher when it is DOWN, and the spins will generally flip back, first one then the other. There is an even smaller probability that three interacting spins will flip DOWN. The existence of two DOWN spins makes it more likely for the third to do so. If the first two spins were neighbors,than the third spin can have only one of them as its neighbor. So it still costs some energy to flip DOWN the third spin. If there are three spins flipped DOWN in an L shape,the spin that completes a 2 × 2 square has two neighbors that are +1 and two neighbors that are –1,so the interactions with its neighbors cancel. The external field then gives a preference for it to be DOWN. There is still a high probability that several of the spins that are DOWN will flip UP and the little cluster will then disappear. Fig. 1.6.9 shows various clusters and their energies compared to a uniform region of +1 spins. As more spins are added,the internal region of the cluster becomes composed of spins that have four neighbors that are all DOWN. Beyond a certain size (see Question 1.6.14) the cluster of DOWN spins will grow, because adding spins lowers the energy of the system. At some point the growing region of DOWN spins encounters another region of DOWN spins and the whole system reaches its new equilibrium state, where most spins are DOWN.

Question 1.6.14 Using an estimate of how the energy of large clusters of DOWN spins grows, show that large enough clusters must have a lower energy than the same region if it were composed of UP spins.

Solution 1.6.14 The energy of a cluster of DOWN spins is given by its interaction with the external magnetic field and the number of antialigned bonds that form its boundary. The change in energy due to the external magnetic field is exactly $2hN_c$, which is proportional to the number of spins in the

Figure 1.6.9 Illustration of small clusters of DOWN spins shown as filled dark squares residing in a background of UP spins on a square lattice. The energies for creating the clusters are shown. The magnetic field, h, is negative. The formation of such clusters is the first step towards nucleation of a DOWN region when the system undergoes a first-order transition from UP to DOWN. The energy is counted by the number of spins that are DOWN times the magnetic field strength, plus the interaction strength times the number of antialigned neighboring spins, which is the length of the boundary of the cluster. In a first-order transition, as the size of the clusters grows the gain from orienting toward the magnetic field eventually becomes greater than the loss from the boundary energy. Then the cluster becomes more likely to grow than shrink. See Question 1.6.14 and Fig. 1.6.10. ∎


$2h+8J$
$4h+12J$
$6h+16J$
$8h+16J$
$8h+20J$
$10h+20J$
$10h+24J$
$12h+22J$
$12h+24J$
$12h+26J$

cluster N_c. This is negative since h is negative. The energy of the boundary is proportional to the number of antialigned bonds, and it is always positive. Because every additional antialigned bond raises the cluster energy, the boundary of the cluster tends to be smooth at low temperatures. Therefore, we can estimate the boundary energy using a simple shape like a square or circular cluster in 2-d (a cube or ball in 3-d). Either way the energy will increase as $fJN_c^{(d\text{-}1)/d}$, where d is the dimensionality and f is a constant accounting for the shape. Since the negative contribution to the energy increases, in proportion to the area (volume) of the cluster, and the positive contribution to the energy increases in proportion to the perimeter (surface area) of the cluster, the negative term eventually wins.Once a cluster is large enough so that its energy is dominated by the interaction with the magnetic field,then, on-average, adding an additional spin to the cluster will lower the system energy. ∎

Question 1.6.15 Without looking at Fig. 1.6.9, construct all of the different possible clusters of as many as five DOWN spins.Label them with their energy.

Solution 1.6.15 See Fig. 1.6.9. ∎

The scenario just described, known as nucleation and growth, is generally responsible for the kinetics of first-order transitions. We can illustrate the process schematically (Fig. 1.6.10) using a one dimensional plot indicating the energy per spin of a cluster as a function of the number of atoms in the cluster. The energy of the cluster increases at first when there are very few spins in the cluster, and then decreases once it is large enough. Eventually the energy decreases linearly with the number of spins in the cluster. The decrease per spin is the energy difference per spin between the two phases. The first cluster size that is "over the hump" is known as the critical cluster. The process of reaching this cluster is known as nucleation.A first estimate of the time to nucleate a critical cluster at a particular place in space is given by the inverse of the Boltzmann factor of the highest energy barrier in Fig. 1.6.10. This corresponds to the rate of transition over the barrier given by a two-state system with this same barrier (see Eq. (1.4.38) and Eq. (1.4.44)). The size of the critical cluster depends on the magnitude of the magnetic field.A larger magnetic field implies a smaller critical cluster. Once the critical cluster is reached,the kinetics corresponds to the biased diffusion described at the end of Section 1.4. The primary difficulty with an illustration such as Fig. 1.6.10 is that it is one-dimensional. We would need to show the energy of each type of cluster and all of the ways one cluster can transform into another. Moreover, the clusters themselves may move in space and merge or separate. In Fig. 1.6.11 we show frames from a simulation of nucleation in the Ising model using Glauber dynamics. The frames illustrate the process of nucleation and growth.

Experimental studies of nucleation kinetics are sometimes quite difficult. In physical systems,impurities often lower the barrier to nucleation and therefore control the rate at which the first-order transition occurs. This can be a problem for the investigation of the inherent nucleation because of the need to study highly purified

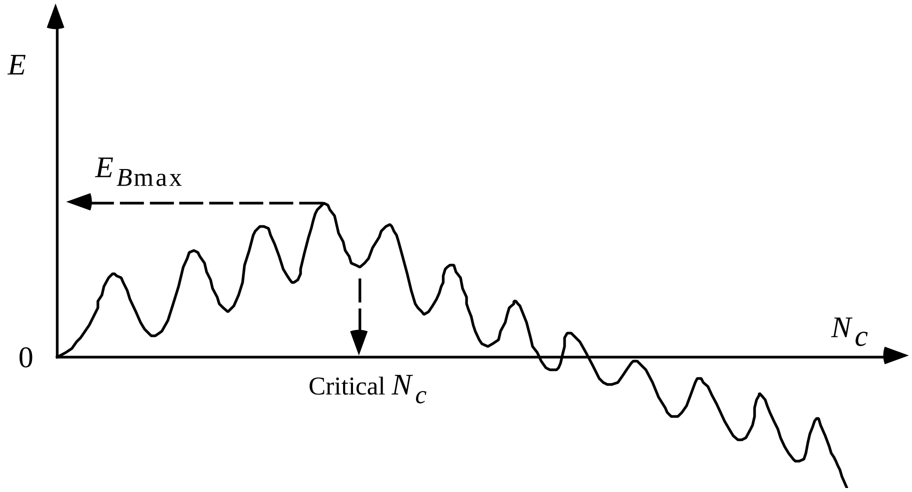


Figure 1.6.10 Schematic illustration of the energies that control the kinetics of a first-order phase transition. The horizontal axis is the size of a cluster of DOWN spins N_c that are the equilibrium phase. The cluster is in a background of UP spins that are the metastable phase. The vertical axis is the energy of the cluster. Initially the energy increases with cluster size until the cluster reaches the critical cluster size. Then the energy decreases. Each spin flip has its own barrier to overcome, leading to a washboard potential. The highest barrier $E_{B\max}$ that the system must overcome to create a critical nucleus controls the rate of nucleation. This is similar to the relaxation of a two-level system discussed in Section 1.4. However, this simple picture neglects the many different possible clusters and the many ways they can convert into each other by the flipping of spins. A few different types of clusters are shown in Fig. 1.6.9. ∎

systems. However, this sensitivity should be understood as an opportunity for control over the kinetics. It is similar to the sensitivity of electrical properties to dopant impurities in a semiconductor, which enables the construction of semiconductor devices. There is at least one direct example of the control of the kinetics of a first-order transition. Before describing the example, we review a few properties of the water-to-ice transition. The temperature of the water-to-ice transition can be lowered significantly by the addition of impurities. The freezing temperature of salty ocean water is lower than that of pure water. This suppression is thermodynamic in origin, which means that the T_c is actually lower. There exist fish that live in sub-zero-degrees ocean water whose blood has less salt than the surrounding ocean. These fish use a family of so-called antifreeze proteins that are believed to kinetically suppress the freezing of their blood. Instead of lowering the freezing temperature,these proteins suppress ice nucleation.

The existence of a long nucleation time implies that it is often possible to create metastable materials. For example, supercooled water is water whose temperature has been lowered below its freezing point. For many years, particle physicists used a superheated fluid to detect elementary particles. Ultrapure liquids in large tanks were

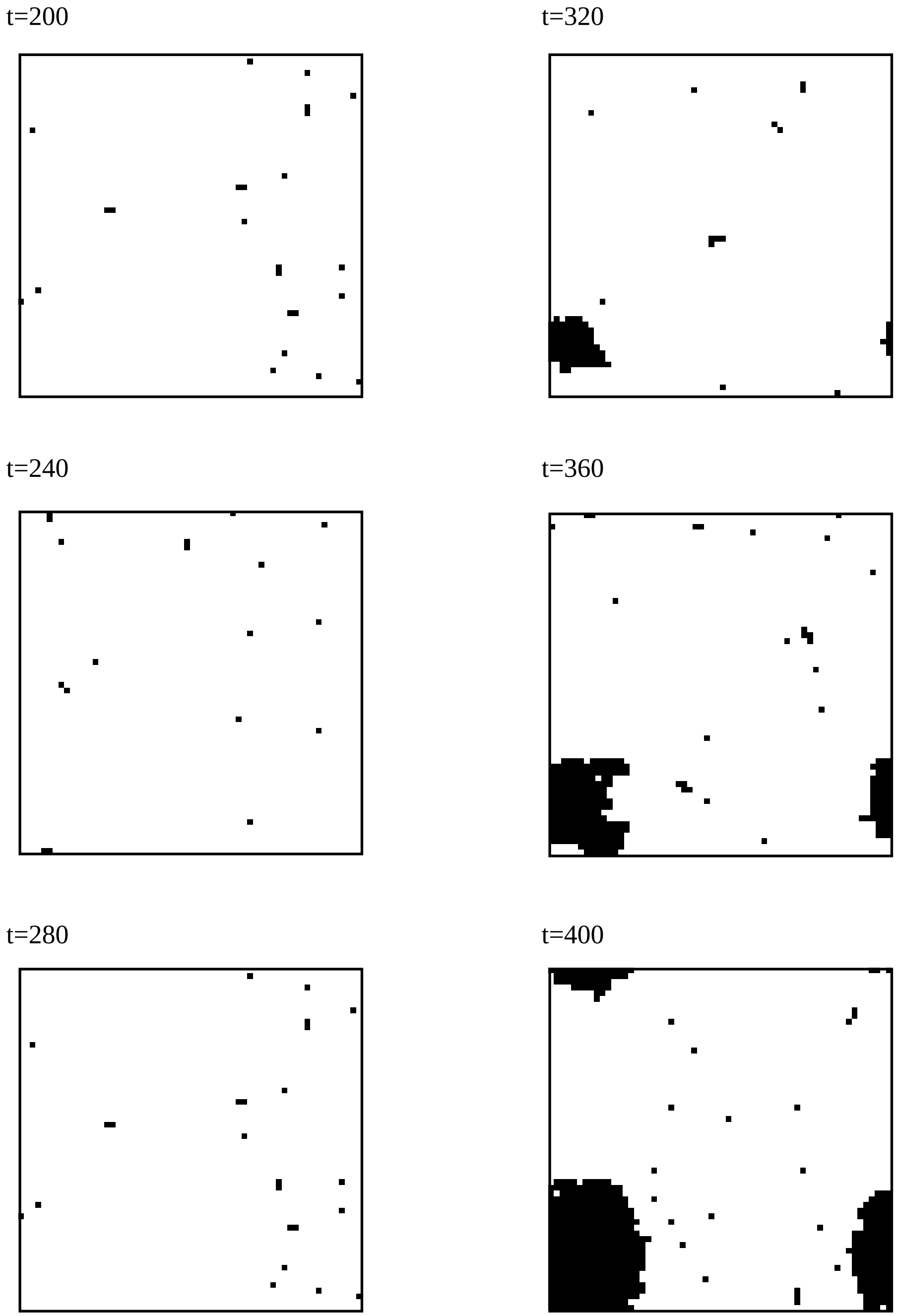
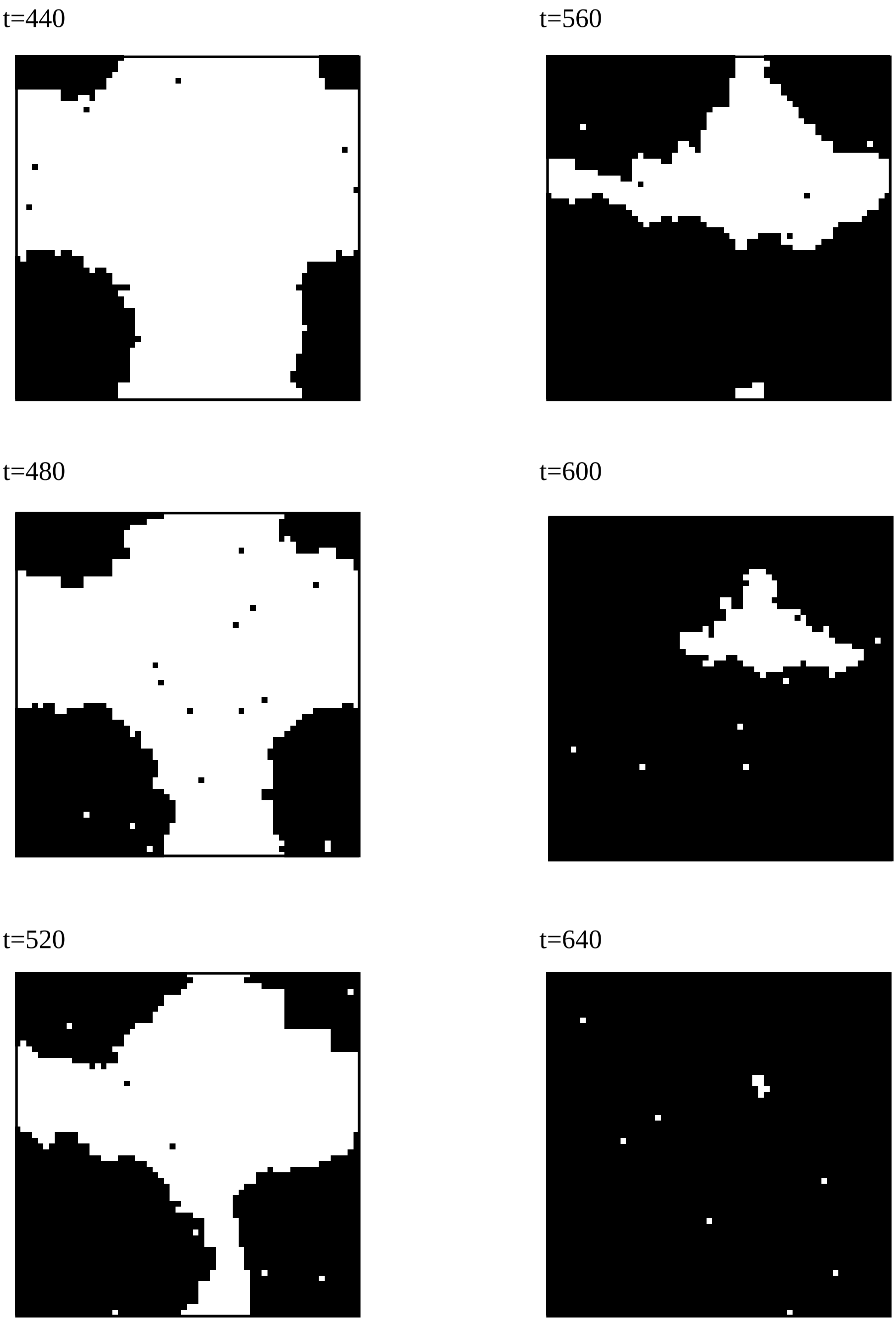


Figure 1.6.11 Frames from a simulation illustrating nucleation and growth in an Ising model in 2-d. The temperature is $T = zJ/3$ and the magnetic field is $h = -0.25$. Glauber dynamics was used. Each time step consists of N updates where the space size is $N = 60 \times 60$. Frames shown are in intervals of 40 time steps. The first frame shown is at $t = 200$ steps after the beginning of the simulation. Black squares are DOWN spins and white areas are UP spins. The



metastability of the UP phase is seen in the existence of only a few DOWN spins until the frame at $t = 320$. All earlier frames are qualitatively the same as the frames at $t = 200,240$ and 280. A critical nucleus forms between $t = 280$ and $t = 320$. This nucleus grows systematically until the final frame when the whole system is in the equilibrium DOWN phase. ▮

suddenly shifted above their boiling temperature. Small bubbles would then nucleate along the ionization trail left by charged particles moving through the tank. The bubbles could be photographed and the tracks of the particles identified. Such detectors were called bubble chambers. This methodology has been largely abandoned in favor of electronic detectors. There is a limit to how far a system can be supercooled or superheated. The limit is easy to understand in the Ising model. If a system with a positive magnetization *m* is subject to a negative magnetic field of magnitude greater than *zJm*, then each individual spin will flip DOWN independent of its neighbors. This is the ultimate limit for nucleation kinetics.

1.6.9 *Connections between CA and the Ising model*

Our primary objective throughout this section is the investigation of the equilibrium properties of interacting systems. It is useful, once again, to consider the relationship between the equilibrium ensemble and the kinetic CA we considered in Section 1.5. When a deterministic CA evolves to a unique steady state independent of the initial conditions, we can identify the final state as the $T = 0$ equilibrium ensemble. This is, however, not the way we usually consider the relationship between a dynamic system and its equilibrium condition. Instead, the equilibrium state of a system is generally regarded as the time average over microscopic dynamics. Thus when we use the CA to represent a microscopic dynamics, we could also identify a long time average of a CA as the equilibrium ensemble. Alternatively, we can consider a stochastic CA that evolves to a unique steady-state distribution where the steady state is the equilibrium ensemble of a suitably defined energy function.

1.7 Computer Simulations (Monte Carlo, Simulated Annealing)

Computer simulations enable us to investigate the properties of dynamical systems by directly studying the properties of particular models. Originally, the introduction of computer simulation was viewed by many researchers as an undesirable adjunct to analytic theory. Currently, simulations play such an important role in scientific studies that many analytic results are not believed unless they are tested by computer simulation. In part, this reflects the understanding that analytic investigations often require approximations that are not necessary in computer simulations. When a series of approximations has been made as part of an analytic study, a computer simulation of the original problem can directly test the approximations. If the approximations are validated, the analytic results often generalize the simulation results. In many other cases, simulations can be used to investigate systems where analytic results are unknown.

1.7.1 *Molecular dynamics and deterministic simulations*

The simulation of systems composed of microscopic Newtonian particles that experience forces due to interparticle interactions and external fields is called molecular dynamics. The techniques of molecular dynamics simulations, which integrate

Newton's laws for individual particles,have been developed to optimize the efficiency of computer simulation and to take advantage of parallel computer architectures. Typically, these methods implement a discrete iterative map (Section 1.1) for the particle positions. The most common (Verlet) form is:

$$r(t) = 2r(t - \Delta t) - r(t - 2\Delta t) + \Delta t^2 a(t - \Delta t) \tag{1.7.1}$$

where $a(t) = F(t)/m$ is the force on the particle calculated from models for interparticle and external forces. As in Section 1.1, time would be measured in units of the time interval Δt for convenience and efficiency of implementation. Eq. (1.7.1) is algebraically equivalent to the iterative map in Question 1.1.4, which is written as an update of both position and velocity:

$$\begin{aligned} r(t) &= r(t - \Delta t) + \Delta t v(t - \Delta t/2) \\ v(t + \Delta t/2) &= v(t - \Delta t/2) + \Delta t a(t) \end{aligned} \tag{1.7.2}$$

As indicated, the velocity is interpreted to be at half integral times, though this does not affect the result of the iterative map.

For most such simulations of physical systems,the accuracy is limited by the use of models for interatomic interactions. Modern efforts attempt to improve upon this approach by calculating forces from quantum mechanics. However, such simulations are very limited in the number of particles and the duration of a simulation.A useful measure of the extent of a simulation is the product Nt_{max} of the amount of physical time t_{max}, and the number of particles that are simulated N. Even without quantum mechanical forces,molecular dynamics simulations are still far from being able to describe systems on a space and time scale comparable to human senses. However, there are many questions that can be addressed regarding microscopic properties of molecules and materials.

The development of appropriate simplified macroscopic descriptions of physical systems is an essential aspect of our understanding of these systems. These models may be based directly upon macroscopic phenomenology obtained from experiment. We may also make use of the microscopic information obtained from various sources, including both theory and experiment, to inform our choice of macroscopic models. It is more difficult, but important as a strategy for the description of both simple and complex systems, to develop systematic methods that enable macroscopic models to be obtained directly from microscopic models. The development of such methods is still in its infancy, and it is intimately related to the issues of emergent simplicity and complexity discussed in Chapter 8.

Abstract mathematical models that describe the deterministic dynamics for various systems, whether represented in the form of differential equations or deterministic cellular automata (CA, Section 1.5), enable computer simulation and study through integration of the differential equations or through simulation of the CA. The effects of external influences, not incorporated in the parameters of the model, may be modeled using stochastic variables (Section 1.2). Such models, whether of fluids or of galaxies, describe the macroscopic behavior of physical systems by assuming that the microscopic (e.g., molecular) motion is irrelevant to the macroscopic phenomena being described. The microscopic behavior is summarized by parameters such as density, elasticity or viscosity. Such model simulations enable us to describe macroscopic phenomena on a large range of spatial and temporal scales.

1.7.2 *Monte Carlo simulations*

In our investigations of various systems, we are often interested in average quantities rather than a complete description of the dynamics. This was particularly apparent in Section 1.3, when equilibrium thermodynamic properties of systems were discussed. The ergodic theorem (Section 1.3.5) suggested that we can use an ensemble average instead of the space-time average of an experiment. The ensemble average enables us to treat problems analytically, when we cannot integrate the dynamics explicitly. For example, we studied equilibrium properties of the Ising model in Section 1.6 without reference to its dynamics. We were able to obtain estimates of its free energy, energy and magnetization by averaging various quantities using ensemble probabilities.

However, we also found that there were quite severe limits to our analytic capabilities even for the simplest Ising model. It was necessary to use the mean field approximation to obtain results analytically. The essential difficulty that we face in performing ensemble averages for complex systems,and even for the simple Ising model, is that the averages have to be performed over the many possible states of the system. For as few as one hundred spins,the number of possible states of the system—2^{100}—is so large that we cannot average over all of the possible states. This suggests that we consider approximate numerical techniques for studying the ensemble averages. In order to perform the averages without summing over all the states, we must find some way to select a representative sample of the possible states.

Monte Carlo simulations were developed to enable numerical averages to be performed efficiently. They play a central role in the use of computers in science. Monte Carlo can be thought of as a general way of estimating averages by selecting a limited sample of states of the system over which the averages are performed. In order to optimize convergence of the average, we take advantage of information that is known about the system to select the limited sample. As we will see, under some circumstances,the sequence of states selected in a Monte Carlo simulation may itself be used as a model of the dynamics of a system. Then,if we are careful about designing the Monte Carlo, we can separate the time scales of a system by treating the fast degrees of freedom using an ensemble average and still treat explicitly the dynamic degrees of freedom.

To introduce the concept of Monte Carlo simulation,we consider finding the average of a function $f(s)$, where the system variable s has the probability $P(s)$. For simplicity, we take s to be a single real variable in the range $[-1,+1]$. The average can be approximated by a sum over equally spaced values s_i:

$$< f(s) > = \int_{-1}^{1} f(s)P(s)ds \approx \sum_{s_i} f(s_i)P(s_i)\Delta s = \frac{1}{M}\sum_{n=-M}^{M} f(n/M)P(n/M) \qquad (1.7.3)$$

This formula works well if the functions $f(s)$ and $P(s)$ are reasonably smooth and uniform in magnitude. However, when they are not smooth,this sum can be a very inefficient way to perform the integral. Consider this integral when $P(s)$ is a Gaussian,and $f(s)$ is a constant:

$$< f(s) > \quad \int_{-1}^{1} e^{-s^2/2\sigma^2} ds \approx \frac{1}{M} \sum_{n=-M}^{M} e^{-(n/M)^2/2\sigma^2} \tag{1.7.4}$$

A plot of the integrand in Fig. 1.7.1 shows that for σ<<1 we are performing the integral by summing many values that are essentially zero. These values contribute nothing to the result and require as much computational effort as the comparatively few points that do contribute to the integral near $s = 0$, where the function is large. The few points near $s = 0$ will not give a very accurate estimate of the integral. Thus,most of the computational work is being wasted and the integral is not accurately evaluated. If we want to improve the accuracy of the sum, we have to increase the value of M. This means we will be summing many more points that are almost zero.

To avoid this problem, we would like to focus our attention on the region in Eq. (1.7.4) where the integrand is large. This can be done by changing how we select the points where we perform the average. Instead of picking the points at equal intervals along the line, we pick them with a probability given by $P(s)$. This is the same as saying that we have an ensemble representing the system with the state variable s. Then we perform the ensemble average:

$$< f(s) > = \int f(s) P(s) ds = \frac{1}{N} \sum_{s:P(s)}^{N} f(s) \tag{1.7.5}$$

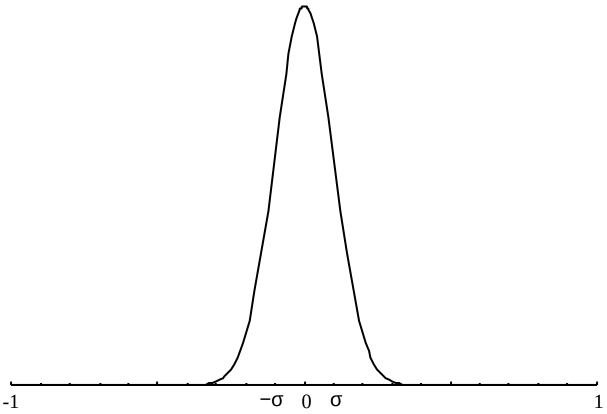


Figure 1.7.1 Plot of the Gaussian distribution illustrating that an integral that is performed by uniform sampling will use a lot of points to represent regions where the Gaussian is vanishingly small. The problem gets worse as σ becomes smaller compared to the region over which the integral must be performed. It is much worse in typical multidimensional averages where the Boltzmann probability is used. Monte Carlo simulations make such integrals computationally feasible by sampling the integrand in regions of high probability. ∎

The latter expression represents the sum over N values of s, where these values have the probability distribution $P(s)$. We have implicitly assumed that the function $f(s)$ is relatively smooth compared to $P(s)$. In Eq. (1.7.5) we have replaced the integral with a sum over an ensemble. The problem we now face is to obtain the members of the ensemble with probability $P(s)$. To do this we will invert the ergodic theorem of Section 1.3.5.

Since Section 1.3 we have described an ensemble as representing a system, if the dynamics of the system satisfied the ergodic theorem. We now turn this around and say that the ensemble sum in Eq.(1.7.5) can be represented by any dynamics that satisfies the ergodic theorem, and which has as its equilibrium probability $P(s)$. To do this we introduce a time variable t that, for our current purposes, just indicates the order of terms in the sum we are performing. The value of s appearing in the tth term would be $s(t)$. We then rewrite the ergodic theorem by considering the time average as an approximation to the ensemble average (rather than the opposite):

$$< f(s) > = \frac{1}{T}\sum_{t=1}^{T} f(s(t)) \tag{1.7.6}$$

The problem remains to sequentially generate the states $s(t)$, or, in other words, to specify the dynamics of the system. If we know the probability $P(s)$,and s is a few binary or real variables,this may be done directly with the assistance of a random number generator (Question 1.7.1). However, often the system coordinate s represents a large number of variables.A more serious problem is that for models of physical systems, we generally don’t know the probability distribution explicitly.

Thermodynamic systems are described by the Boltzmann probability (Section 1.3):

$$\begin{aligned} P(\{x,p\}) &= \frac{1}{Z} e^{-E(\{x,p\})/kT} \\ Z &= \sum_{\{x,p\}} e^{-E(\{x,p\})/kT} \end{aligned} \tag{1.7.7}$$

where $\{x,p\}$ are the microscopic coordinates of the system,and $E(\{x,p\})$ is the microscopic energy. An example of a quantity we might want to calculate would be the average energy:

$$U = \frac{1}{Z}\sum_{\{x,p\}} E(\{x,p\}) e^{-E(\{x,p\})/kT} \tag{1.7.8}$$

In many cases,as discussed in Section 1.4, the quantity that we would like to find the average of depends only on the position of particles and not on their momenta. We then write more generally

$$\begin{aligned} P(s) &= \frac{1}{Z_s} e^{-F(s)/kT} \\ Z_s &= \sum_{s} e^{-F(s)/kT} \end{aligned} \tag{1.7.9}$$

where we use the system state variable s to represent the relevant coordinates of the system. We make no assumption about the dimensionality of the coordinate s which may, for example, be the coordinates $\{x\}$ of all of the particles. $F(s)$ is the free energy of the set of states associated with the coordinate s. A precise definition, which indicates both the variable s and its value s', is given in Eq. (1.4.27):

$$F_s(s') = -kT \ln(\sum_{\{x,p\}} \delta_{s,s'}\, e^{-E(\{x,p\})/kT}) \tag{1.7.10}$$

We note that Eq. (1.7.9) is often written using the notation $E(s)$ (the energy of s) instead of $F(s)$ (the free energy of s),though $F(s)$ is more correct. An average we might calculate, of a quantity $Q(s)$, would be:

$$U = \frac{1}{Z}\sum_{s} Q(s) e^{-F(s)/kT} \tag{1.7.11}$$

where $Q(s)$ is assumed to depend only on the variable s and not directly on $\{x,p\}$.

The problem with the evaluation of either Eq. (1.7.8) or Eq. (1.7.11) is that the Boltzmann probability does not explicitly give us the probability of a particular state. In order to find the actual probability, we need to find the partition function Z. To calculate Z we need to perform a sum over all states of the system, which is computationally impossible. Indeed, if we were able to calculate Z, then, as discussed in Section 1.3, we would know the free energy and all the other thermodynamic properties of the system. So a prescription that relies upon knowing the actual value of the probability doesn't help us. However, it turns out that we don't need to know the actual probability in order to construct a dynamics for the system, only the relative probabilities of particular states. The relative probability of two states, $P(s) / P(s')$,is directly given by the Boltzmann probability in terms of their relative energy:

$$P(s) / P(s') = e^{-(F(s)-F(s'))/kT} \tag{1.7.12}$$

This is the key to Monte Carlo simulations. It is also a natural result, since a system that is evolving in time does not know global properties that relate to all of its possible states. It only knows properties that are related to the energy it has,and how this energy changes with its configuration. In classical mechanics, the change of energy with configuration would be the force experienced by a particle.

Our task is to describe a dynamics that generates a sequence of states of a system $s(t)$ with the proper probability distribution, $P(s)$. The classical (Newtonian) approach to dynamics implies that a deterministic dynamics exists which is responsible for generating the sequence of states of a physical system. In order to generate the equilibrium ensemble, however, there must be contact with a thermal reservoir. Energy transfer between the system and the reservoir introduces an external interaction that disrupts the system's deterministic dynamics.

We will make our task simpler by allowing ourselves to consider a stochastic Markov chain (Section 1.2) as the dynamics of the system. The Markov chain is described by the probability $P_s(s'|s'')$ of the system in a state $s = s''$ making a transition to the state $s = s'$. A particular sequence $s(t)$ is generated by starting from one configuration and choosing its successors using the transition probabilities.

The general formulation of a Markov chain includes the classical Newtonian dynamics and can also incorporate the effects of a thermal reservoir. However, it is generally convenient and useful to use a Monte Carlo simulation to evaluate averages that do not depend on the momenta,as in Eq.(1.7.11). There are some drawbacks to this approach. It limits the properties of the system whose averages can be evaluated. Systems where interactions between particles depend on their momenta cannot be easily included. Moreover, averages of quantities that depend on both the momentum and the position of particles cannot be performed. However, if the energy separates into potential and kinetic energies as follows:

$$E(\{x,p\}) = V(\{x\}) + \sum_i p_i^2/2m \tag{1.7.13}$$

then averages over all quantities that just depend on momenta (such as the kinetic energy) can be evaluated directly without need for numerical computation. These averages are the same as those of an ideal gas. Monte Carlo simulations can then be used to perform the average over quantities that depend only upon position $\{x\}$, or more generally, on position-related variables s. Thus,in the remainder of this section we focus on describing Markov chains for systems described only by position-related variables s.

As described in Section 1.2 we can think about the Markov dynamics as a dynamics of the probability rather than the dynamics of a system. Then the dynamics are specified by

$$P_s(s';t) = \sum_{s'} P_s(s'|s'')P_s(s'';t-1) \tag{1.7.14}$$

In order for the stochastic dynamics to represent the ensemble, we must have the time average over the probability distribution $P_s(s',t)$ equal to the ensemble probability. This is true for a long enough time average if the probability converges to the ensemble probability distribution, which is a steady-state distribution of the Markov chain:

$$P_s(s') = P_s(s';\infty) = \sum_{s'} P_s(s'|s'')P_s(s'';\infty) \tag{1.7.15}$$

Thermodynamics and stochastic Markov chains meet when we construct the Markov chain so that the Boltzmann probability, Eq. (1.7.9), is the limiting distribution.

We now make use of the Perron-Frobenius theorem (see Section 1.7.4 below), which says that a Markov chain governed by a set of transition probabilities $P_s(s'|s'')$ converges to a unique limiting probability distribution as long as it is irreducible and acyclic. Irreducible means that there exist possible paths between each state and all other possible states of the system. This does not mean that all states of the system are connected by nonzero transition probabilities. There can be transition probabilities that are zero. However, it must be impossible to separate the states into two sets for which there are no transitions from one set to the other. Acyclic means that the system is not ballistic—the states are not organized by the transition matrix into a ring with a deterministic flow around it. There may be currents, but they must not be deterministic. It is sufficient for there to be a single state which has a nonzero probability of making a transition to itself for this condition to be satisfied,thus it is often assumed and unstated.

We can now summarize the problem of identifying the desired Markov chain. We must construct a matrix $P_s(s'|s'')$ that satisfies three properties.First,it must be an allowable transition matrix. This means that it must be nonnegative, $P_s(s''|s')\geq 0$, and satisfy the normalization condition (Eq (1.2.4)):

$$\sum_{s''} P_s(s''|s') = 1 \tag{1.7.16}$$

Second,it must have the desired probability distribution, Eq.(1.7.9),as a fixed point. Third, it must not be reducible—it is possible to construct a path between any two states of the system.

These conditions are sufficient to guarantee that a long enough Markov chain will be a good approximation to the desired ensemble. There is no guarantee that the convergence will be rapid. As we have seen in Section 1.4,in the case of the glass transition,the ergodic theorem may be violated on all practical time scales for systems that are following a particular dynamics. This applies to realistic or artificial dynamics. In general such violations of the ergodic theorem, or even just slow convergence of averages, are due to energy barriers or entropy “bottlenecks” that prevent the system from reaching all possible configurations of the system in any reasonable time. Such obstacles must be determined for each system that is studied, and are sometimes but not always apparent. It should be understood that different dynamics will satisfy the conditions of the ergodic theorem over very different time scales. The equivalence of results of an average performed using two distinct dynamics is only guaranteed ifthey are both simulated for long enough so that each satisfies the ergodic theorem.

Our discussion here also gives some additional insights into the conditions under which the ergodic theorem applies to the actual dynamics of physical systems. We note that any proof of the applicability of the ergodic theorem to a real system requires considering the actual dynamics rather than a model stochastic process. When the ergodic theorem does not apply to the actual dynamics, then the use of a Monte Carlo simulation for performing an average must be considered carefully. It will not give the same results if it satisfies the ergodic theorem while the real system does not.

We are still faced with the task of selecting values for the transition probabilities $P_s(s'|s'')$ that satisfy the three requirements given above. We can simplify our search for transition probabilities $P_s(s'|s'')$ for use in Monte Carlo simulations by imposing the additional constraint of microscopic reversibility, also known as detailed balance:

$$P_s(s''|s')P_s(s';\infty) = P_s(s'|s'')\, P_s(s'';\infty) \tag{1.7.17}$$

This equation implies that the transition currents between two states of the system are equal and therefore cancel in the steady state, Eq.(1.7.15). It corresponds to true equilibrium,as would be present in a physical system. Detailed balance implies the steady-state condition, but is not required by it.Steady state can also include currents that do not change in time. We can prove that Eq.(1.7.17) implies Eq. (1.7.15) by summing over s':

$$\sum_{s'} P_s(s''|s')P_s(s';\infty) = \sum_{s'} P_s(s'|s'')P_s(s'';\infty) = P_s(s'';\infty) \tag{1.7.18}$$

We do not yet have an explicit prescription for $P_s(s'|s'')$. There is still a tremendous flexibility in determining the transition probabilities.One prescription that enables direct implementation, called Metropolis Monte Carlo, is:

$$\begin{array}{lll} P_s(s'|s'') = \lambda(s'|s'') & P_s(s')/P_s(s'') \geq 1 & s'' \neq s' \\ P_s(s'|s'') = \lambda(s'|s'')P_s(s')/P_s(s'') & P_s(s')/P_s(s'') < 1 & s'' \neq s' \\ P_s(s''|s'') = 1 - \sum\limits_{s' \neq s'} P_s(s'|s'') & & \end{array} \tag{1.7.19}$$

These expressions specify the transition probability $P_s(s'|s'')$ in terms of a symmetric stochastic matrix $\lambda(s'|s'')$. $\lambda(s'|s'')$ is independent of the limiting equilibrium distribution. The constraint associated with the limiting distribution has been incorporated explicitly into Eq. (1.7.19). It satisfies detailed balance by direct substitution in Eq. (1.7.17), since for $P_s(s') \geq P_s(s'')$ (similarly for the opposite) we have

$$\begin{aligned} P_s(s''|s')P_s(s') &= \lambda(s''|s')P_s(s') = \lambda(s'|s'')P_s(s') \\ &= \left(\lambda(s'|s'')P_s(s')/P_s(s'')\right)P_s(s'') = P_s(s'|s'')P_s(s'') \end{aligned} \tag{1.7.20}$$

The symmetry of the matrix $\lambda(s'|s'')$ is essential to the proof of detailed balance. One must often be careful in the design of specific algorithms to ensure this property. It is also important to note that the limiting probability appears in Eq. (1.7.19) only in the form of a ratio $P_s(s')/P_s(s'')$ which can be given directly by the Boltzmann distribution.

To understand Metropolis Monte Carlo, it is helpful to describe a few examples. We first describe the movement of the system in terms of the underlying stochastic process specified by $\lambda(s'|s'')$, which is independent of the limiting distribution. This means that the limiting distribution of the underlying process is uniform over the whole space of possible states.

A standard way to choose the matrix $\lambda(s'|s'')$ is to set it to be constant for a few states s' that are near s''. For example, the simplest random walk is such a case, since it allows a probability of 1/2 for the system to move to the right and to the left. If s is a continuous variable,we could choose a distance r_0 and allow the walker to take a step anywhere within the distance r_0 with equal probability. Both the discrete and continuous random walk have d-dimensional analogs or, for a system of interacting particles, N-dimensional analogs. When there is more than one dimension,we can choose to move in all dimensions simultaneously. Alternatively, we can choose to move in only one of the dimensions in each step. For an Ising model (Section 1.6), we could allow equal probability for any one of the spins to flip.

Once we have specified the underlying stochastic process, we generate the sequence of Monte Carlo steps by applying it. However, we must modify the probabilities according to Eq. (1.7.19). This takes the form of choosing a step, but sometimes rejecting it rather than taking it. When a step is rejected,the system does not change its state. This gives rise to the third equation in Eq.(1.7.19) where the system does not move. Specifically, we can implement the Monte Carlo process according to the following prescription:

1. Pick one of the possible moves allowed by the underlying process. The selection is random from all of the possible moves. This guarantees that we are selecting it with the underlying probability $\lambda(s'|s'')$.

2. Calculate the ratio of probabilities between the location we are going to, compared to the location we are coming from

$$P_s(s'')\ /\ P_s(s') = e^{-(E(s')-E(s'))/kT} \tag{1.7.21}$$

If this ratio of probabilities is greater than one, which means the energy is lower where we are going, the step is accepted. This gives the probability for the process to occur as $\lambda(s'|s'')$, which agrees with the first line of Eq.(1.7.19). However, if this ratio is less than one, we accept it with a probability given by the ratio. For example,if the ratio is 0.6, we accept the move 60% of the time. If the move is rejected,the system stays in its original location. Thus,if the energy where we are trying to go is higher, we do not accept it all the time, only some of the time. The likelihood that we accept it decreases the higher the energy is.

The Metropolis Monte Carlo prescription makes logical sense. It tends to move the system to regions of lower energy. This must be the case in order for the final distribution to satisfy the Boltzmann probability. However, it also allows the system to climb up in energy so that it can reach, with a lower probability, states of higher energy. The ability to climb in energy also enables the system to get over barriers such as the one in the two-state system in Section 1.4.

For the Ising model, we can see that the Monte Carlo dynamics that uses all single spin flips as its underlying stochastic process is not the same as the Glauber dynamics (Section 1.6.7), but is similar. Both begin by selecting a particular spin. After selection of the spin, the Monte Carlo will set the spin to be the opposite with a probability:

$$\min(1,e^{-(E(1)-\mathrm{E}(-1))\,/\,kT}) \tag{1.7.22}$$

This means that if the energy is lower for the spin to flip, it is flipped. If it is higher, it may still flip with the indicated probability. This is different from the Glauber prescription, which sets the selected spin to UP or DOWN according to its equilibrium probability (Eq. (1.6.61)–Eq. (1.6.63)). The difference between the two schemes can be shown by plotting the probability of a selected spin being UP as a function of the energy difference between UP and DOWN, $E_+ = E(1) - E(-1)$ (Fig. 1.7.2). The Glauber dynamics prescription is independent of the starting value of the spin. The Metropolis Monte Carlo prescription is not. The latter causes more changes, since the spin is more likely to flip. Unlike the Monte Carlo prescription, the Glauber dynamics explicitly requires knowledge of the probabilities themselves. For a single spin flip in an Ising system this is fine, because there are only two possible states and the probabilities depend only on E_+. However, this is difficult to generalize when a system has many more possible states.

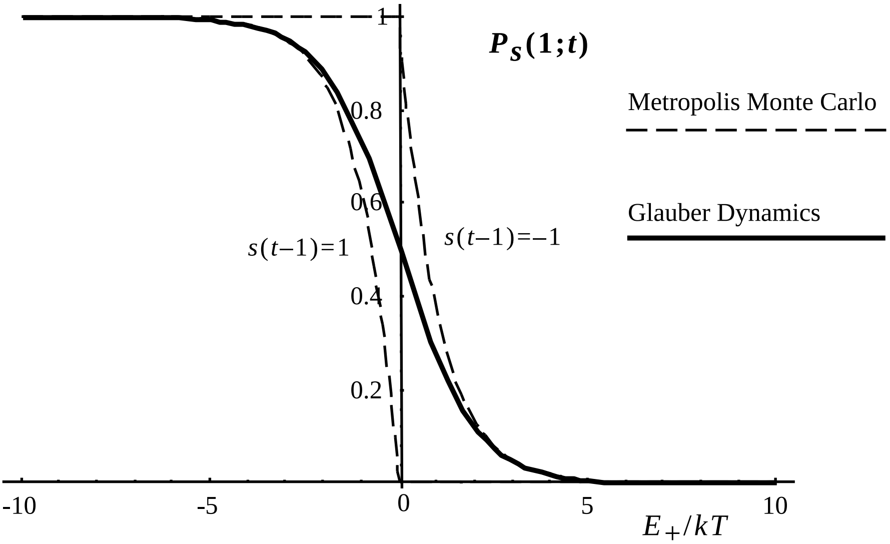


Figure 1.7.2 Illustration of the difference between Metropolis Monte Carlo and Glauber dynamics for the update of a spin in an Ising model. The plots show the probability $P_s(1;t)$ of a spin being UP at time t. The Glauber dynamics probability does not depend on the starting value of the spin. There are two curves for the Monte Carlo probability, for $s(t-1)=1$ and $s(t-1)=-1$. ▮

There is a way to generalize further the use of Monte Carlo by recognizing that we do not even have to use the correct equilibrium probability distribution when generating the time series. The generalized expression for an arbitrary probability distribution $P'(s)$ is:

$$< f(s) >_{P(s)} = \int \frac{f(s)P(s)}{P'(s)} P'(s)ds = \frac{1}{N} \sum_{s:P'(s)}^{N} \frac{f(s)P(s)}{P'(s)} \tag{1.7.23}$$

The subscript $P(s)$ indicates that the average assumes that s has the probability distribution $P(s)$ rather than $P'(s)$. This equation generalizes Eq. (1.7.5). The problem with this expression is that it requires that we know explicitly the probabilities $P(s)$ and $P'(s)$. This can be remedied. We illustrate for a specific case, where we use the Boltzmann distribution at one temperature to evaluate the average at another temperature:

$$< f(s) >_{P(s)} = \frac{1}{N} \sum_{s:P'(s)}^{N} \frac{f(s)P(s)}{P'(s)} = \left(\frac{Z'}{Z}\right) \frac{1}{N} \sum_{s:P'(s)}^{N} f(s)e^{-E(s)(1/kT-1/kT')} \tag{1.7.24}$$

The ratio of partition functions can be directly evaluated as an average:

$$\left(\frac{Z}{Z'}\right) = \frac{\sum_{s} e^{-E(s)/kT}}{\sum_{s} e^{-E(s)/kT'}} = \frac{\sum_{s} e^{-E(s)(1/kT-1/kT')} e^{-E(s)/kT'}}{\sum_{s} e^{-E(s)/kT'}}$$

$$= \left\langle e^{-E(s)(1/kT-1/kT')} \right\rangle_{P'(s)} = \frac{1}{N} \sum_{s:P'(s)}^{N} e^{-E(s)(1/kT-1/kT')} \tag{1.7.25}$$

Thus we have the expression:

$$< f(s) >_{P(s)} = \sum_{s:P'(s)}^{N} f(s) e^{-E(s)(1/kT-1/kT')} \Bigg/ \sum_{s:P'(s)}^{N} e^{-E(s)(1/kT-1/kT')} \tag{1.7.26}$$

This means that we can obtain the average at various temperatures using only a single Monte Carlo simulation. However, the whole point of using the ensemble average is to ensure that the average converges rapidly. This may not happen if the ensemble temperature T' is much different from the temperature T. On the other hand, there are circumstances where the function $f(s)$ may have an energy dependence that makes it better to perform the average using an ensemble that is not the equilibrium ensemble.

The approach of Monte Carlo simulations to the study of statistical averages ensures that we do not have to be concerned that the dynamics we are using for the system is a real dynamics. The result is the same for a broad class of artificial dynamics. The generality provides a great flexibility; however, this is also a limitation. We cannot use the Monte Carlo dynamics to study dynamics. We can only use it to perform statistical averages. Must we be resigned to this limitation? The answer, at least in part, is no. The reason is rooted in the central limit theorem. For example,the implementations of Metropolis Monte Carlo and the Glauber dynamics are quite different. We know that in the limit of long enough times, the distribution of configurations generated by both is the same. We expect that since each of them flips only one spin,if we are interested in changes in many spins,the two should give comparable results in the sense of the central limit theorem. This means that aside from an overall scale factor, the time evolution of the distribution of probabilities for long times is the same. Since we already know that the limiting distribution is the same in both cases, we are asserting that the approach to this limiting distribution, which is the long time dynamics, is the same.

The claim that for a large number of steps all dynamics is the same is not true about all possible Monte Carlo dynamics. If we allow all of the spins in an Ising model to change their values in one step of the underlying dynamics $\lambda(s'|s'')$, then this step would be equivalent to many steps in a dynamics that allows only one spin to flip at a time. In order for two different dynamics to give the same results,there are two types of constraints that are necessary. First, both must have similar kinds of allowed steps. Specifically, we define steps to the naturally proximate configurations as local moves. As long as the Monte Carlo allows only local moves, the long time dynamics should be the same. Such dynamics correspond to a local diffusion in the space of possible configurations of the system. More generally, two different dynamics should be the same if configuration changes that require many steps in one also require many steps in the other. The second type of constraint is related to symmetries of the problem. A lack of bias in the random walk was necessary to guarantee that the Gaussian distribution resulted from a generalized random walk in Section 1.2. For systems with more than one dimension, we must also ensure that there is no relative bias between motion in different directions.

We can think about Monte Carlo dynamics as diffusive dynamics of a system that interacts frequently with a reservoir. There are properties of more realistic dynamics that are not reproduced by such configuration Monte Carlo simulations. Correlations between steps are not incorporated because of the assumptions underlying Markov chains. This rules out ballistic motion,and exact or approximate momentum conservation. Momentum conservation can be included if both position and momentum are included as system coordinates. The method called Brownian dynamics incorporates both ballistic and diffusive dynamics in the same simulation. However, if correlations in the dynamics of a system have a shorter range than the motion we are interested in, momentum conservation may not matter to results that are of interest, and conventional Monte Carlo simulations can be used directly.

In summary, Monte Carlo simulations are designed to reproduce an ensemble rather than the dynamics of a particular system. As such,they are ideally suited to investigating the equilibrium properties of thermodynamic systems. However, Monte Carlo dynamics with local moves often mimic the dynamics of real systems. Thus, Monte Carlo simulations may be used to investigate the dynamics of systems when they are appropriately designed. This property will be used in Chapter 5 to simulate the dynamics of long polymers.

There is a flip side to the design of Monte Carlo dynamics to simulate actual dynamics. If our objective is the traditional objective of a Monte Carlo simulation, of obtaining an ensemble average, then the ability to simulate dynamics may not be an advantage. In some systems, the real dynamics is slow and we would prefer to speed up the process. This can often be done by knowingly introducing nonlocal moves that displace the state of the system by large distances in the space of conformations. Such nonlocal Monte Carlo dynamics have been designed for various systems. In particular, both local and nonlocal Monte Carlo dynamics for the problem of polymer dynamics will be described in Chapter 5.

Question 1.7.1 In order to perform Monte Carlo simulations, we must be able to choose steps at random and accept or reject steps with a certain probability. These operations require the availability of random numbers. We might think of the source of these random numbers as a thermal reservoir. Computers are specifically deigned to be completely deterministic. This means that inherently there is no randomness in their operation. To obtain random numbers in a computer simulation requires a deterministic algorithm that generates a sequence of numbers that look random but are not random. Such sequences are called pseudo-random numbers. Random numbers should not be correlated to each other. However, using pseudo-random numbers, if we start a program over again we must get exactly the same sequence of numbers. The difficulties associated with the generation of random numbers are central to performing Monte Carlo computer simulations. If we assume that we have random numbers, and they are not really uncorrelated, then our results may very well be incorrect. Nevertheless, pseudo-random numbers often give results that are consistent with those expected from random numbers.

There are a variety of techniques to generate pseudo-random numbers. Many of these pseudo-random number generators are designed to provide, with equal "probability," an integer between 0 and the maximal integer possible. The maximum integer used by a particular routine on a particular machine should be checked before using it in a simulation. Some use a standard short integer which is represented by 16 bits (2 bytes).One bit represents the unused sign of the integer. This leaves 15 bits for the magnitude of the number. The pseudo-random number thus ranges up to $2^{15} - 1 = 32767$. An example of a routine that provides pseudo-random integers is the subroutine `rand()` in the ANSI C library, which is executed using a line such as:

$$k = \texttt{rand()}; \tag{1.7.27}$$

The following three questions discuss how to use such a pseudo-random number generator. Assume that it provides a standard short integer.

1. Explain how to use a pseudo-random number generator to choose a move in a Metropolis Monte Carlo simulation, Eq. (1.7.19).
2. Explain how to use a pseudo-random number generator to accept or reject a move in a Metropolis Monte Carlo simulation, Eq. (1.7.19).
3. Explain how to use a pseudo-random number generator to provide values of x with a probability $P(x)$ for x in the interval [0,1]. Hint: Use two pseudo-random numbers every step.

Solution 1.7.1

1. Given the necessity of choosing one out of M possible moves, we create a one-to-one mapping between the M moves and the integers $\{0, \ldots, M - 1\}$ If M is smaller than 2^{15} we can use the value of $k = \texttt{rand()}$ to determine which move is taken next. If k is larger than $M - 1$, we don't make any move. If M is much smaller than 2^{15} then we can use only some of the bits of k. This avoids making many unused calls to `rand()`. Fewer bits can be obtained using a modulo operation. For example, if $M = 10$ we might use k modulo 16. We could also ignore values above 32759,and use k modulo 10. This also causes each move to occur with equal frequency. However, a standard word of caution about using only a few bits is that we shouldn't use the lowest order bits (i.e., the units, twos and fours bits), because they tend to be more correlated than the

higher order bits. Thus it may be best first to divide k by a small number, like 8 (or equivalently to shift the bits to the right),if it is desired to use fewer bits. If M is larger than 2^{15} it is necessary to use more than one call to `rand()` (or a random number generator that provides a 4-byte integer) so that all possible moves are accounted for.

2. Given the necessity of determining whether to accept a move with the probability P, we compare 2^{15} P with a number given by k = `rand()`. If the former is bigger we accept the move, and if it is smaller we reject the move.
3. One way to do this is to generate two random numbers r_1 and r_2. Dividing both by 32767 (or 2^{15}), we use the first random number to be the location in the interval $x = r_1/32767$. However, we use this location only if the second random number $r_2/32767$ is smaller than $P(x)$. If the random number is not used, we generate two more and proceed. This means that we will use the position x with a probability $P(x)$ as desired. Because it is necessary to generate many random numbers that are rejected, this method for generating numbers for use in performing the integral Eq. (1.7.3) is only useful if evaluations of the function $f(x)$ are much more costly than random number generation. ▮

Question 1.7.2 To compare the errors that arise from conventional numerical integration and Monte Carlo sampling, we return to Eq.(1.7.4) and Eq. (1.7.5) in this and the following question. We choose two integrals that can be evaluated analytically and for which the errors can also be evaluated analytically.

Evaluate two examples of the integral $\int P(x)f(x)dx$ over the interval x [1,1]. For the first example (1) take $f(x) = 1$, and for the second (2) $f(x) = x$. In both cases assume the probability distribution is an exponential

$$P(x) = Ae^{-\lambda x} = \left[\frac{\lambda}{e^{\lambda} - e^{-\lambda}}\right]e^{-\lambda x} \qquad (1.7.28)$$

where the normalization constant A is given by the expression in square brackets.

Calculate the two integrals exactly (analytically). Then evaluate approximations to the integrals using sums over N equally spaced points, Eq.(1.7.4). These sums can also be evaluated analytically. To improve the result of the sum, you can use Simpson's rule. This modifies Eq.(1.7.4) only by subtracting 1/2 of the value of the integrand at the first and last points. The errors in evaluation of the same integral by Monte Carlo simulation are to be calculated in Question 1.7.3.

Solution 1.7.2

1. The value of the integral of $P(x)$ is unity as required by normalization. If we use a sum over equally spaced points we would have:

$$A\int_{-1}^{1}dxe^{-\lambda x}\approx\frac{A}{M}\sum_{n=-M}^{M}e^{-\lambda(n/M)}=\frac{A}{M}\sum_{n=-M}^{M}a^{n}\tag{1.7.29}$$

where we used the temporary definition $a = e^{-\lambda/M}$ to obtain

$$A\int_{-1}^{1}dxe^{-\lambda x}\approx\frac{A}{M}\frac{(a^{M+1}-a^{-M})}{a-1}=\frac{A}{M}\frac{(e^{\lambda}e^{\lambda/M}-e^{-\lambda})}{e^{\lambda/M}-1}\tag{1.7.30}$$

Expanding the answer in powers of λ/M gives:

$$\begin{aligned}A\int_{-1}^{1}dxe^{-\lambda x}&\approx A\frac{(e^{\lambda}-e^{-\lambda})}{\lambda}+A\frac{(e^{\lambda}+e^{-\lambda})}{2M}+A\frac{\lambda(e^{\lambda}-e^{-\lambda})}{2M^{2}}+\\&=1+\frac{\lambda}{2M}\tanh(\lambda)+\frac{\lambda^{2}}{2M^{2}}+\end{aligned}\tag{1.7.31}$$

The second term can be eliminated by noting that the sum could be evaluated using Simpson's rule by subtracting 1/2 of the contribution of the end points. Then the third term gives an error of $\lambda^2/2M^2$. This is the error in the numerical approximation to the average of $f(x) = 1$.

2. For $f(x) = x$ the exact integral is:

$$\begin{aligned}A\int_{-1}^{1}dx\,xe^{-\lambda x}&=-A\frac{d}{d\lambda}\int_{-1}^{1}dxe^{-\lambda x}=-A\frac{d}{d\lambda}\left[\frac{e^{\lambda}-e^{-\lambda}}{\lambda}\right]\\&=-\coth(\lambda)+(1/\lambda)\end{aligned}\tag{1.7.32}$$

while the sum is:

$$\begin{aligned}A\int_{-1}^{1}dx\,xe^{-\lambda x}&\approx\frac{A}{M^{2}}\sum_{n=-M}^{M}ne^{-\lambda(n/M)}=\frac{A}{M^{2}}\sum_{n=-M}^{M}na^{n}\\&=\frac{A}{M^{2}}a\frac{d}{da}\sum_{n=-M}^{M}a^{n}=\frac{A}{M^{2}}a\frac{d}{da}\frac{(a^{M+1}-a^{-M})}{a-1}\\&=\frac{A}{M^{2}}\frac{((M+1)a^{M+1}+Ma^{-M})}{a-1}+\frac{A}{M^{2}}\frac{a(a^{M+1}-a^{-M})}{(a-1)^{2}}\\&=\frac{A}{M}\frac{(e^{\lambda}e^{\lambda/M}+e^{-\lambda})}{e^{\lambda/M}-1}+\frac{A}{M^{2}}\frac{(e^{\lambda}e^{\lambda/M})}{e^{\lambda/M}-1}+\frac{A}{M^{2}}\frac{e^{\lambda/M}(e^{\lambda}e^{\lambda/M}-e^{-\lambda})}{(e^{\lambda/M}-1)^{2}}\end{aligned}\tag{1.7.33}$$

With some assistance from Mathematica,the expansion to second order in λ/M is:

$$\begin{aligned}&=-A\frac{(e^{\lambda}-e^{-\lambda})}{\lambda^{2}}+A\frac{(e^{\lambda}+e^{-\lambda})}{\lambda}+\frac{A}{M}\frac{(e^{\lambda}-e^{-\lambda})}{2}+\frac{A}{M^{2}}\frac{11}{12}(e^{\lambda}-e^{-\lambda})+\\&=-1/\lambda+\coth(\lambda)+\frac{\lambda}{2M}+\frac{11}{12}\frac{\lambda}{M^{2}}+\end{aligned}\tag{1.7.34}$$

The first two terms are the correct result. The third term can be seen to be eliminated using Simpson's rule. The fourth term is the error. ▮

Question 1.7.3 Estimate the errors in performing the same integrals as in Question 1.7.2 using a Monte Carlo ensemble sampling with N terms as in Eq.(1.7.5). It is not necessary to evaluate the integrals to evaluate the errors.

Solution 1.7.3

1. The errors in performing the integral for $f(x) = 1$ are zero, since the Monte Carlo sampling would be given by the expression:

$$< 1 >_{P(s)} = \frac{1}{N}\sum_{s:P(s)}^{N} 1 = 1 \tag{1.7.35}$$

One way to think about this result is that Monte Carlo takes advantage of the normalization of the probability, which the technique of summing the integrand over equally spaced points cannot do. This knowledge makes this integral trivial, but it is also of use in performing other integrals.

2. To evaluate the error for the integral over $f(x) = x$ we use an argument based on the sampling error of different regions of the integral. We break up the domain $[-1,1]$ into q regions of size $\Delta x = 2/q$. Each region is assumed to have a significant number of samples. The number of these samples is approximately given by:

$$NP(x)\Delta x \tag{1.7.36}$$

If this were the exact number of samples as q increased,then the integral would be exact. However, since we are picking the points at random, there will be a deviation in the number of these from this ideal value. The typical deviation, according to the discussion in Section 1.2 of random walks, is the square root of this number. Thus the error in the sum

$$\sum_{s:P(s)}^{N} f(x) \tag{1.7.37}$$

from a particular interval Δx is

$$(NP(x)\Delta x)^{1/2} f(x) \tag{1.7.38}$$

Since this error could have either a positive or negative sign, we must take the square root of the sum of the squares of the error in each region to give us the total error:

$$\left|\int P(x)f(x) - \frac{1}{N}\sum_{s:P(s)}^{N} f(x)\right| \approx \frac{1}{N}\sqrt{\sum NP(x)\Delta x\, f(x)^2} \approx \frac{1}{\sqrt{N}}\sqrt{\int P(x)f(x)^2} \tag{1.7.39}$$

For $f(x) = x$ the integral in the square root is:

$$\int Ae^{-\lambda x} f(x)^2 dx = \int Ae^{-\lambda x} x^2 dx = A\frac{d^2}{d\lambda^2}\frac{(e^{\lambda} - e^{-\lambda})}{\lambda} = \frac{2}{\lambda^2} - \frac{2\coth(\lambda)}{\lambda} + 1 \tag{1.7.40}$$

The approach of Monte Carlo is useful when the exponential is rapidly decaying. In this case, $\lambda >> 1$, and we keep only the third term and have an error that is just of magnitude $1/\sqrt{N}$. Comparing with the sum over equally spaced points from Question 1.7.2, we see that the error in Monte Carlo is independent of λ for large λ, while it grows for the sum over equally spaced points. This is the crucial advantage of the Monte Carlo method. However, for a fixed value of λ we also see that the error is more slowly decreasing with N than the sum over equally spaced points. So when a large number of samples is possible, the sum over equally spaced points is more rapidly convergent. ▮

Question 1.7.4 How would the discrete nature of the integer random numbers described in Question 1.7.1 affect the ensemble sampling? Answer qualitatively. Is there a limit to the accuracy of the integral in this case?

Solution 1.7.4 The integer random numbers introduce two additional sources of error, one due to the sampling interval along the x axis and the other due to the imperfect approximation of $P(x)$. In the limit of a large number of samples, each of the possible values along the x axis would be sampled equally. Thus, the ensemble sum would reduce to a sum of the integrand over equally spaced points. The number of points is given by the largest integer used (e.g., 2^{15}). This limits the accuracy accordingly. ▮

1.7.3 *Perron-Frobenius theorem*

The Perron-Frobenius theorem is tied to our understanding of the ergodic theorem and the use of Monte Carlo simulations for the representation of ensemble averages. The theorem only applies to a system with a finite space of possible states. It says that a transition matrix that is irreducible must ultimately lead to a stable limiting probability distribution. This distribution is unique, and thus depends only on the transition matrix and not on the initial conditions. The Perron-Frobenius theorem assumes an irreducible matrix,so that starting from any state,there is some path by which it is possible to reach every other state of the system. If this is not the case,then the theorem can be applied to each subset of states whose transition matrix is irreducible.

In a more general form than we will discuss,the Perron-Frobenius theorem deals with the effect of matrix multiplication when all of the elements of a matrix are positive. We will consider it only for the case of a transition matrix in a Markov chain, which also satisfies the normalization condition, Eq. (1.7.16). In this case, the proof of the Perron-Frobenius theorem follows from the statement that there cannot be any eigenvalues of the transition matrix that are larger than one. Otherwise there would be a vector that would increase everywhere upon matrix multiplication. This is not possible, because probability is conserved. Thus if the probability increases in one place it must decrease someplace else, and tend toward the limiting distribution.

A difficulty in the proof of the theorem arises from dealing with the case in which there are deterministic currents through the system: e.g., ballistic motion in a circular path. An example for a two-state system would be

$$\begin{aligned} &P(1|1) = 0 \quad P(1|-1) = 1 \\ &P(-1|1) = 1 \; P(-1|-1) = 0 \end{aligned} \tag{1.7.41}$$

In this case, a system in the state $s = +1$, goes into $s = -1$, and a system in the state s $= -1$ goes into $s = +1$. The limiting behavior of this Markov chain is of two probabilities that alternate in position without ever settling down into a limiting distribution. An example with three states would be

$$\begin{array}{lll} P(1|1) = 0 & P(1|2) = 1 & P(1|3) = 1 \\ P(2|1) = .5 & P(2|2) = 0 & P(2|3) = 0 \\ P(3|1) = .5 & P(3|2) = 0 & P(3|3) = 0 \end{array} \tag{1.7.42}$$

Half of the systems with $s = 1$ make transitions to $s = 2$ and half to $s = 3$. All systems with $s = 2$ and $s = 3$ make transitions to $s = 1$. In this case there is also a cyclical behavior that does not disappear over time. These examples are special cases, and the proof shows that they are special. It is sufficient, for example, for there to be a single state where there is some possibility of staying in the same state. Once this is true, these examples of cyclic currents do not apply and the system will settle down into a limiting distribution.

We will prove the Perron-Frobenius theorem in a few steps enumerated below. The proof is provided for completeness and reference, and can be skipped without significant loss for the purposes of this book. The proof relies upon properties of the eigenvectors and eigenvalues of the transition matrix. The eigenvectors need not always be positive, real or satisfy the normalization condition that is usually applied to probability distributions, $P(s)$. Thus we use $v(s)$ to indicate complex vectors that have a value at every possible state of the system.

Given an irreducible real nonnegative matrix ($P(s'|s) \geq 0$) satisfying

$$\sum_{s'} P(s'|s) = 1 \tag{1.7.43}$$

we have:

1. Applying $P(s'|s)$ cannot increase the value of all elements of a nonnegative vector, $v(s') \geq 0$:

$$\min_{s'} \left(\frac{1}{v(s')} \sum_{s} P(s'|s) v(s) \right) \leq 1 \tag{1.7.44}$$

To avoid infinities, we can assume that the minimization only includes s' such that $v(s') \neq 0$.

Proof: Assume that Eq. (1.7.44) is not true. In this case

$$\sum_s P(s'|s)v(s) > v(s') \tag{1.7.45}$$

for all $v(s') \neq 0$, which implies

$$\sum_{s'} \sum_s P(s'|s)v(s) > \sum_{s'} v(s') \tag{1.7.46}$$

Using Eq. (1.7.43), the left is the same as the right and the inequality is impossible.

2. The magnitude of eigenvalues of $P(s'|s)$ is not greater than one.

Proof: Let $v(s)$ be an eigenvector of $P(s'|s)$ with eigenvalue λ:

$$\sum_s P(s'|s)v(s) = \lambda v(s') \tag{1.7.47}$$

Then:

$$\sum_s P(s'|s)\left|v(s)\right| \geq \left|\lambda\right|\left|v(s')\right| \tag{1.7.48}$$

This inequality follows because each term in the sum on the left has been made positive. If all terms started with the same phase, then equality holds. Otherwise, inequality holds. Comparing Eq. (1.7.48) with Eq. (1.7.44), we see that $|\lambda| \leq 1$.

If $|\lambda| = 1$,then equality must hold in Eq.(1.7.48), and this implies that $|v(s)|$, the vector whose elements are the magnitudes of $v(s)$,is an eigenvector with eigenvalue 1. Steps 3–5 show that there is one such vector which is strictly positive (greater than zero) everywhere.

3. $P(s'|s)$ has an eigenvector with eigenvalue $\lambda = 1$. We use the notation $v_1(s)$ for this vector.

Proof: The existence of such an eigenvector follows from the existence of an eigenvector of the transpose matrix with eigenvalue $\lambda = 1$. Eq.(1.7.43) implies that the vector $v(s) = 1$ (one everywhere) is an eigenvector of the transpose matrix with eigenvalue $\lambda = 1$. Thus $v_1(s)$ exists, and by step 2 we can take it to be real and nonnegative, $v_1(s) \geq 0$. We can, however, assume more, as the following shows.

4. An eigenvector of $P(s'|s)$ with eigenvalue 1 must be strictly positive, $v_1(s) > 0$.

Proof: Define a new Markov chain given by the transition matrix

$$Q(s'|s) = (P(s'|s) + \delta_{s,s'}) / 2 \tag{1.7.49}$$

Applying $Q(s'|s)$ $N - 1$ times to any vector $v_1(s) \geq 0$ must yield a vector that is strictly positive. This follows because $P(s'|s)$ is irreducible. Starting with unit probability at any one value of s, after $N - 1$ steps we will move some probability everywhere. Also, by the construction of $Q(s'|s)$, any s which has a nonzero probability at one time will continue to have a nonzero probability at all later times. By linear superposition,this applies to any initial probability distribution. It also applies to any unnormalized vector $v_1(s) \geq 0$. Moreover, if $v_1(s)$ is an eigenvector of $P(s'|s)$ with eigenvalue one,then it is also an eigenvector of $Q(s'|s)$ with the same eigenvalue. Since applying $Q(s'|s)$ to $v_1(s)$ changes nothing, applying it $N - 1$ times also changes nothing. We have just proven that $v_1(s)$ must be strictly positive.

5. There is only one linearly independent eigenvector of $P(s'|s)$ with eigenvalue $\lambda = 1$.

Proof: Assume there are two such eigenvectors: $v_1(s)$ and $v_2(s)$. Then we can make a linear combination $c_1v_1(s) + c_2v_2(s)$,so that at least one of the elements is zero and others are positive. This linear combination is also an eigenvector of $P(s'|s)$ with eigenvalue $\lambda = 1$, which violates step 4. Thus there is exactly one eigenvector of $P(s'|s)$ with eigenvalue $\lambda = 1$, $v_1(s)$:

$$\sum_s P(s'|s)v_1(s) = v_1(s') \tag{1.7.50}$$

6. Either $P(s'|s)$ has only one eigenvalue with $|\lambda| = 1$ (in which case $\lambda = 1$), or it can be written as a cyclical flow.

Proof: Steps 2 and 5 imply that all eigenvectors of $P(s'|s)$ with eigenvalues λ satisfying $|\lambda| = 1$ can be written as:

$$v_i(s) = D_i(s)v_1(s) = e^{i\phi_i(s)}v_1(s) \tag{1.7.51}$$

As indicated, $D_i(s)$ is a vector with elements of magnitude one, $|D_i(s)| = 1$. We can write

$$\sum_s P(s'|s)D_i(s)v_1(s) = \lambda_i D_i(s')v_1(s') \tag{1.7.52}$$

There cannot be any terms in the sum on the left of Eq.(1.7.52) that add terms of different phase. If there were, then we would have a smaller magnitude than adding the absolute values, which would not agree with Eq.(1.7.50). Thus we can assign all of the elements of $D_i(s)$ into groups that have the same phase. $P(s'|s)$ cannot allow transitions to occur from any two of these groups into the same group. Since $P(s'|s)$ is irreducible, the only remaining possibility is that the different groups are connected in a ring with the first mapped onto the second, and the second mapped onto the third, and so on until we return to the first group. In particular, if there are any transitions between a site and itself this would violate the requirements and we could have no complex eigenvalues.

7. A Markov chain governed by an irreducible transition matrix, which has only one eigenvector, $v_1(s)$ with $|\lambda| = 1$,has a limiting distribution over long enough times which is proportional to this eigenvector. Using $P^t(s'|s)$ to represent the effect of applying $P(s'|s)$ t times, we must prove that:

$$\lim_{t\to\infty} v(s;t) = \lim_{t\to\infty} \sum_s P^t(s'|s)v(s) = cv_1(s') \tag{1.7.53}$$

for $v(s) \geq 0$. The coefficient c depends on the normalization of $v(s)$ and $v_1(s)$. If both are normalized so that the total probability is one, then conservation of probability implies that $c = 1$.

Proof: We write the matrix $P(s'|s)$ in the Jordan normal form using a similarity transformation. In matrix notation:

$$\mathbf{P} = \mathbf{S}^{-1}\mathrm{J}\mathbf{S} \tag{1.7.54}$$

J consists of a block diagonal matrix.Each of the block matrices along the diagonal is of the form

$$\mathbf{N} = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & \ddots & 0 \\ 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \tag{1.7.55}$$

where λ is an eigenvalue of **P**. In this block the only nonzero elements are λs on the diagonal, and 1s just above the diagonal.

Since $\mathbf{P}^t = \mathbf{S}^{-1}\mathbf{J}^t\mathbf{S}$, we consider $\mathbf{J}^t$, which consists of diagonal blocks $\mathbf{N}^t$. We prove that $\mathbf{N}^t \to 0$ as $t \to \infty$ for $|\lambda| < 1$. This can be shown by evaluating explicitly the matrix elements. The qth element above the diagonal of $\mathbf{N}^t$ is:

$$\lambda^{t-q}\binom{t}{q} \tag{1.7.56}$$

which vanishes as $t \to \infty$.

Since 1 is an eigenvalue with only one eigenvector, there must be one 1 × 1 block along the diagonal of **J** for the eigenvalue 1. Then $\mathbf{J}^t$ as $t \to \infty$ has only one nonzero element which is a 1 on the diagonal. Eq.(1.7.53) follows, because applying the matrix $\mathbf{P}^t$ always results in the unique column of $\mathbf{S}^{-1}$ that corresponds to the nonzero diagonal element of $\mathbf{J}^t$. By our assumptions,this column must be proportional to $v_1(s)$. This completes our proof and discussion of the Perron-Frobenius theorem.

1.7.4 *Minimization*

At low temperatures, a thermodynamic system in equilibrium will be found in its minimum energy configuration. For this and other reasons,it is often useful to identify the minimum energy configuration of a system without describing the full ensemble. There are also many other problems that can be formulated as minimization or optimization problems.

Minimization problems are often described in a d-dimensional space of continuous variables. When there is only a single valley in the parameter space of the problem,there are a variety of techniques that can be used to obtain this minimum. They may be classified into direct search and gradient-based techniques. In this section we focus on the single-valley problem. In Section 1.7.5 we will discuss what happens when there is more than one valley.

Direct search techniques involve evaluating the energy at various locations and closing in on the minimum energy. In one dimension, search techniques can be very effective. The key to a search is bracketing the minimum energy. Then each energy evaluation is used to geometrically shrink the possible domain of the minimum.

We start in one dimension by looking at the energy at two positions s_1 and s_2 that are near each other. If the left of the two positions s_1 is higher in energy $E(s_1) > E(s_2)$, then the minimum must be to its right. This follows from our assumption that there is only a single valley—the energy rises monotonically away from the minimum and therefore cannot be lower than $E(s_2)$,anywhere to the left of s_1. Evaluating the energy at a third location s_3 to the right of s_2 further restricts the possible locations of the minimum. If $E(s_3)$ is also greater than the middle energy location $E(s_3) > E(s_2)$, then the minimum must lie between s_1 and s_3. Thus, we have successfully bracketed the minimum. Otherwise, we have that $E(s_3) < E(s_2)$, and the minimum must lie to the right of s_2. In this case we look at the energy at a location s_4 to the right of s_3. This process is continued until the energy minimum is bracketed. To avoid taking many steps to the right,the size of the steps to the right can be taken to be an increasing geometric series, or may be based on an extrapolation of the function using the values that are available.

Once the energy minimum is bracketed, the segment is bisected again and again to locate the energy minimum. This is an iterative process. We describe a simple version of this process that can be easily implemented. An iteration begins with three locations $s_1 < s_2 < s_3$. The values of the energy at these locations satisfy $E(s_1)$, $E(s_3) > E(s_2)$. Thus the minimum is between s_1 and s_3. We choose a new location s_4, which in even steps is $s_4 = (s_1 + s_2) / 2$ and in odd steps is $s_4 = (s_2 + s_3)/2$. Then we eliminate either s_1 or s_3. The one that is eliminated is the one next to s_2 if $E(s_2) > E(s_4)$, or the one next to s_4 if $E(s_2) < E(s_4)$. The remaining three locations are relabled to be s_1, s_2, s_3 for the next step. Iterations stop when the distance between s_1 and s_3 is smaller than an error tolerance which is set in advance. More sophisticated versions of this algorithm use improved methods for selecting s_4 that accelerate the convergence.

In higher-dimension spaces,direct search can be used. However, mapping a multidimensional energy surface is much more difficult. Moreover, the exact logic that enables an energy minimum to be bracketed within a particular domain in one dimension is not possible in higher-dimension spaces. Thus, techniques that make use of a gradient of the function are typically used even if the gradient must be numerically evaluated. The most common gradient-based minimization techniques include steepest descent, second order and conjugate gradient.

Steepest descent techniques involve taking steps in the direction of the most rapid descent direction as determined by the gradient of the energy. This is the same as using a first-order expansion of the energy to determine the direction of motion toward lower energy. Illustrating first in one dimension,we start from a position s_1 and write the expansion as:

$$E(s) = E(s_1) + (s - s_1)\left.\frac{dE(s)}{ds}\right|_{s_1} + O((s - s_1)^2) \qquad (1.7.57)$$

We now take a step in the direction of the minimum by setting:

$$s_2 = s_1 - c\left.\frac{dE(s)}{ds}\right|_{s_1} \qquad (1.7.58)$$

From the expansion we see that for small enough $c, E(s_2)$ must be smaller than $E(s_1)$. The problem is to carefully select c so that we do not go too far. If we go too far we may reach beyond the energy minimum and increase the energy. We also do not want to make such a small step that many steps will be needed to reach the minimum. We can think of the sequence of configurations we pick as a time sequence, and the process we use to pick the next location as an iterative map. Then the minimum energy configuration is a fixed point of the iterative map given by Eq.(1.7.58). From a point near to the minimum we can have all of the behaviors described in Section 1.1—stable (converging) and unstable (diverging), both of these with or without alternation from side to side of the minimum.Of particular relevance is the discussion in Question 1.1.12 that suggests how c may be chosen to stabilize the iterative map and obtain rapid convergence.

When s is a multidimensional variable, Eq. (1.7.57) and Eq. (1.7.58) both continue to apply as long as the derivative is replaced by the gradient:

$$E(s) = E(s_1) + (s - s_1)\cdot \;\; {}_sE(s)\Big|_{s_1} + O((s - s_1)^2) \qquad (1.7.59)$$

$$s_2 = s_1 - c \;\; {}_sE(s)\Big|_{s_1} \qquad (1.7.60)$$

Since the direction opposite to the gradient is the direction in which the energy decreases most rapidly, this is known as a steepest descent technique. For the multidimensional case it is more difficult to choose a consistent value of c, since the behavior of the function may not be the same in different directions. The value of c may be chosen "on the fly" by making sure that the new energy is smaller than the old. If the current value of c gives a value $E(s_2)$ which is larger than $E(s_1)$ then c is reduced. We can improve upon this by looking along the direction of the gradient and considering the energy to be a function of c:

$$E(s_1 - c \;\; {}_sE(s)\Big|_{s_1}) \qquad (1.7.61)$$

Then c can be chosen by finding the actual minimum in this direction using the search technique that works well in one dimension.

Gradient techniques work well when different directions in the energy have the same behavior in the vicinity of the minimum energy. This means that the second derivative in different directions is approximately the same. If the second derivatives are very different in different directions,then the gradient technique tends to bounce back and forth perpendicular to the direction in which the second derivative is very small, without making much progress toward the minimum (Fig. 1.7.3). Improvements of the gradient technique fall into two classes. One class of techniques makes direct use of the second derivatives, the other does not. If we expand the energy to second order at the present best guess for the minimum energy location s_1 we have

$$E(s)=E(s_1)+(s-s_1)\;\; {}_sE(s)\Big|_{s_1}+(s-s_1)\;\; \overset{\leftarrow}{{}_s}\;\overset{\rightarrow}{{}_s}E(s)\Big|_{s_1}(s-s_1)+O((s-s_1)^3)$$

$$(1.7.62)$$

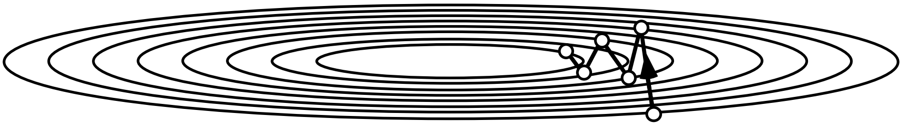


Figure 1.7.3 Illustration of the difficulties in finding a minimum energy by steepest descent when the second derivative is very different in different directions. The steps tend to oscillate and do not make progress toward the minimum along the flat direction. ▮

Setting the gradient of this expression to zero gives the next approximation for the minimum energy location s_2 as:

$$s_2 = s_1 - \frac{1}{2}\left(\vec{\ }_s \; {}_s E(s)\Big|_{s_1} \right)^{-1} \; {}_s E(s)\Big|_{s_1} \qquad (1.7.63)$$

This, in effect, gives a better description of the value of c for Eq. 1.7.60, which turns out to be a matrix inversely related to the second-order derivatives. Steps are large in directions in which the second derivative is small. If the second derivatives are not easily available, approximate second derivatives are used that may be improved upon as the minimization is being performed. Because of the need to evaluate the matrix of second-order derivatives and invert the matrix,this approach is not often convenient. In addition, the use of second derivatives assumes that the expansion is valid all the way to the minimum energy. For many minimization problems, this is not valid enough to be a useful approximation. Fortunately, there is a second approach called the conjugate gradient technique that often works as well and sometimes better.

Conjugate gradient techniques make use of the gradient but are designed to avoid the difficulties associated with long narrow wells where the steepest descent techniques result in oscillations. This is done by starting from a steepest descent in the first step of the minimization. In the second step, the displacement is taken to be along a direction that does not include the direction taken in the first step. Explicitly, let v_i be the direction taken in the ith step, then the first two directions would be:

$$\begin{aligned} v_1 &= - \; {}_s E(s)\Big|_{s_1} \\ v_2 &= - \; {}_s E(s)\Big|_{s_2} + v_1 \frac{\left(v_1 \quad {}_s E(s)\Big|_{s_2}\right)}{v_1 \;\; v_1} \end{aligned} \qquad (1.7.64)$$

This ensures that v_2 is orthogonal to v_1. Subsequent directions are made orthogonal to some number of previous steps. The use of orthogonal directions avoids much of the problem of bouncing back and forth in the energy well.

Monte Carlo simulation can also be used to find minimum energy configurations if the simulations are done at zero temperature. A zero temperature Monte Carlo means that the steps taken always reduce the energy of the system. This approach works not only for continuous variables, but also for the discrete variables like in the Ising model. For the Ising model,the zero temperature Monte Carlo described above and the zero temperature Glauber dynamics are the same. Every selected spin is placed in its low energy orientation—aligned with the local effective field.

None of these techniques are suited to finding the minimum energy configuration if there are multiple energy minima, and we do not know if we are located near the correct minimum energy location. One way to address this problem is to start from various initial configurations and to look for the local minimum nearby. By doing this many times it might be possible to identify the global minimum energy. This works when there are only a few different energy minima. There are no techniques that guarantee finding the global minimum energy for an arbitrary energy function $E(s)$. However, by using Monte Carlo simulations that are not at $T = 0$, a systematic approach called simulated annealing has been developed to try to identify the global minimum.

1.7.5 *Simulated annealing*

Simulated annealing was introduced relatively recently as an approach to finding the global minimum when the energy or other optimization function contains many local minima. The approach is based on the physical process of heating a system and cooling it down slowly. The minimum energy for many simple materials is a crystal. If a material is heated to a liquid or vapor phase and cooled rapidly, the material does not crystallize. It solidifies as a glass or amorphous solid. On the other hand, if it is cooled slowly, crystals may form. If the material is formed out of several different kinds of atoms, the cooling may also result in phase separation into particular compounds or atomic solids. The separated compounds are lower in energy than a rapidly cooled mixture.

Simulated annealing works in much the same way. A Monte Carlo simulation is started at a high temperature. Then the temperature is lowered according to a cooling schedule until the temperature is so low that no additional movements are likely. If the procedure is effective,the final energy should be the lowest energy of the simulation. We could also keep track of the energy during the simulation and take the lowest value, and the configuration at which the lowest value was reached.

In general, simulated annealing improves upon methods that find only a local minimum energy, such as steepest descent, discussed in the previous section. For some problems, the improvement is substantial. Even if the minimum energy that is found is not the absolute minimum in energy of the system, it may be close. For example, in problems where there are many configurations that have roughly the same low energy, simulated annealing may find one of the low-energy configurations.

However, simulated annealing does not work well for all problems,and for some problems it fails completely. It is also true that annealing of physical materials does not always result in the lowest energy conformation. Many materials, even when cooled slowly, result in polycrystalline materials, disordered solids and mixtures. When it is important for technological reasons to reach the lowest energy state, special techniques are often used. For example, the best crystal we know how to make is silicon. In order to form a good silicon crystal, it is grown using careful nonuniform cooling. A single crystal can be gradually pulled from a liquid that solidifies only on the surfaces of the existing crystal. Another technique for forming crystals is growth from the vapor phase, where atoms are deposited on a previously formed crystal that serves as a template for the continuing growth. The difficulties inherent in obtaining materials in their lowest energy state are also apparent in simulations.

In Section 1.4 we considered the cooling of a two-state system as a model of a glass transition. We can think about this simulation to give us clues about why both physical and simulated annealing sometimes fail to find low energy states of the system. We saw that using a constant cooling rate leaves some systems stuck in the higher energy well. When there are many such high energy wells then the system will not be successful in finding a low energy state. The problem becomes more difficult if the height of the energy barrier between the two wells is much larger than the energy difference between the upper and lower wells. In this case, at higher temperatures the system does not care which well it is in. At low temperatures when it would like to be in the lower energy well, it cannot overcome the barrier. How well the annealing works in finding a low energy state depends on whether we care about the energy scale characteristic of the barrier, or characteristic of the energy difference between the two minima.

There is another characteristic of the energy that can help or hurt the effectiveness of simulated annealing. Consider a system where there are many local minimum energy states (Fig. 1.7.4). We can think about the effect of high temperatures as placing the system in one of the many wells of the energy minima. These wells are called basins of attraction.A system in a particular basin of attraction will go into the minimum energy configuration of the basin if we suddenly cool to zero temperature. We

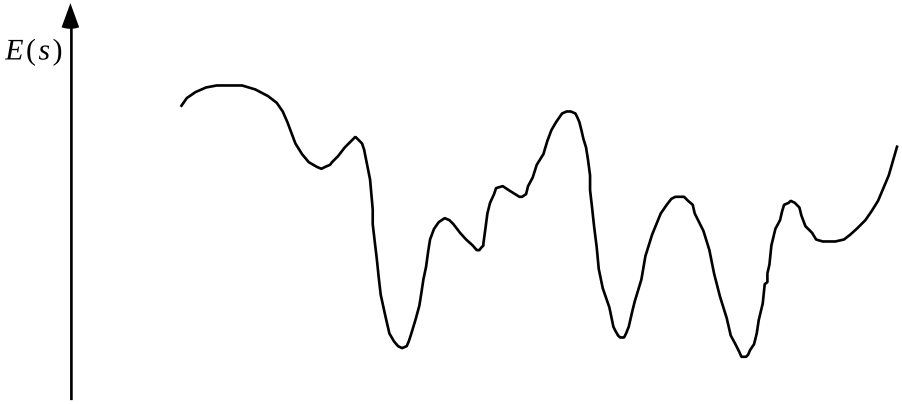


Figure 1.7.4 Schematic plot of a system energy $E(s)$ as a function of a system coordinate s. In simulated annealing, the location of a minimum energy is sought by starting from a high temperature Monte Carlo and cooling the system to a low temperature. At the high temperature the system has a high kinetic energy and explores all of the possible configurations. As the temperature is cooled it descends into one of the wells, called basins of attraction, and cannot escape. Finally, when the temperature is very low it loses all kinetic energy and sits in the bottom of the well. Minima with larger basins of attraction are more likely to capture the system. Simulated annealing works best when the lowest-energy minima have the largest basins of attraction. ∎

also can see that the gradual cooling in simulated annealing will result in low energy states if the size of the basin of attraction increases with the depth of the well. This means that at high temperatures the system is more likely to be in the basin of attraction of a lower energy minimum. Thus,simulated annealing works best when energy varies in the space in such a way that deep energy minima also have large basins of attraction. This is sometimes but not always true both in physical systems and in mathematical optimization problems.

Another way to improve the performance of simulated annealing is to introduce nonlocal Monte Carlo steps. If we understand the characteristics of the energy, we can design steps that take us through energy barriers. The problem with this approach is that if we don't know the energy surface well enough, then moving around in the space by arbitrary nonlocal steps will result in attempts to move to locations where the energy is high. These steps will be rejected by the Monte Carlo and the nonlocal moves will not help. An example where nonlocal Monte Carlo moves can help is treatments of low-energy atomic configurations in solids. Nonlocal steps can allow atoms to move through each other, switching their relative positions, instead of trying to move gradually around each other.

Finally, for the success of simulated annealing, it is often necessary to design carefully the cooling schedule.Generally, the slower the cooling the more likely the simulation will end up in a low energy state. However, given a finite amount of computer and human time,it is impossible to allow an arbitrarily slow cooling. Often there are particular temperatures where the cooling rate is crucial. This happens at phase transitions, such as at the liquid-to-solid phase boundary. If we know of such a transition, then we can cool rapidly down to the transition, cool very slowly in its vicinity and then speed up thereafter. The most difficult problems are those where there are barriers of varying heights leading to a need to cool slowly at all temperatures.

For some problems the cooling rate should be slowed as the temperature becomes lower. One way to achieve this is to use a logarithmic cooling schedule. For example, we set the temperature $T(t)$ at time step t of the Monte Carlo, to be:

$$T(t) = T_0 / \ln(t / t_0 + 1) \tag{1.7.65}$$

where t_0 and T_0 are parameters that must be chosen for the particular problem. In Question 1.7.5 we show that for the two-state system,if $kT_0 > (E_B - E_1)$,then the system will always relax into its ground state.

Question 1.7.5: Show that by using a logarithmic cooling schedule, Eq. (1.7.65), where $kT_0 > (E_B - E_1)$,the two-state system of Section 1.4 always relaxes into the ground state. To simplify the problem, consider an incremental time Δt during which the temperature is fixed.Show that the system will still relax to the equilibrium probability over this incremental time, even at low temperatures.

Solution 1.7.5: We write the solution of the time evolution during the incremental time Δt from Eq. (1.4.45) as:

$$P(1;t + \Delta t) - P(1;\infty) = (P(1;t) - P(1;\infty))e^{-t/\tau(t)} \tag{1.7.66}$$

where $P(1;\infty)$ is the equilibrium value of the probability for the temperature $T(t)$. $\tau(t)$is the relaxation time for the temperature$T(t)$. In order for relaxation to occur we must have that $e^{-t/\tau(t)} << 1$, equivalently:

$$t / \tau(t) >> 1 \tag{1.7.67}$$

We calculate $\tau(t)$ from Eq. (1.4.44):

$$\begin{aligned} 1/\tau(t) &= \nu\left(e^{-(E_B-E_1)/kT(t)} + e^{-(E_B-E_{-1})/kT(t)}\right) \\ &> \nu e^{-(E_B-E_1)/kT(t)} = \nu(t/t_0+1)^{-\gamma} \end{aligned} \tag{1.7.68}$$

where we have substituted Eq. (1.7.65) and defined $\gamma = (E_B - E_1)/kT_0$. We make the reasonable assumption that we start our annealing at a high temperature where relaxation is not a problem. Then by the time we get to the low temperatures that are of interest, $t >> t_0$, so:

$$1/\tau(t) > 2\nu(t/t_0)^{-\gamma} \tag{1.7.69}$$

and

$$t/\tau(t) > \nu t_0^{\gamma} t^{1-\gamma} \tag{1.7.70}$$

For $\gamma < 1$ the right-hand side increases with time and thus the relaxation improves with time according to Eq. (1.7.67). If relaxation occurs at higher temperatures, it will continue to occur at all lower temperatures despite the increasing relaxation time. ∎

1.8 Information

Ultimately, our ability to quantify complexity (How complex is it?) requires a quantification of information (How much information does it take to describe it?). In this section, we discuss information. We will also need computation theory described in Section 1.9 to discuss complexity in Chapter 8.A quantitative theory of information was developed by Shannon to describe the problem of communication. Specifically, how much information can be communicated through a transmission channel (e.g., a telephone line) with a specified alphabet of letters and a rate at which letters can be transmitted. The simplest example is a binary alphabet consisting of two characters (digits) with a fixed rate of binary digits (bits) per second. However, the theory is general enough to describe quite arbitrary alphabets,letters of variable duration such as are involved in Morse code, or even continuous sound with a specified band-width. We will not consider many of the additional applications,our objective is to establish the basic concepts.

1.8.1 *The amount of information in a message*

We start by considering the information content of a string of digits $s = (s_1 s_2 ... s_N)$.One might naively expect that information is contained in the state of each digit. However, when we receive a digit, we not only receive information about what the digit is, but also what the digit is not. Let us assume that a digit in the string of digits we receive is the number 1. How much information does this provide? We can contrast two different scenarios—binary and hexadecimal digits:

1. There were two possibilities for the number, either 0 or 1.

2. There were sixteen possibilities for the number {0, 1, 2,3,4,5, 6, 7,8, 9, A, B, C, D, E, F}.

In which of these did the "1" communicate more information? Since the first case provides us with the information that it is "not 0," while the second provides us with the information that it is "not 0," "not 2," "not 3," etc., the second provides more information. Thus there is more information in a digit that can have sixteen states than a digit that can have only two states. We can quantify this difference if we consider a binary representation of hexadecimal digits {0000,0001,0010,0011,…,1111}. It takes four binary digits to represent one hexadecimal digit. The hexadecimal number 1 is represented as 0001 in binary form and uses four binary digits. Thus a hexadecimal 1 contains four times as much information as a binary 1.

We note that the amount of information does not depend on the particular value that is taken by the digit. For hexadecimal digits, consider the case of a digit that has the value 5. Is there any difference in the amount of information given by the 5 than if it were 1? No, either number contains the same amount of information.

This illustrates that information is actually contained in the distinction between the state of a digit compared to the other possible states the digit may have. In order to quantify the concept of information, we must specify the number of possible states. Counting states is precisely what we did when we defined the entropy of a system in Section 1.3. We will see that it makes sense to define the information content of a string in the same way as the entropy—the logarithm of the number of possible states of the string:

$$I(s) = \log_2(\Omega) \qquad (1.8.1)$$

By convention, the information is defined using the logarithm base two. Thus, the information contained in a single binary digit which has two possible states is $\log_2(2) = 1$. More generally, the number of possible states in a string of N bits, with each bit taking one of two values (0 or 1) is 2^N. Thus the information in a string of N bits is (in what follows the function log( ) will be assumed to be base two):

$$I(s) = \log(2^N) = N \qquad (1.8.2)$$

Eq.(1.8.2) says that each bit provides one unit of information. This is consistent with the intuition that the amount of information grows linearly with the length of the string. The logarithm is essential, because the number of possible states grows exponentially with the length of the string, while the information grows linearly.

It is important to recognize that the definition of information we have given assumes that each of the possible realizations of the string has equal a priori probability. We use the phrase a priori to emphasize that this refers to the probability prior to receipt of the string—once the string has arrived there is only one possibility.

To think about the role of probability we must discuss further the nature of the message that is being communicated. We construct a scenario involving a sender and a receiver of a message. In order to make sure that the recipient of the message could not have known the message in advance (so there is information to communicate), we assume that the sender of the information is sending the result of a random occurrence, like the flipping of a coin or the throwing of a die. To enable some additional flexibility, we assume that the random occurrence is the drawing of a ball from a bag. This enables us to construct messages that have different probabilities. To be specific, we assume there are ten balls in the bag numbered from 0 to 9. All of them are red except the ball marked 0, which is green. The person communicating the message only reports ifthe ball drawn from the bag is red (using the digit 1) or green (using the digit 0). The recipient of the message is assumed to know about the setup. If the recipient receives the number 0,he then knows exactly which ball was selected,and all that were not selected. However, if he receives a 1, this provides less information, because he only knows that one of nine was selected,not which one. We notice that the digit 1 is nine times as likely to occur as the digit 0.This suggests that a higher probability digit contains less information than a lower probability digit.

We generalize the definition of the information content of a string of digits to allow for the possibility that different strings have different probabilities. We assume that the string is one of an ensemble of possible messages, and we define the information as:

$$I(s) = -\log(P(s)) \tag{1.8.3}$$

where $P(s)$ is the probability of the occurrence of the message s in the ensemble. Note that in the case of equal a priori probability $P(s) = 1/\Omega$, Eq. (1.8.3) reduces to Eq. (1.8.1). The use of probabilities in the definition of information makes sense in one of two cases:(1) The recipient knows the probabilities that represent the conventions of the transmission, or (2) A large number of independent messages are sent, and we are considering the information communicated by one of them. Then we can approximate the probability of a message by its proportion of appearance among the messages sent. We will discuss these points in greater detail later.

Question 1.8.1 Calculate the information, according to Eq.(1.8.3),that is provided by a single digit in the example given in the text of drawing red and green balls from a bag.

Solution 1.8.1 For the case of a 0,the information is the same as that of a decimal digit:

$$I(0) = -\log(1/10) \approx 3.32 \tag{1.8.4}$$

For the case of a 1 the information is

$$I(0) = -\log(9/10) \approx 0.152 \tag{1.8.5}$$ ∎

We can specialize the definition of information in Eq. (1.8.3) to a message $s = (s_1 s_2 ... s_N)$ composed of individual characters (bits, hexadecimal characters, ASCII characters, decimals, etc.) that are completely independent of each other (for example, each corresponding to the result of a separate coin toss). This means that the total probability of the message is the product of the probability of each character, $P(s) = \prod_i P(s_i)$. Then the information content of the message is given by:

$$I(s) = -\sum_i \log(P(s_i)) \tag{1.8.6}$$

If all of the characters have equal probability and there are k possible characters in the alphabet, then $P(s_i) = 1/k$, and the information content is:

$$I(s) = N \log(k) \tag{1.8.7}$$

For the case of binary digits,this reduces to Eq.(1.8.2). For other cases like the hexadecimal case, $k = 16$,this continues to make sense:the information $I = 4N$ corresponds to the requirement of representing each hexadecimal digit with four bits. Note that the previous assumption of equal a priori probability for the whole string is stronger than the independence of the digits and implies it.

Question 1.8.2 Apply the definition of information content in Eq.(1.8.3) to each of the following cases. Assume messages consist of a total of N bits subject to the following constraints (aside for the constraints assume equal probabilities):

1. Every even bit is 1.
2. Every (odd, even) pair of bits is either 11 or 00.
3. Every eighth bit is a parity bit (the sum modulo 2 of the previous seven bits).

Solution 1.8.2: In each case, we first give an intuitive argument, and then we show that Eq. (1.8.3) or Eq. (1.8.6) give the same result.

1. The only information that is transferred is the state of the odd bits. This means that only half of the bits contain information. The total information is $N / 2$. To apply Eq.(1.8.6), we see that the even bits, which always have the value 1, have a probability $P(1) = 1$ which contributes no information. Note that we never have to consider the case $P(0) = 0$ for these bits, which is good, because by the formula it would give infinite information. The odd bits with equal probabilities, $P(1) = P(0) = 1/2$, give an information of one for either value received.
2. Every pair of bits contains only two possibilities, giving us the equivalent of one bit of information rather than two. This means that total information is $N/2$. To apply Eq.(1.8.6), we have to consider every (odd, even) pair of bits as a single character. These characters can never have the value 01 or 10, and they have the value 11 or 00 with probability $P(11) = P(00) = 1/2$, which gives the expected result. We will see later that there is another way to think about this example by using conditional probabilities.
3. The number of independent pieces of information is $7N / 8$. To see this from Eq. (1.8.6), we group each set of eight bits together and consider

them as a single character (a byte). There are only 2^7 different possibilities for each byte, and each one has equal probability according to our constraints and assumptions. This gives the desired result.

Note: Such representations are used to check for noise in transmission. If there is noise,the redundancy of the eighth bit provides additional information. The noise-dependent amount of additional information can also be quantified; however, we will not discuss it here. ▮

Question 1.8.3 Consider a transmission of English characters using an ASCII representation. ASCII characters are the conventional method for computer representation of English text including small and capital letters, numerals and punctuation. Discuss (do not evaluate for this question) how you would determine the information content of a message. We will evaluate the information content of English in a later question.

Solution 1.8.3 In ASCII, characters are represented using eight bits. Some of the possible combinations of bits are not used at all. Some are used very infrequently. One way to determine the information content of a message is to assume a model where each of the characters is independent. To calculate the information content using this assumption, we must find the probability of occurrence of each character in a sample text. Using these probabilities,the formula Eq.(1.8.6) could be applied. However, this assumes that the likelihood of occurrence of a character is independent of the preceding characters, which is not correct. ▮

Question 1.8.4: Assume that you know in advance that the number of ones in a long binary message is M. The total number of bits is N. What is the information content of the message? Is it similar to the information content of a message of N independent binary characters where the probability that any character is one is $P(1) = M / N$?

Solution 1.8.4: We count the number of possible messages with M ones and take the logarithm to obtain the information as

$$I = \log(\binom{N}{M}) = \log(\frac{N!}{M!(N-M)!}) \tag{1.8.8}$$

We can show that this is almost the same as the information of a message of the same length with a particular probability of ones, $P(1) = M / N$, by use of the first two terms of Sterling's approximation Eq. (1.2.36). Assuming $1 << M << N$ (A correction to this would grow logarithmically with N and can be found using the additional terms in (Eq. (1.2.36)):

$$\begin{aligned} I &\approx N(\log(N) - 1) - M(\log(M) - 1) - (N - M)(\log(N - M) - 1) \\ &= -N[P(1)\log(P(1)) + (1 - P(1))\log(1 - P(1))] \end{aligned} \tag{1.8.9}$$

This is the information from a string of independent characters where $P(1) = M / N$. For such a string, the number of ones is approximately $NP(1)$ and the number of zeros $N(1 - P(1))$ (see also Question 1.8.7). ▮

1.8.2 *Characterizing sources of information*

The information content of a particular message is defined in terms of the probability that it,out of all possible messages, will be received. This means that we are characterizing not just a message but the source of the message.A direct characterization of the source is not the information of a particular message, but the average information over the ensemble of possible messages. For a set of possible messages with a given probability distribution $P(s)$ this is:

$$< I > = -\sum_s P(s)\log(P(s)) \tag{1.8.10}$$

If the messages are composed out of characters $s = (s_1 s_2 ... s_N)$,and each character is determined independently with a probability $P(s_i)$, then we can write the information content as:

$$< I > = -\sum_s \left(\prod_i P(s_i)\right)\log\left(\prod_{i'} P(s_{i'})\right) = -\sum_s \left(\prod_i P(s_i)\right)\sum_{i'} \log\big(P(s_{i'})\big) \tag{1.8.11}$$

We can move the factor in parenthesis inside the inner sum and interchange the order of the summations.

$$< I > = -\sum_{i'} \sum_s \left(\prod_i P(s_i)\right)\log\big(P(s_{i'})\big) = -\sum_{i'} \left(\sum_{\{s_i\}_{i\neq i'}} \prod_{i\neq i'} P(s_i)\right)\sum_{s_{i'}} P(s_{i'})\log\big(P(s_{i'})\big) \tag{1.8.12}$$

The latter expression results from recognizing that the sum over all possible states is a sum over all possible values of each of the letters. The sum and product can be interchanged:

$$\sum_{\{s_i\}_{i\neq i'}} \prod_{i\neq i'} P(s_i) = \prod_{i\neq i'} \sum_{s_i} P(s_i) = 1 \tag{1.8.13}$$

giving the result:

$$< I > = -\sum_{i'} \sum_{s_{i'}} P(s_{i'})\log\big(P(s_{i'})\big) \tag{1.8.14}$$

This shows that the average information content of the whole message is the average information content of each character summed over the whole character string. If the characters have the same probability, this is just the average information content of an individual character times the number of characters. If all letters of the alphabet have the same probability, this reduces to Eq. (1.8.7).

The average information content of a binary variable is given by:

$$< I > = -(P(1)\log(P(1)) + P(0)\log(P(0))) \qquad (1.8.15)$$

Aside from the use of a logarithm base two, this is the same as the entropy of a spin (Section 1.6) with two possible states $s = \pm 1$ (see Question 1.8.5). The maximum information content occurs when the probabilities are equal, and the information goes to zero when one of the two becomes one,and the other zero (see Fig. 1.8.1). The information reflects the uncertainty in, or the lack of advance knowledge about, the value received.

Question 1.8.5 Show that the expression for the entropy S given in Eq. (1.6.16) of a set of noninteracting binary spins is the same as the information content defined in Eq.(1.8.15) aside from a normalization constant $k \ln(2)$. Consider the binary notation $s_i = 0$ to be the same as $s_i = -1$ for the spins.

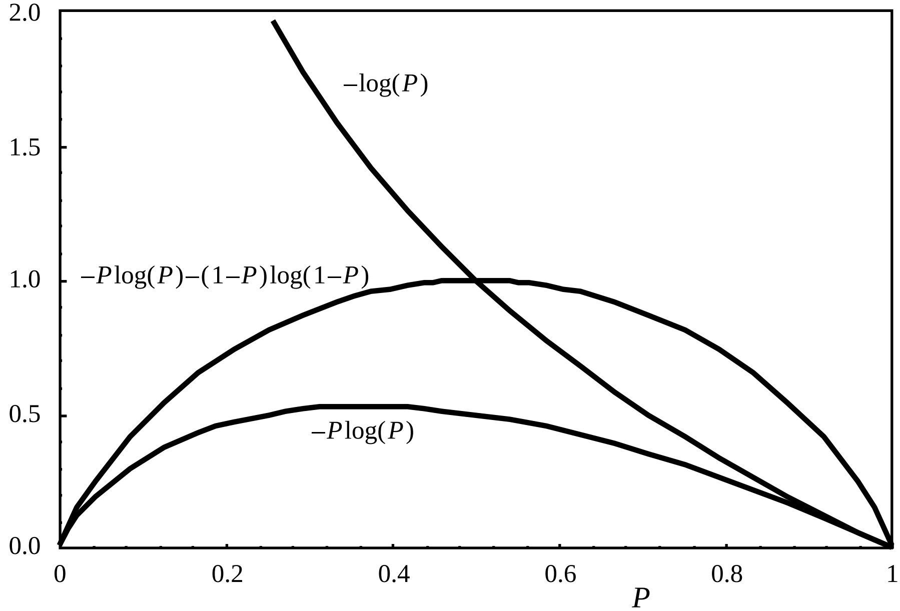


Figure 1.8.1 Plots of functions related to the information content of a message with probability P. $-\log(P)$ is the information content of a single message of probability P. $-P\log(P)$ is the contribution of this message to the average information given by the source. While the information content of a message diverges as P goes to zero, it appears less frequently so its contribution to the average information goes to zero. If there are only two possible messages, or two possible (binary) characters with probability P and $1 - P$ then the average information given by the source per message or per character is given by $-P\log(P) - (1 - P)\log(1 - P)$. ∎

Solution 1.8.5 The local magnetization m_i is the average value of a particular spin variable:

$$m_i = P_{s_i}(1) - P_{s_i}(-1) \tag{1.8.16}$$

Using $P_{s_i}(1) + P_{s_i}(-1) = 1$ we have:

$$\begin{aligned} P_{s_i}(1) &= (1 + m_i) / 2 \\ P_{s_i}(-1) &= (1 - m_i) / 2 \end{aligned} \tag{1.8.17}$$

Inserting these expressions into Eq. (1.8.15) and summing over a set of binary variables leads to the expression:

$$I = \left[N - \frac{1}{2}\sum_i \left((1+m_i)\log\left(1+m_i\right) + (1-m_i)\log\left(1-m_i\right) \right) \right] = S/k\ln(2) \tag{1.8.18}$$

The result is more general than this derivation suggests and will be discussed further in Chapter 8. ▮

Question 1.8.6 For a given set of possible messages, prove that the ensemble where all messages have equal probability provides the highest average information.

Solution 1.8.6 Since the sum over all probabilities is a fixed number (1), we consider what happens when we transfer some probability from one message to another. We start with the information given by

$$< I > = -P(s')\ln(P(s')) - P(s'')\ln(P(s'')) - \sum_{s \neq s',s''} P(s)\ln(P(s)) \tag{1.8.19}$$

and after shifting a probability of δ from one to the other we have:

$$< I' > = -(P(s') - \delta)\ln(P(s') - \delta) - (P(s'') + \delta)\ln(P(s'') + \delta) - \sum_{s \neq s',s'} P(s)\ln(P(s)) \tag{1.8.20}$$

We need to expand the change in information to first nonzero order in δ. We simplify the task by using the expression:

$$<I'> - <I> = f(P(s') + \delta) - f(P(s')) + f(P(s'') - \delta) - f(P(s'')) \tag{1.8.21}$$

where

$$f(x) = -x\log(x) \tag{1.8.22}$$

Taking a derivative, we have

$$\frac{d}{dx} f(x) = -(\log(x) + 1) \tag{1.8.23}$$

This gives the result:

$$<I'> - <I> = -(\log(P(s')) - \log(P(s'')))\delta \tag{1.8.24}$$

Since $\log(x)$ is a monotonic increasing function, we see that the average information increases $((<I'> - <I>) > 0)$ when probability $\delta > 0$ is transferred from a higher-probability character to a lower-probability character $(P(s'')$ $> P(s')$ $\quad -(\log(P(s')) - \log(P(s'')) > 0)$. Thus,any change of the probability toward a more uniform probability distribution increases the average information. ▮

Question 1.8.7 A source produces strings of characters of length N. Each character that appears in the string is independently selected from an alphabet of characters with probabilities $P(s_i)$. Write an expression for the probability $P(s)$ of a typical string of characters. Show that this expression implies that the string gives N times the average information content of an individual character. Does this mean that every string must give this amount of information?

Solution 1.8.7 For a long string, each character will appear $NP(s_i)$ times. The probability of such a string is:

$$P(s) = \prod_{s_i} P(s_i)^{NP(s_i)} \tag{1.8.25}$$

The information content is:

$$I(s) = -\log(P(s)) = -N \sum_{s_i} P(s_i)\log(P(s_i)) \tag{1.8.26}$$

which is N times the average information of a single character. This is the information of a typical string. A particular string might have information significantly different from this. However, as the number of characters in the string increases, by the central limit theorem (Section 1.2), the fraction of times a particular character appears (i.e., the distance traveled in a random walk divided by the total number of steps) becomes more narrowly distributed around the expected probability $P(s_i)$. This means the proportion of messages whose information content differs from the typical value decreases with increasing message length. ▮

1.8.3 *Correlations between characters*

Thus far we have considered characters that are independent of each other. We can also consider characters whose values are correlated. We describe the case of two correlated characters. Because there are two characters,the notation must be more complete. As discussed in Section 1.2, we use the notation $P_{s_1,s_2}(s'_1, s'_2)$ to denote the probability that in the same string the character s_1 takes the value s'_1 and the variable s_2 takes the value s'_2. The average information contained in the two characters is given by:

$$<I_{s_1,s_2}> = -\sum_{s'_1,s'_2} P_{s_1,s_2}(s'_1,s'_2)\log(P_{s_1,s_2}(s'_1,s'_2)) \tag{1.8.27}$$

Note that the notation $I(s_1,s_2)$ is often used for this expression. We use $<I_{s_1,s_2}>$ because it is not a function of the values of the characters—it is the average information carried by the characters labeled by s_1 and s_2. We can compare the information content of the two characters with the information content of each character separately:

$$\begin{aligned} P_{s_1}(s'_1) &= \sum_{s'_2} P_{s_1,s_2}(s'_1,s'_2) \\ P_{s_2}(s'_2) &= \sum_{s'_1} P_{s_1,s_2}(s'_1,s'_2) \end{aligned} \tag{1.8.28}$$

$$\begin{aligned} <I_{s_1}> &= -\sum_{s'_1,s'_2} P_{s_1,s_2}(s'_1,s'_2)\log(\sum_{s'_2} P_{s_1,s_2}(s'_1,s''_2)) \\ <I_{s_2}> &= -\sum_{s'_1,s'_2} P_{s_1,s_2}(s'_1,s'_2)\log(\sum_{s''_1} P_{s_1,s_2}(s''_1,s'_2)) \end{aligned} \tag{1.8.29}$$

It is possible to show (see Question 1.8.8) the inequalities:

$$<I_{s_2}> + <I_{s_1}> \quad \geq \quad <I_{s_1,s_2}> \quad \geq \quad <I_{s_2}>,<I_{s_1}> \tag{1.8.30}$$

The right inequality means that we receive more information from both characters than from either one separately. The left inequality means that information we receive from both characters together cannot exceed the sum of the information from each separately. It can be less if the characters are dependent on each other. In this case, receiving one character reduces the information given by the second.

The relationship between the information from a character s_1 and the information from the same character after we know another character s_2 can be investigated by defining a contingent or conditional probability:

$$P_{s_1,s_2}(s'_1|s'_2) = \frac{P_{s_1,s_2}(s'_1,s'_2)}{\sum_{s''_1} P_{s_1,s_2}(s''_1,s'_2)} \tag{1.8.31}$$

This is the probability that s_1 takes the value s'_1 assuming that s_2 takes the value s'_2. We used this notation in Section 1.2 to describe the transitions from one value to the next in a chain of events (random walk). Here we are using it more generally. We could recover the previous meaning by writing the transition probability as $P_s(s'_1|s'_2) = P_{s(t),s(t-1)}(s'_1|s'_2)$. In this section we will be concerned with the more general definition, Eq. (1.8.31).

We can find the information content of the character s_1 when s_2 takes the value s'_2

$$\begin{aligned} <I_{s_1}>\big|_{s_2=s'_2} &= -\sum_{s'_1} P_{s_1,s_2}(s'_1|s'_2)\log\left(P_{s_1,s_2}(s'_1|s'_2)\right) \\ &= \frac{-\sum_{s'_1} P_{s_1,s_2}(s'_1,s'_2)\left(\log(P_{s_1,s_2}(s'_1,s'_2)) - \log(\sum_{s'''_1} P_{s_1,s_2}(s'''_1,s'_2))\right)}{\sum_{s'_1} P_{s_1,s_2}(s''_1,s'_2)} \end{aligned} \tag{1.8.32}$$

This can be averaged over possible values of s_2, giving us the average information content of the character s_1 when the character s_2 is known.

$$
\begin{aligned}
<<I_{s_1|s_2}>> &\equiv <<I_{s_1}>\Big|_{s_2=s_2'}> \\
&= -\sum_{s_2'} P_{s_2}(s_2') \sum_{s_1'} P_{s_1,s_2}(s_1' \mid s_2') \log\left(P_{s_1,s_2}(s_1' \mid s_2')\right) \\
&= -\sum_{s_2'} \sum_{s_1''} P_{s_1,s_2}(s_1'',s_2') \sum_{s_1'} \frac{P_{s_1,s_2}(s_1',s_2')}{\sum_{s_1''} P_{s_1,s_2}(s_1'',s_2')} \log\left(P_{s_1,s_2}(s_1' \mid s_2')\right) \\
&= -\sum_{s_1',s_2'} P_{s_1,s_2}(s_1',s_2') \log\left(P_{s_1,s_2}(s_1' \mid s_2')\right)
\end{aligned}
\tag{1.8.33}
$$

The average we have taken should be carefully understood. The unconventional double average notation is used to indicate that the two averages are of a different nature. One way to think about it is as treating the information content of a dynamic variable s_1 when s_2 is a quenched (frozen) random variable. We can rewrite this in terms of the information content of the two characters,and the information content of the character s_2 by itself as follows:

$$
\begin{aligned}
<<I_{s_1|s_2}>> &= -\sum_{s_1',s_2'} P_{s_1,s_2}(s_1',s_2') \left(\log(P_{s_1,s_2}(s_1',s_2')) - \log(\sum_{s_1''} P_{s_1,s_2}(s_1'',s_2')) \right) \\
&= <I_{s_1,s_2}> - <I_{s_2}>
\end{aligned}
\tag{1.8.34}
$$

Thus we have:

$$
<I_{s_1,s_2}> = <I_{s_1}> + <<I_{s_2|s_1}>> = <I_{s_2}> + <<I_{s_1|s_2}>> \tag{1.8.35}
$$

This is the intuitive result that the information content given by both characters is the same as the information content gained by sequentially obtaining the information from the characters.Once the first character is known,the second character provides only the information given by the conditional probabilities. There is no reason to restrict the use of Eq.(1.8.27) – Eq.(1.8.35) to the case where s_1 is a single character and s_2 is a single character. It applies equally well if s_1 is one set of characters,and s_2 is another set of characters.

Question 1.8.8 Prove the inequalities in Eq. (1.8.30).

Hints for the left inequality:

1. It is helpful to use Eq. (1.8.35).
2. Use convexity ($f(\ x\ ) > \ f(x)\ $) of the function $f(x) = -x\log(x)$.

Solution 1.8.8 The right inequality in Eq. (1.8.30) follows from the inequality:

$$
P_{s_1}(s_1'') = \sum_{s_1''} P_{s_1,s_2}(s_1'',s_2') > P_{s_1,s_2}(s_1',s_2') \tag{1.8.36}
$$

The logarithm is a monotonic increasing function, so we can take the logarithm:

$$\log(\sum_{s_1''} P_{s_1,s_2}(s_1'',s_2')) > \log(P_{s_1,s_2}(s_1',s_2')) \tag{1.8.37}$$

Changing sign and averaging leads to the desired result:

$$\begin{aligned} < I_{s_2} > &= -\sum_{s_1',s_2'} P_{s_1,s_2}(s_1',s_2')\log(\sum_{s_1''} P_{s_1,s_2}(s_1'',s_2')) \\ &< \ -\sum_{s_1',s_2'} P_{s_1,s_2}(s_1',s_2')\log(P_{s_1,s_2}(s_1',s_2')) = < I_{s_1,s_2} > \end{aligned} \tag{1.8.38}$$

The left inequality in Eq. (1.8.30) may be proven from Eq. (1.8.35) and the intuitive inequality

$$(< I_{s_1} >) > (<< I_{s_1|s_2} >>) \tag{1.8.39}$$

To prove this inequality we make use of the convexity of the function $f(x) = -x\log(x)$. Convexity of a function means that its value always lies above line segments (secants) that begin and end at points along its graph. Algebraically:

$$f((ax + by) / (a + b)) > (af(x) + bf(y)) / (a + b) \tag{1.8.40}$$

More generally, taking a set of values of x and averaging over them gives:

$$f(\ x\ ) > \ f(x) \tag{1.8.41}$$

Convexity of $f(x)$ follows from the observation that

$$\frac{d^2f}{dx^2} = -\frac{1}{x\ln(2)} < 0 \tag{1.8.42}$$

for all $x > 0$, which is where the function $f(x)$ is defined.

We then note the relationship:

$$P_{s_1}(s_1') = \sum_{s_2'} P_{s_2}(s_2')P_{s_1,s_2}(s_1' | s_2') = < P_{s_1,s_2}(s_1' | s_2') >_{s_2} \tag{1.8.43}$$

where, to simplify the following equations, we use a subscript to indicate the average with respect to s_2. The desired result follows from applying convexity as follows:

$$\begin{aligned} < I_{s_1} > &= -\sum_{s_1'} P_{s_1}(s_1')\log(P_{s_1}(s_1')) = \sum_{s_1'} f(P_{s_1}(s_1')) = \sum_{s_1'} f(< P_{s_1,s_2}(s_1' | s_2') >_{s_2}) \\ &> \ \sum_{s_1'} < f(P_{s_1,s_2}(s_1' | s_2')) >_{s_2} \\ &= -\sum_{s_2'} P_{s_2}(s_2')\sum_{s_1'} P_{s_1,s_2}(s_1' | s_2')\log\left(P_{s_1,s_2}(s_1' | s_2')\right) = << I_{s_1|s_2} >> \end{aligned} \tag{1.8.44}$$

the final equality following from the definition in Eq. (1.8.33). We can now make use of Eq. (1.8.35) to obtain the desired result. ▮

1.8.4 *Ergodic sources*

We consider a source that provides arbitrarily long messages, or simply continues to give characters at a particular rate. Even though the messages are infinitely long, they are still considered elements of an ensemble. It is then convenient to measure the average information per character. The characterization of such an information source is simplified if each (long) message contains within it a complete sampling of the possibilities. This means that if we wait long enough, the entire ensemble of possible character sequences will be represented in any single message. This is the same kind of property as an ergodic system discussed in Section 1.3. By analogy, such sources are known as ergodic sources. For an ergodic source, not only the characters appear with their ensemble probabilities, but also the pairs of characters, the triples of characters, and so on.

For ergodic sources, the information from an ensemble average over all possible messages is the same as the information for a particular long string. To write this down we need a notation that allows variable length messages. We write $\underline{s}_N = (s_1 s_2 ... s_N)$, where N is the length of the string. The average information content per character may be written as:

$$< i_s > = \lim_{N\to\infty} \frac{< I_{\underline{s}_N} >}{N} = -\lim_{N\to\infty} \frac{1}{N} \sum_{\underline{s}_N} P(\underline{s}_N)\log(P(\underline{s}_N)) = -\lim_{N\to\infty} \frac{1}{N} \log(P(\underline{s}_N)) \tag{1.8.45}$$

The rightmost equality is valid for an ergodic source. An example of an ergodic source is a source that provides independent characters—i.e., selects each character from an ensemble. For this case, Eq.(1.8.45) was shown in Question 1.8.7. More generally, for a source to be ergodic, long enough strings must break up into independent substrings, or substrings that are more and more independent as their length increases.

Assuming that N is large enough, we can use the limit in Eq. (1.8.45) and write:

$$P(\underline{s}_N) \approx 2^{-N<i_s>} \tag{1.8.46}$$

Thus, for large enough N, there are a set of strings that are equally likely to be generated by the source. The number of these strings is

$$2^{N<i_s>} \tag{1.8.47}$$

Since any string of characters is possible, in principle, this statement must be formally understood as saying that the total probability of all other strings becomes arbitrarily small.

If the string of characters is a Markov chain (Section 1.2), so that the probability of each character depends only on the previous character, then there are general conditions that can ensure that the source is ergodic. Similar to the discussion of Monte Carlo simulations in Section 1.7, for the source to be ergodic, the transition probabilities between characters must be irreducible and acyclic. Irreducibility guarantees that all characters are accessible from any starting character. The acyclic property guarantees that starting from one substring, all other substrings are accessible. Thus, if we can reach any particular substring, it will appear with the same frequency in all long strings.

We can generalize the usual Markov chain by allowing the probability of a character to depend on several (n) previous characters. A Markov chain may be constructed to represent such a chain by defining new characters, where each new character is formed out of a substring of n characters. Then each new character depends only on the previous one. The essential behavior of a Markov chain that is important here is that correlations measured along the chain of characters disappear exponentially. Thus,the statistical behavior of the chain in one place is independent of what it was in the sufficiently far past. The number of characters over which the correlations disappear is the correlation length. By allowing sufficiently many correlation lengths along the string—segments that are statistically independent—the average properties of one string will be the same as any other such string.

Question 1.8.9 Consider ergodic sources that are Markov chains with two characters $s_i = \pm 1$ with transition probabilities:

a. $P(1|1) = .999, P(-1|1) = .001, P(-1|-1) = 0.5, P(1|-1) = 0.5$

b. $P(1|1) = .999, P(-1|1) = .001, P(-1|-1) = 0.999, P(1|-1) = 0.001$

c. $P(1|1) = .999, P(-1|1) = .001, P(-1|-1) = 0.001, P(1|-1) = 0.999$

d. $P(1|1) = .001, P(-1|1) = .999, P(-1|-1) = 0.5, P(1|-1) = 0.5$

e. $P(1|1) = .001, P(-1|1) = .999, P(-1|-1) = 0.999, P(1|-1) = 0.001$

f. $P(1|1) = .001, P(-1|1) = .999, P(-1|-1) = 0.001, P(1|-1) = 0.999$

Describe the appearance of the strings generated by each source, and (roughly) its correlation length.

Solution 1.8.9 (a) has long regions of 1s of typical length 1000. In between there are short strings of –1s of average length 2 = 1 + 1/2 + 1/4 + ...(there is a probability of 1/2 that a second character will be –1 and a probability of 1/4 that both the second and third will be –1, etc.). (b) has long regions of 1s and long regions of –1s, both of typical length 1000. (c) is like (a) except the regions of –1s are of length 1. (d) has no extended regions of 1 or –1 but has slightly longer regions of –1s. (e) inverts (c). (f) has regions of alternating 1 and –1 of length 1000 before switching to the other possibility (odd and even indices are switched). We see that the characteristic correlation length is of order 1000 in (a),(b),(c),(e) and (f) and of order 2 in (d). ▮

We have considered in detail the problem of determining the information content of a message, or the average information generated by a source, when the characteristics of the source are well defined. The source was characterized by the ensemble of possible messages and their probabilities. However, we do not usually have a well-defined characterization of a source of messages,so a more practical question is to determine the information content from the message itself.The definitions that we have provided do not guide us in determining the information of an arbitrary message. We must have a model for the source. The model must be constructed out of the information we have—the string of characters it produces.One possibility is to model the source as ergodic. An ergodic source can be modeled in two ways, as a source of independent substrings or as a generalized Markov chain where characters depend on a certain number of previous characters. In each case we construct not one,but an infinite sequence of models. The models are designed so that if the source is ergodic then the information estimates given by the models converge to give the correct information content.

There is a natural sequence of independent substring models indexed by the number of characters in the substrings n. The first model is that of a source producing independent characters with a probability specified by their frequency of occurrence in the message. The second model would be a source producing pairs of correlated characters so that every pair of characters is described by the probability given by their occurrence (we allow character pairs to overlap in the message). The third model would be that of a source producing triples of correlated characters,and so on. We use each of these models to estimate the information. The nth model estimate of the information per character given by the source is:

$$< i_s >_{1,n} = \lim_{N \to \infty} \frac{1}{n} \sum_{\underline{s}_n} \tilde{P}_N(\underline{s}_n) \log(\tilde{P}_N(\underline{s}_n)) \qquad (1.8.48)$$

where we indicate using the subscript 1,n that this is an estimate obtained using the first type of model (independent substring model) using substrings of length n. We also make use of an approximate probability for the substring defined as

$$\tilde{P}_N(\underline{s}_n) = N(\underline{s}_n)/(N - n + 1) \qquad (1.8.49)$$

where $N(\underline{s}_n)$ is the number of times s_n appears in the string of length N. The information of the source might then be estimated as the limit $n \to \infty$ of Eq. (1.8.48):

$$< i_s > = \lim_{n \to \infty} \lim_{N \to \infty} \frac{1}{n} \sum_{\underline{s}_n} \tilde{P}_N(\underline{s}_n) \log(\tilde{P}_N(\underline{s}_n)) \qquad (1.8.50)$$

For an ergodic source, we can see that this converges to the information of the message. The n limit converges monotonically from above. This is because the additional information in $\underline{s}_{n+1}$ given by s_{n+1} is less than the information added by each previous character (see Eq. 1.8.59 below). Thus, the estimate of information per character based on $\underline{s}_n$ is higher than the estimate based on $\underline{s}_{n+1}$. Therefore, for each value of n the estimate $<i_s>_{1,n}$ is an upper bound on the information given by the source.

How large does N have to be? Since we must have a reasonable sample of the occurrence of substrings in order to estimate their probability, we can only estimate probabilities of substrings that are much shorter than the length of the string. The number of possible substrings grows exponentially with n as k^n, where k is the number of possible characters. If substrings occur with roughly similar probabilities, then to estimate the probability of a substring of length n would require at least a string of length k^n characters. Thus,taking the large N limit should be understood to correspond to N greater than k^n. This is a very severe requirement. This means that to study a model of English character strings of length $n = 10$ (ignoring upper and lower case, numbers and punctuation) would require $26^{10} \sim 10^{14}$ characters. This is roughly the number of characters in all of the books in the Library of Congress (see Question 1.8.15).

The generalized Markov chain model assumes a particular character is dependent only on n previous characters. Since the first n characters do not provide a significant amount of information for a very long chain ($N >> n$), we can obtain the average information per character from the incremental information given by a character. Thus, for the nth generalized Markov chain model we have the estimate:

$$< i_s >_{2,n} = << I_{s_n|\underline{s}_{n-1}} >> = \lim_{N\to\infty} \sum_{\underline{s}_{n-1}} \tilde{P}_N(\underline{s}_{n-1}) \sum_{s_n} \tilde{P}(s_n \mid \underline{s}_{n-1}) \log(\tilde{P}(s_n \mid \underline{s}_{n-1})) \tag{1.8.51}$$

where we define the approximate conditional probability using:

$$\tilde{P}_N(s_n \mid \underline{s}_{n-1}) = N(\underline{s}_{n-1}s_n)/N(\underline{s}_{n-1}) \tag{1.8.52}$$

Taking the limit $n \to \infty$ we have an estimate of the information of the source per character:

$$< i_s > = \lim_{n\to\infty} \lim_{N\to\infty} \sum_{\underline{s}_{n-1}} \tilde{P}_N(\underline{s}_{n-1}) \sum_{s_n} \tilde{P}(s_n \mid \underline{s}_{n-1}) \log(\tilde{P}(s_n \mid \underline{s}_{n-1})) \tag{1.8.53}$$

This also converges from above as a function of n for large enough N. For a given n, a Markov chain model takes into account more correlations than the previous independent substring model and thus gives a better estimate of the information (Question 1.8.10).

Question 1.8.10 Prove that the Markov chain model gives a better estimate of the information for ergodic sources than the independent substring model for a particular n. Assume the limit $N \to \infty$ so that the estimated probabilities become actual and we can substitute $\tilde{P}_N \to P$ in Eq.(1.8.48) and Eq. (1.8.51).

Solution 1.8.10 The information in a substring of length n is given by the sum of the information provided incrementally by each character, where the previous characters are known. We derive this statement algebraically (Eq. (1.8.59)) and use it to prove the desired result. Taking the N limit in Eq. (1.8.48), we define the nth approximation using the independent substring model as:

$$< i_s >_{1,n} = \frac{1}{n} \sum_{\underline{s}_n} P(\underline{s}_n) \log(P(\underline{s}_n)) \tag{1.8.54}$$

and for the *n*th generalized Markov chain model we take the same limit in Eq. (1.8.51):

$$< i_s >_{2,n} = \sum_{\underline{s}_{n-1}} P(\underline{s}_{n-1}) \sum_{s_n} P(s_n \,|\, \underline{s}_{n-1}) \log(P(s_n \,|\, \underline{s}_{n-1})) \qquad (1.8.55)$$

To relate these expressions to each other, follow the derivation of Eq.(1.8.34), or use it with the substitutions $s_1 \rightarrow \underline{s}_{n-1}$ and $s_2 \rightarrow s_n$, to obtain

$$< i_s >_{2,n} = -\sum_{\underline{s}_{n-1}, s_n} P(\underline{s}_{n-1} s_n) \left(\log(P(\underline{s}_{n-1} s_n)) - \log(\sum_{s_n} P(\underline{s}_{n-1} s_n)) \right) \qquad (1.8.56)$$

Using the identities

$$\begin{aligned} &P(\underline{s}_{n-1} s_n) = P(\underline{s}_n) \\ &P(\underline{s}_{n-1}) = \sum_{s_n} P(\underline{s}_{n-1} s_n) \end{aligned} \qquad (1.8.57)$$

this can be rewritten as:

$$< i_s >_{2,n} = n < i_s >_{1,n} - (n-1) < i_s >_{1,n-1} \qquad (1.8.58)$$

This result can be summed over n from 1 to n (the $n = 1$ case is $<i_s>_{2,1} = <i_s>_{1,1}$) to obtain:

$$\sum_{n'=1}^{n} < i_s >_{2,n'} = n < i_s >_{1,n} \qquad (1.8.59)$$

since $< i_s >_{2,n}$ is monotonic decreasing and $< i_s >_{1,n}$ is seen from this expression to be an average over $< i_s >_{2,n}$ with lower values of n, we must have that

$$< i_s >_{2,n} \le < i_s >_{1,n} \qquad (1.8.60)$$

as desired. ▮

Question 1.8.11 We have shown that the two models—the independent substring models and the generalized Markov chain model—are upper bounds to the information in a string. How good is the upper bound? Think up an example that shows that it can be terrible for both, but better for the Markov chain.

Solution 1.8.11 Consider the example of a long string formed out of a repeating substring, for example (000000010000000100000001…). The average information content per character of this string is zero. This is because once the repeat structure has become established,there is no more information. Any model that gives a nonzero estimate of the information content per character will make a great error in its estimate of the information content of the string, which is N times as much as the information per character.

For the independent substring model,the estimate is never zero. For the Markov chain model it is nonzero until n reaches the repeat distance. A Markov model with n the same size or larger than the repeat length will give the correct answer of zero information per character. This means that even for the Markov chain model, the information estimate does not work very well for n less than the repeat distance. ▮

Question 1.8.12 Write a computer program to estimate the information in English and find the estimate. For simple,easy-to-compute estimates, use single-character probabilities,two-character probabilities,and a Markov chain model for individual characters. These correspond to the above definitions of $< i_s >_{2,1} = < i_s >_{1,1}$, $< i_s >_{1,2}$, and $< i_s >_{2,2}$ respectively.

Solution 1.8.12 A program that evaluates the information content using single-character probabilities applied to the text (excluding equations) of Section 1.8 of this book gives an estimate of information content of 4.4 bits/character. Two-character probabilities gives 3.8 bits/character, and the one-character Markov chain model gives 3.3 bits/character. A chapter of a book by Mark Twain gives similar results. These estimates are decreasing in magnitude, consistent with the discussion in the text. They are also still quite high as estimates of the information in English per character.

The best estimates are based upon human guessing of the next character in a written text. Such experiments with human subjects give estimates of the lower and upper bounds of information content per character of English text. These are 0.6 and 1.2 bits/character. This range is significantly below the estimates we obtained using simple models. Remarkably, these estimates suggest that it is enough to give only one in four to one in eight characters of English in order for text to be decipherable. ▮

Question 1.8.13 Construct an example illustrating how correlations can arise between characters over longer than,say, ten characters. These correlations would not be represented by any reasonable character-based Markov chain model. Is there an example of this type relevant to the English language?

Solution 1.8.13 Example 1: If we have information that is read from a matrix row by row, where the matrix entries have correlations between rows, then there will be correlations that are longer than the length of the matrix rows.

Example 2: We can think about successive English sentences as rows of a matrix. We would expect to find correlations between rows (i.e., between words found in adjacent sentences) rather than just between letters. ▮

Question 1.8.14 Estimate the amount of information in a typical book (order of magnitude is sufficient). Use the best estimate of information content per character of English text of about 1 bit per character.

Solution 1.8.14 A rough estimate can be made using as follows: A 200 page novel with 60 characters per line and 30 lines per page has 4 × 10^5 characters. Textbooks can have several times this many characters.A dictionary, which is significantly longer than a typical book, might have 2 × 10^7 characters. Thus we might use an order of magnitude value of 10^6 bits per book. ▮

Question 1.8.15 Obtain an estimate of the number of characters (and thus the number of bits of information) in the Library of Congress. Assume an average of 10^6 characters per book.

Solution 1.8.15 According to information provided by the Library of Congress,there are presently (in 1996) 16 million books classified according to the Library of Congress classification system, 13 million other books at the Library of Congress, and approximately 80 million other items such as newspapers, maps and films. Thus with 10^7–10^8 book equivalents, we estimate the number of characters as 10^{13}–10^{14}. ▮

Inherent in the notion of quantifying information content is the understanding that the same information can be communicated in different ways, as long as the amount of information that can be transmitted is sufficient. Thus we can use binary, decimal, hexadecimal or typed letters to communicate both numbers and letters. Information can be communicated using any set of (two or more) characters. The presumption is that there is a way of translating from one to another. Translation operations are called codes; the act of translation is encoding or decoding. Among possible codes are those that are invertible.Encoding a message cannot add information, it might,however, lose information (Question 1.8.16). Invertible codes must preserve the amount of information.

Once we have determined the information content, we can compare different ways of writing the same information. Assume that one source generates a message of length N characters with information I. Then a different source may transmit the same information using fewer characters. Even if characters are generated at the same rate,the information may be more rapidly transmitted by one source than another. In particular, regardless of the value of N, by definition of information content, we could have communicated the same information using a binary string of length I. It is,however, impossible to use fewer than I bits because the maximum information a binary message can contain is equal to its length. This amount of information occurs for a source with equal a priori probability.

Encoding the information in a shorter form is equivalent to data compression. Thus a completely compressed binary data string would have an amount of information given by its length. The source of such a message would be characterized as a source of messages with equal a priori probability—a random source. We see that randomness and information are related. Without a translation (decoding) function it would be impossible to distinguish the completely compressed information from random numbers. Moreover, a random string could not be compressed.

Question 1.8.16 Prove that an encoding operation that takes a message as input and converts it into another well-defined message (i.e., for a particular input message, the same output message is always given) cannot add information but may reduce it. Describe the necessary conditions for it to keep the same amount of information.

Solution 1.8.16 Our definition of information relies upon the specification of the ensemble of possible messages. Consider this ensemble and assume that each message appears in the ensemble a number of times in proportion to its probability, like the bag with red and green balls. The effect of a coding operation is to label each ball with the new message (code) that will be delivered after the coding operation. The amount of information depends not on the nature of the label, but rather on the number of balls with the same label. The requirement that a particular message is encoded in a well-defined way means that two balls that start with the same message cannot be labeled with different codes. However, it is possible for balls with different original messages to be labeled the same. The average information is not changed if and only if all distinct messages are labeled with distinct codes. If any distinct messages become identified by the same label, the information is reduced.

We can prove this conclusion algebraically using the result of Question 1.8.8, which showed that transferring probability from a less likely to a more likely case reduced the information content. Here we are,in effect, transferring all of the probability from the less likely to the more likely case. The change in information upon labeling two distinct messages with the same code is given by ($f(x) = -x\log(x)$, as in Question 1.8.8):

$$\begin{aligned}\Delta I &= f(P(s_1) + P(s_2)) - (f(P(s_1)) + f(P(s_2))) \\ &= (f(P(s_1) + P(s_2)) + f(0)) - (f(P(s_1)) + f(P(s_2))) < 0\end{aligned} \tag{1.8.61}$$

where the inequality follows because $f(x)$ is convex in the range $0 < x < 1$. ▮

1.8.5 *Human communication*

The theory of information, like other theories, relies upon idealized constructs that are useful in establishing the essential concepts, but do not capture all features of real systems. In particular, the definition and discussion of information relies upon sources that transmit the result of random occurrences, which, by definition, cannot be known by the recipient. The sources are also completely described by specifying the nature of the random process. This model for the nature of the source and the recipient does not adequately capture the attributes of communication between human beings. The theory of information can be applied directly to address questions about information channels and the characterization of communication in general. It can also be used to develop an understanding of the complexity of systems. In this section, however, we will consider some additional issues that should be kept in mind when applying the theory to the communication between human beings. These issues will arise again in Chapter 8.

The definition of information content relies heavily on the concepts of probability, ensembles, and processes that generate arbitrarily many characters. These concepts are fraught with practical and philosophical difficulties—when there is only one transmitted message,how can we say there were many that were possible? A book may be considered as a single communication.A book has finite length and, for a particular author and a particular reader, is a unique communication. In order to understand both the strengths and the limitations of applying the theory of information,it is necessary to recognize that the information content of a message depends on the information that the recipient of the message already has. In particular, information that the recipient has about the source. In the discussion above,a clear distinction has been made. The only information that characterizes the source is in the ensemble probabilities $P(s)$. The information transmitted by a single message is distinct from the ensemble probabilities and is quantified by $I(s)$. It is assumed that the characterization of the source is completely known to the recipient. The content of the message is completely unknown (and unknowable in advance) to the recipient.

A slightly more difficult example to consider is that of a recipient who does not know the characterization of the source. However, such a characterization in terms of an ensemble $P(s)$ does exist. Under these circumstances, the amount of information transferred by a message would be more than the amount of information given by $I(s)$. However, the maximum amount of information that could be transferred would be the sum of the information in the message,and the information necessary to characterize the source by specifying the probabilities $P(s)$. This upper bound on the information that can be transferred is only useful if the amount of information necessary to characterize the source is small compared to the information in the message.

The difficulty with discussing human communication is that the amount of information necessary to fully characterize the source (one human being) is generally much larger than the information transmitted by a particular message. Similarly, the amount of information possessed by the recipient (another human being) is much larger than the information contained in a particular message. Thus it is reasonable to assume that the recipient does not have a full characterization of the source. It is also reasonable to assume that the model that the recipient has about the source is more sophisticated than a typical Markov chain model, even though it is a simplified model of a human being. The information contained in a message is, in a sense, the additional information not contained in the original model possessed by the recipient. This is consistent with the above discussion, but it also recognizes that specifying the probabilities of the ensemble may require a significant amount of information. It may also be convenient to summarize this information by a different type of model than a Markov chain model.

Once the specific model and information that the recipient has about the source enters into an evaluation of the information transfer, there is a certain and quite reasonable degree of relativity in the amount of information transferred. An extreme example would be if the recipient has already received a long message and knows the same message is being repeated,then no new information is being transmitted.A person who has memorized the Gettysburg Address will receive very little new information upon hearing or reading it again. The prior knowledge is part of the model possessed by the recipient about the source.

Can we incorporate this in our definition of information? In every case where we have measured the information of a message, we have made use of a model of the source of the information. The underlying assumption is that this model is possessed by the recipient. It should now be recognized that there is a certain amount of information necessary to describe this model. As long as the amount of information in the model is small compared to the amount of information in the message, we can say that we have an absolute estimate of the information content of the message. As soon as the information content of the model approaches that of the message itself, then the amount of information transferred is sensitive to exactly what information is known. It might be possible to develop a theory of information that incorporates the information in the model,and thus to arrive at a more absolute measure of information. Alternatively, it might be necessary to develop a theory that considers the recipient and source more completely, since in actual communication between human beings, both are nonergodic systems possessed of a large amount of information. There is significant overlap of the information possessed by the recipient and the source. Moreover, this common information is essential to the communication itself.

One effort to arrive at a universal definition of information content of a message has been made by formally quantifying the information contained in models. The resulting information measure, Kolmogorov complexity, is based on computation theory discussed in the next section. While there is some success with this approach,two difficulties remain. In order for a universal definition of information to be agreed upon,models must still have an information content which is less than the message—knowledge possessed must be smaller than that received. Also, to calculate the information contained in a particular message is essentially impossible, since it requires computational effort that grows exponentially with the length of the message. In any practical case,the amount of information contained in a message must be estimated using a limited set of models of the source. The utilization of a limited set of models means that any estimate of the information in a message is an upper bound.

1.9 Computation

The theory of computation describes the operations that we perform on numbers, including addition, subtraction, multiplication and division. More generally, a computation is a sequence of operations each of which has a definite/unique/well-defined result. The fundamental study of such operations is the theory of logic. Logical operations do not necessarily act upon numbers, but rather upon abstract objects called statements. Statements can be combined together using operators such as AND and OR, and acted upon by the negation operation NOT. The theory of logic and the theory of computation are at root the same. All computations that have been conceived of can be constructed out of logical operations. We will discuss this equivalence in some detail.

We also discuss a further equivalence, generally less well appreciated, between computation and deterministic time evolution. The theory of computation strives to describe the class of all possible discrete deterministic or causal systems. Computations are essentially causal relationships. Computation theory is designed to capture all such possible relationships. It is thus essential to our understanding not just of the behavior of computers, or of human logic, but also to the understanding of causal relationships in all physical systems. A counterpoint to this association of computation and causality is the recognition that certain classes of deterministic dynamical systems are capable of the property known as universal computation.

One of the central findings of the theory of computation is that many apparently different formulations of computation turn out to be equivalent. The sense in which they are equivalent is that each one can simulate the other. In the early years of computation theory, there was an effort to describe sets of operations that would be more powerful than others. When all of them were shown to be equivalent it became generally accepted (the Church-Turing hypothesis) that there is a well-defined set of possible computations realized by any of several conceptual formulations. This has become known as the theory of universal computation.

1.9.1 *Propositional logic*

Logic is the study of reasoning, inference and deduction. Propositional logic describes the manipulation of statements that are either true or false. It assumes that there exists a set of statements that are either true or false at a particular time, but not both. Logic then provides the possibility of using an assumed set of relationships between the statements to determine the truth or falsehood of other statements.

For example,the statements Q_1 = "I am standing" and Q_2 = "I am sitting" may be related by the assumption: Q_1 is true implies that Q_2 is not true. Using this assumption,it is understood that a statement "Q_1 AND Q_2" must be false. The falsehood depends only on the relationship between the two sentences and not on the particular meaning of the sentences. This suggests that an abstract construction that describes mechanisms of inference can be developed. This abstract construction is propositional logic.

Propositional logic is formed out of statements (propositions) that may be true (T) or false (F), and operations. The operations are described by their actions upon statements. Since the only concern of logic is the truth or falsehood of statements, we can describe the operations through tables of truth values (truth tables) as follows. NOT (^) is an operator that acts on a single statement (a unary operator) to form a new statement. If Q is a statement then ^Q (read "not Q") is the symbolic representation of “It is not true that Q.” The truth of ^Q is directly (causally) related to the truth of Q by the relationship in the table:

Q	^Q
T	F
F	T

(1.9.1)

The value of the truth or falsehood of Q is shown in the left column and the corresponding value of the truth or falsehood of ^Q is given in the right column.

Similarly, we can write the truth tables for the operations AND (&) and OR (|):

Q_1	Q_2	Q_1&Q_2
T	T	T
T	F	F
F	T	F
F	F	F

(1.9.2)

Q_1	Q_2	Q_1|Q_2
T	T	T
T	F	T
F	T	T
F	F	F

(1.9.3)

As the tables show, Q_1&Q_2 is only true if both Q_1 is true and Q_2 is true. $Q_1|Q_2$ is only false if both Q_1 is false and Q_2 is false.

Propositional logic includes logical theorems as statements. For example, the statement Q_1 is true if and only if Q_2 is true can also be written as a binary operation $Q_1 \equiv Q_2$ with the truth table:

Q_1	Q_2	$Q_1 \equiv Q_2$
T	T	T
T	F	F
F	T	F
F	F	T

(1.9.4)

Another binary operation is the statement Q_1 implies Q_2, Q_1 Q_2. When this statement is translated into propositional logic, there is a difficulty that is usually bypassed by the following convention:

Q_1	Q_2	Q_1 Q_2
T	T	T
T	F	F
F	T	T
F	F	T

(1.9.5)

The difficulty is that the last two lines suggest that when the antecedent Q_1 is false, the implication is true, whether or not the consequent Q_2 is true. For example, the statement "If I had wings then I could fly" is as true a statement as "If I had wings then I couldn't fly," or the statement "If I had wings then potatoes would be flat." The problem originates in the necessity of assuming that the result is true or false in a unique way based upon the truth values of Q_1 and Q_2. Other information is not admissible, and a third choice of "nonsense" or "incomplete information provided" is not allowed within propositional logic. Another way to think about this problem is to say that there are many operators that can be formed with definite outcomes. Regardless of how we relate these operators to our own logical processes, we can study the system of operators that can be formed in this way. This is a model, but not a complete one, for human logic. Or, if we choose to define logic as described by this system, then human thought (as reflected by the meaning of the word "implies") is not fully characterized by logic (as reflected by the meaning of the operation "   ").

In addition to unary and binary operations that can act upon statements to form other statements, it is necessary to have parentheses that differentiate the order of operations to be performed. For example a statement $((Q_1 \equiv Q_2)\&(\wedge Q_3)|Q_1)$ is a series of operations on primitive statements that starts from the innermost parenthesis and progresses outward. As in this example, there may be more than one innermost parenthesis. To be definite, we could insist that the order of performing these operations is from left to right. However, this order does not affect any result.

Within the context of propositional logic, it is possible to describe a systematic mechanism for proving statements that are composed of primitive statements. There are several conclusions that can be arrived at regarding a particular statement. A tautology is a statement that is always true regardless of the truth or falsehood of its component statements. Tautologies are also called theorems. A contradiction is a statement that is always false. Examples are given in Question 1.9.1.

Question 1.9.1 Evaluate the truth table of:

a. $(Q_1 \quad Q_2)|((\wedge Q_2)\&Q_1)$

b. $(\wedge(Q_1 \quad Q_2))\equiv((\wedge Q_1)|Q_2)$

Identify which is a tautology and which is a contradiction.

Solution 1.9.1 Build up the truth table piece by piece:

a. Tautology:

Q_1	Q_2	$Q_1 \quad Q_2$	$(\wedge Q_2)\&Q_1$	$(Q_1 \quad Q_2)\|((\wedge Q_2)\&Q_1)$
T	T	T	F	T
T	F	F	T	T
F	T	T	F	T
F	F	T	F	T

(1.9.6)

b. Contradiction:

Q_1	Q_2	$\hat{}(Q_1 \quad Q_2)$	$(\hat{}Q_1)\vert Q_2$	$(\hat{}(Q_1 \quad Q_2)) \equiv ((\hat{}Q_1)\vert Q_2)$
T	T	F	T	F
T	F	T	F	F
F	T	F	T	F
F	F	F	T	F

(1.9.7) ▮

Question 1.9.2: Construct a theorem (tautology) from a contradiction.

Solution 1.9.2: By negation. ▮

1.9.2 *Boolean algebra*

Propositional logic is a particular example of a more general symbolic system known as a Boolean algebra. Set theory, with the operators complement,union and intersection, is another example of a Boolean algebra. The formulation of a Boolean algebra is convenient because within this more general framework a number of important theorems can be proven. They then hold for propositional logic,set theory and other Boolean algebras.

A Boolean algebra is a set of elements $B=\{Q_1,Q_2, \ldots\}$, a unary operator ($\hat{}$), and two binary operators, for which we adopt the notation (+,•),that satisfy the following properties for all Q_1, Q_2, Q_3 in B:

1. Closure: $\hat{}Q_1$, Q_1+Q_2, and $Q_1 \bullet Q_2$ are in B
2. Commutative law: $Q_1+Q_2=Q_2+Q_1$, and $Q_1 \bullet Q_2=Q_2 \bullet Q_1$
3. Distributive law: $Q_1 \bullet (Q_2+Q_3)=(Q_1 \bullet Q_2)+(Q_1 \bullet Q_3)$ and $Q_1+(Q_2 \bullet Q_3)=(Q_1+Q_2)\bullet(Q_1+Q_3)$
4. Existence of identity elements, 0 and 1: $Q_1+0=Q_1$, and $Q_1 \bullet 1=Q_1$
5. Complementarity law: $Q_1+(\hat{}Q_1)=1$ and $Q_1 \bullet (\hat{}Q_1)=0$

The statements of properties 2 through 5 consist of equalities. These equalities indicate that the element of the set that results from operations on the left is the same as the element resulting from operations on the right. Note particularly the second part of the distributive law and the complementarity law that would not be valid if we interpreted + as addition and • as multiplication.

Assumptions 1 to 5 allow the proof of additional properties as follows:

6. Associative property: $Q_1+(Q_2+Q_3)=(Q_1+Q_2)+Q_3$ and $Q_1 \bullet (Q_2 \bullet Q_3)=(Q_1 \bullet Q_2)\bullet Q_3$
7. Idempotent property: $Q_1+Q_1=Q_1$ and $Q_1 \bullet Q_1=Q_1$
8. Identity elements are nulls: $Q_1+1=1$ and $Q_1 \bullet 0=0$
9. Involution property: $\hat{}(\hat{}Q_1)=Q_1$

10. Absorption property: Q_1+(Q_1•Q_2)=Q_1 and Q_1•(Q_1+Q_2)=Q_1
11. DeMorgan's Laws: ^(Q_1+Q_2)=(^Q_1)•(^Q_2) and ^(Q_1•Q_2)=(^Q_1)+(^Q_2)

To identify propositional logic as a Boolean algebra we use the set B={T,F} and map the operations of propositional logic to Boolean operations as follows:(^ to ^), (| to +) and (& to •). The identity elements are mapped:(1 to T) and (0 to F). The proof of the Boolean properties for propositional logic is given as Question 1.9.3.

Question 1.9.3: Prove that the identification of propositional logic as a Boolean algebra is correct.

Solution 1.9.3: (1) is trivial; (2) is the invariance of the truth tables of Q_1&Q_2, Q_1|Q_2 to interchange of values of Q_1 and Q_2; (3) requires comparison of the truth tables of Q_1|(Q_2&Q_3) and (Q_1|Q_2)&(Q_1|Q_3) (see below). Comparison of the truth tables of Q_1&(Q_2|Q_3) and (Q_1&Q_2)|(Q_1&Q_3) is done similarly.

Q_1	Q_2	Q_3	$Q_2\&Q_3$	$Q_1\vert(Q_2\&Q_3)$	$Q_1\vert Q_2$	$Q_1\vert Q_3$	$(Q_1\vert Q_2)\&(Q_1\vert Q_3)$
T	T	T	T	T	T	T	T
T	T	F	F	T	T	T	T
T	F	T	F	T	T	T	T
T	F	F	F	T	T	T	T
F	T	T	T	T	T	T	T
F	T	F	F	F	T	F	F
F	F	T	F	F	F	T	F
F	F	F	F	F	F	F	F

(1.9.8)

(4) requires verifying Q_1&T=T, and Q_1|F=F (see the truth tables for & and | above);(5) requires constructing a truth table for Q|^Q and verifying that it is always T (see below). Similarly, the truth table for Q&^Q shows that it is always F.

Q	^Q	Q|^Q
T	F	T
F	T	T

(1.9.9) ▮

1.9.3 *Completeness*

Our objective is to show that an arbitrary truth table, an arbitrary logical statement, can be constructed out of only a few logical operations. Truth tables are also equivalent to numerical functions—specifically, functions of binary variables that have binary results (binary functions of binary variables). This can be seen using the Boolean representation of T and F as {1,0} that is more familiar as a binary notation for numerical functions. For example, we can write the AND and OR operations (functions) also as:

Q_1	Q_2	$Q_1\bullet Q_2$	Q_1+Q_2
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0

(1.9.10)

Similarly for all truth tables,a logical operation is a binary function of a set of binary variables. Thus,the ability to form an arbitrary truth table from a few logical operators is the same as the ability to form an arbitrary binary function of binary variables from these same logical operators.

To prove this ability, we use the properties of the Boolean algebra to systematically discuss truth tables. We first construct an alternative Boolean expression for Q_1+Q_2 by a procedure that can be generalized to arbitrary truth tables. The procedure is to look at each line in the truth table that contains an outcome of 1 and write an expression that provides unity for that line only. Then we combine the lines to achieve the desired table. $Q_1\bullet Q_2$ is only unity for the first line,as can be seen from its column. Similarly, $Q_1\bullet({}^\wedge Q_2)$ is unity for the second line and $({}^\wedge Q_1)\bullet Q_2$ is unity for the third line. Using the properties of + we can then combine the terms together in the form:

$$Q_1\bullet Q_2+Q_1\bullet({}^\wedge Q_2)+({}^\wedge Q_1)\bullet Q_2 \qquad (1.9.11)$$

Using associative and identity properties, this gives the same result as Q_1+Q_2.

We have replaced a simple expression with a much more complicated expression in Eq.(1.9.11). The motivation for doing this is that the same procedure can be used to represent any truth table. The general form we have constructed is called the disjunctive normal form. We can construct a disjunctive normal representation for an arbitrary binary function of binary variables. For example, given a specific binary function of binary variables, $f(Q_1,Q_2,Q_3)$, we construct its truth table, e.g.,

Q_1	Q_2	Q_3	$f(Q_1,Q_2,Q_3)$
1	1	1	1
1	0	1	0
0	1	1	1
0	0	1	0
1	1	0	0
1	0	0	1
0	1	0	0
0	0	0	0

(1.9.12)

The disjunctive normal form is given by:

$$f(Q_1,Q_2,Q_3)=Q_1\bullet Q_2\bullet Q_3+({}^\wedge Q_1)\bullet Q_2\bullet Q_3+Q_1\bullet({}^\wedge Q_2)\bullet({}^\wedge Q_3) \qquad (1.9.13)$$

as can be verified by inspection. An analogous construction can represent any binary function.

We have demonstrated that an arbitrary truth table can be constructed out of the three operations (^,+, •). We say that these form a complete set of operations. Since there are 2^n lines in a truth table formed out of n binary variables, there are 2^{2^n} possible functions of these n binary variables. Each is specified by a particular choice of the 2^n possible outcomes. We have achieved a dramatic simplification by recognizing that all of them can be written in terms of only three operators. We also know that at most $(1/2)n2^n$ (^) operations, $(n-1)\,2^n$ (•) operations and 2^n-1 (+) operations are necessary. This is the number of operations needed to represent the identity function 1 in disjunctive normal form.

It is possible to further simplify the set of operations required. We can eliminate either the + or the • operations and still have a complete set. To prove this we need only display an expression for either of them in terms of the remaining operations:

$$\begin{gathered} Q_1 \bullet Q_2 = \textasciicircum((\textasciicircum Q_1)+(\textasciicircum Q_2)) \\ Q_1 + Q_2 = \textasciicircum((\textasciicircum Q_1)\bullet(\textasciicircum Q_2)) \end{gathered} \tag{1.9.14}$$

Question 1.9.4: Verify Eq. (1.9.14).

Solution 1.9.4: They may be verified using DeMorgan's Laws and the involution property, or by construction of the truth tables, e.g.:

Q_1	Q_2	$\textasciicircum Q_1$	$\textasciicircum Q_2$	$Q_1 \bullet Q_2$	$(\textasciicircum Q_1)+(\textasciicircum Q_2)$
1	1	0	0	1	0
1	0	0	1	0	1
0	1	1	0	0	1
0	0	1	1	0	1

(1.9.15) ▮

It is possible to go one step further and identify binary operations that can represent all possible functions of binary variables. Two possibilities are the NAND ($\hat{\&}$) and NOR ($\hat{|}$) operations defined by:

$$\begin{gathered} Q_1 \,\hat{\&}\, Q_2 = \textasciicircum(Q_1 \& Q_2) \rightarrow \textasciicircum(Q_1 \bullet Q_2) \\ Q_1 \,\hat{|}\, Q_2 = \textasciicircum(Q_1 | Q_2) \rightarrow \textasciicircum(Q_1 + Q_2) \end{gathered} \tag{1.9.16}$$

Both the logical and Boolean forms are written above. The truth tables of these operators are:

Q_1	Q_2	$\textasciicircum(Q_1 \bullet Q_2)$	$\textasciicircum(Q_1+Q_2)$
1	1	0	0
1	0	1	0
0	1	1	0
0	0	1	1

(1.9.17)

We can prove that each is complete by itself (capable of representing all binary functions of binary variables) by showing that they are capable of representing one of the earlier complete sets. We prove the case for the NAND operation and leave the NOR operation to Question 1.9.5.

$$\begin{aligned} \wedge Q_1 &= \wedge(Q_1 \bullet Q_1) = Q_1 \,\hat{\&}\, Q_1 \\ (Q_1 \bullet Q_2) &= \wedge(\wedge(Q_1 \bullet Q_2)) = \wedge(Q_1 \,\hat{\&}\, Q_2) = (Q_1 \,\hat{\&}\, Q_2) \,\hat{\&}\, (Q_1 \,\hat{\&}\, Q_2) \end{aligned} \tag{1.9.18}$$

Question 1.9.5: Verify completeness of the NOR operation.

Solution 1.9.5: We can use the same formulas as in the proof of the completeness of NAND by replacing • with + and $\hat{\&}$ with $\hat{|}$ everywhere. ▮

1.9.4 *Turing machines*

We have found that logical operators can represent any binary function of binary variables. This means that all well-defined mathematical operations on integers can be represented in this way. One of the implications is that we might make machines out of physical elements, each of which is capable of performing a Boolean operation. Such a machine would calculate a mathematical function and spare us a tedious task. We can graphically display the operations of a machine performing a series of Boolean operations as shown in Fig. 1.9.1. This is a simplified symbolic form similar to forms used in the design of computer logic circuits.

By looking carefully at Fig. 1.9.1 we see that there are several additional kinds of actions that are necessary in addition to the elementary Boolean operation. These actions are indicated by the lines that might be thought of as wires. One action is to transfer information from the location where it is input into the system, to the place where it is used. The second is to duplicate the information. Duplication is represented in the figure by a branching of the lines. The branching enables the same

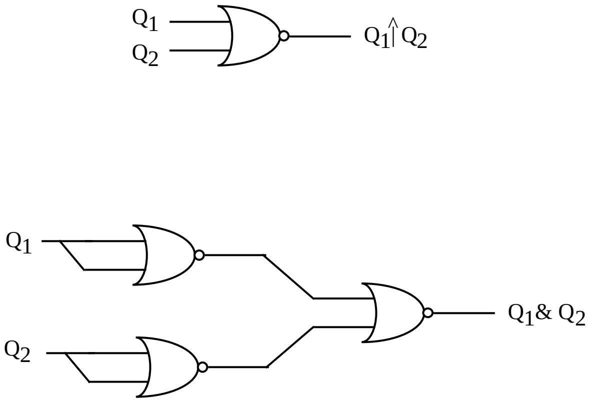


Figure 1.9.1 Graphical representation of Boolean operations. The top figure shows a graphical element representing the NOR operation $Q_1 \hat{|} Q_2 = \wedge(Q_1 | Q_2)$. In the bottom figure we combine several operations together with lines (wires) indicating input, output, data duplication and transfer to form the AND operation, $(Q_1 \hat{|} Q_1) \hat{|} (Q_2 \hat{|} Q_2) = (\wedge Q_1) \hat{|} (\wedge Q_2) = Q_1 \& Q_2$. This equation may be used to prove completeness of the NOR operation. ▮

information to be used in more than one place. Additional implicit actions involve timing, because the representation makes an assumption that time causes the information to be moved and acted upon in a sequence from left to right. It is also necessary to have mechanisms for input and output.

The kind of mathematical machine we just described is limited to performing one prespecified function of its inputs. The process of making machines is time consuming. To physically rearrange components to make a new function would be inconvenient. Thus it is useful to ask whether we might design a machine such that part of its input could include a specification of the mathematical operation to be performed. Both information describing the mathematical function,and the numbers on which it is to be performed, would be encoded in the input which could be described as a string of binary characters.

This discussion suggests that we should systematically consider the properties/ qualities of machines able to perform computations. The theory of computation is a self-consistent discussion of abstract machines that perform a sequence of prespecified well-defined operations. It extends the concept of universality that was discussed for logical operations. While the theory of logic determined that all Boolean functions could be represented using elementary logic operations, the theory of computation endeavors to establish what is possible to compute using a sequence of more general elementary operations. For this discussion many of the practical matters of computer design are not essential. The key question is to establish a relationship between machines that might be constructed and mathematical functions that may be computed. Part of the problem is to define what a computation is.

There are several alternative models of computation that have been shown to be equivalent in a formal sense since each one of them can simulate any other. Turing introduced a class of machines that represent a particular model of computation. Rather than maintaining information in wires, Turing machines (Fig. 1.9.2) use a storage device that can be read and written to. The storage is represented as an infinite one-dimensional tape marked into squares. On the tape can be written characters, one to a square. The total number of possible characters, the alphabet, is finite. These characters are often taken to be digits plus a set of markers (delimiters). In addition to the characters,the tape squares can also be blank. All of the tape is blank except for a finite number of nonblank places. Operations on the tape are performed by a roving read-write head that has a specified (finite) number of internal storage elements and a simple kind of program encoded in it. We can treat the program as a table similar to the tables discussed in the context of logic. The table operation acts upon the value of the tape at the current location of the head,and the value of storage elements within the read head. The result of an operation is not just a single binary value. Instead it corresponds to a change in the state of the tape at the current location (write),a change in the internal memory of the head,and a shift of the location of the head by one square either to the left or to the right.

We can also think about a Turing machine (TM) as a dynamic system. The internal table does not change in time. The internal state $s(t)$,the current location $l(t)$,the current character $a(t)$ and the tape $c(t)$ are all functions of time. The table consists of

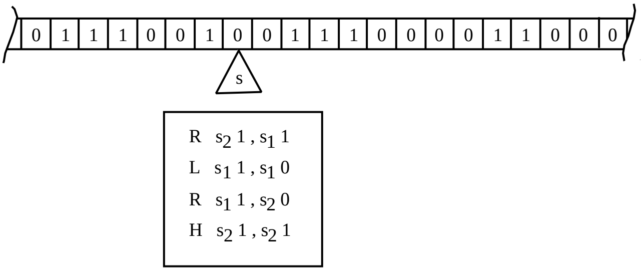


Figure 1.9.2 Turing's model of computation — the Turing machine (TM) — consists of a tape divided into squares with characters of a finite alphabet written on it. A roving "head" indicated by the triangle has a finite number of internal states and acts by reading and writing the tape according to a prespecified table of rules. Each rule consists of a command to read the tape, write the tape, change the internal state of the TM head and move either to the left or right. A simplified table is shown consisting of several rules of the form $\{\mu, s', a', s, a\}$ where a and a' are possible tape characters, s and s' are possible states of the head and μ is a movement of the head right (R), left (L) or halt (H). Each update the TM starts by finding the rule $\{\mu, s', a', s, a\}$ in the table such that a is the character on the tape at the current location of the head, and s is its current state. The tape is written with the corresponding a' and the state of the TM head is changed to s'. Then the TM head moves according to the corresponding μ right or left. The illustration simplifies the characters to binary digits 0 and 1 and the states of the TM head to s_1 and s_2. ▮

a set of instructions or rules of the form $\{\mu,s',a',s,a\}$ corresponding to a deterministic transition matrix. s and a are the current internal state and current tape character respectively. s' and a' are the new internal state and character. μ is the move to be made, either right or left (R or L).

Using either conceptual model, the TM starts from an initial state and location and a specified tape. In each time interval the TM head performs the following operations:

1. Read the current tape character
2. Find the instruction that corresponds to the existing combination of (s,a)
3. Change the internal memory to the corresponding s'
4. Write the tape with the corresponding character a'
5. Move the head to the left or right as specified by μ

When the TM head reaches a special internal state known as the halt state, then the outcome of the computation may be read from the tape. For simplicity, in what follows we will indicate entering the halt state by a move μ = H which is to halt.

The best way to understand the operation of a TM is to construct particular tables that perform particular actions (Question 1.9.6). In addition to logical operations, the possible actions include moving and copying characters. Constructing particular actions using a TM is tedious, in large part because the movements of the head are limited to a single displacement right or left. Actual computers use direct addressing that enables access to a particular storage location in its memory using a number (address) specifying its location. TMs do not generally use this because the tape is arbitrarily long, so that an address is an arbitrarily large number, requiring an arbitrarily large storage in the internal state of the head. Infinite storage in the head is not part of the computational model.

Question 1.9.6 The following TM table is designed to move a string of binary characters (0 and 1) that are located to the left of a special marker M to blank squares on the tape to the right of the M and then to stop on the M. Blank squares are indicated by B. The internal states of the head are indicated by s_1, s_2 . . . These are not italicized, since they are values rather than variables. The movements of the head right and left are indicated by R and L. As mentioned above, we indicate entering the halt state by a movement H. Each line has the form $\{\mu, s', a', s, a\}$.

Read over the program and convince yourself that it does what it is supposed to. Describe how it works. The TM must start from state s_1 and must be located at the leftmost nonblank character. The line numbering is only for convenience in describing the TM, and has no role in its operation.

1.	R	s_2	B	s_1	0	
2.	R	s_3	B	s_1	1	
3.	R	s_2	0	s_2	0	
4.	R	s_2	1	s_2	1	
5.	R	s_2	M	s_2	M	
6.	R	s_3	0	s_3	0	
7.	R	s_3	1	s_3	1	
8.	R	s_3	M	s_3	M	(1.9.19)
9.	L	s_4	0	s_2	B	
10.	L	s_4	1	s_3	B	
11.	L	s_4	0	s_4	0	
12.	L	s_4	1	s_4	1	
13.	L	s_4	M	s_4	M	
14.	R	s_1	B	s_4	B	
15.	H	s_1	M	s_1	M	

Solution 1.9.6 This TM works by (lines 1 or 2) reading a nonblank character (0 or 1) into the internal state of the head; 0 is represented by s_2 and 1 is represented by s_3. The character that is read is set to a blank B. Then the TM moves to the right, ignoring all of the tape characters 0, 1 or M (lines 3 through 8) until it reaches a blank B. It writes the stored character (lines 9 or 10), changing its state to s_4. This state specifies moving to the left,ignoring all characters 0,1 or M (lines 11 through 13) until it reaches a blank B. Then

(line 14) it moves one step right and resets its state to s_1. This starts the procedure from the beginning. If it encounters the marker M in the state s_1 instead of a character to be copied, then it halts (line 15). ∎

Since each character can also be represented by a set of other characters (i.e.,2 in binary is 10), we can allow the TM head to read and write not one but a finite prespecified number of characters without making a fundamental change. The following TM, which acts upon pairs of characters and moves on the tape by two characters at a time, is the same as the one given in Question 1.9.6.

1.	01	01	00	00	01
2.	01	11	00	00	11
3.	01	01	01	01	01
4.	01	01	11	01	11
5.	01	01	10	01	10
6.	01	11	01	11	01
7.	01	11	11	11	11
8.	01	11	10	11	10
9.	10	10	01	01	00
10.	10	10	11	11	00
11.	10	10	01	10	01
12.	10	10	11	10	11
13.	10	10	10	10	10
14.	01	00	00	10	00
15.	00	00	10	00	10

(1.9.20)

The particular choice of the mapping from characters and internal states onto the binary representation is not unique. This choice is characterized by using the left and right bits to represent different aspects. In columns 3 or 5, which represent the tape characters, the right bit represents the type of element (marker or digit), and the left represents which element or marker it is: 00 represents the blank B, 10 represents M, 01 represents the digit 0,and 11 represents the digit 1. In columns 2 or 4, which represent the state of the head,the states s_1 and s_4 are represented by 00 and 10, s_2 and s_3 are represented by 01 and 11 respectively. In column 1, moving right is 01, left is 10, and halt is 00.

The architecture of a TM is very general and allows for a large variety of actions using complex tables. However, all TMs can be simulated by transferring all of the responsibility for the table and data to the tape.A TM that can simulate all TMs is called a universal Turing machine (UTM). As with other TMs,the responsibility of arranging the information lies with the "programmer." The UTM works by representing the table,current state,and current letter on the UTM tape. We will describe the essential concepts in building a UTM but will not explicitly build one.

The UTM acts on its own set of characters with its own set of internal states. In order to use it to simulate an arbitrary TM, we have to represent the TM on the tape of the UTM in the characters that the UTM can operate on. On the UTM tape, we

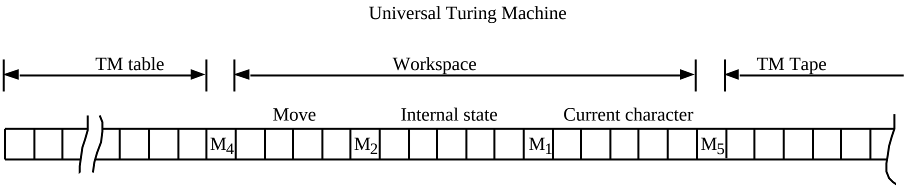


Figure 1.9.3 The universal Turing machine (UTM) is a special TM that can simulate the computation performed by any other TM. The UTM does this by executing the rules of the TM that are encoded on the tape of the UTM. There are three parts to the UTM tape, the part where the TM table is encoded (on the left), the part where the tape of the TM is encoded (on the right) and a workspace (in the middle) where information representing the current state of the TM head, the current character of the TM tape, and the movement command, are encoded. See the text for a description of the operation of the UTM based on its own rule table. ▮

must be able to represent four types of entities: a TM character, the state of the TM head, the movement to be taken by the TM head, and markers that indicate to the UTM what is where on the tape. The markers are special to the UTM and must be carefully distinguished from the other three. For later reference, we will build a particular type of UTM where the tape can be completely represented in binary.

The UTM tape has three parts,the part that represents the table of the TM,a work area,and the part that represents the tape of the TM (Fig. 1.9.3). To represent the tape and table of a particular but arbitrary TM, we start with a binary representation of its alphabet and of its internal states

$$\begin{array}{c} a_1 \rightarrow 00000, a_2 \rightarrow 00001, a_3 \rightarrow 00010, \ldots \\ s_1 \rightarrow 000, s_2 \rightarrow 001, \ldots \end{array} \qquad (1.9.21)$$

where we keep the left zeros, as needed for the number of bits in the longest binary number. We then make a doubled binary representation like that used in the previous example, where each bit becomes two bits with the low order bit a 1. The doubled binary notation will enable us to distinguish between UTM markers and all other entities on the tape. Thus we have:

$$\begin{array}{c} a_1 \rightarrow 01\ 01\ 01\ 01\ 01, a_2 \rightarrow 01\ 01\ 01\ 01\ 11, a_3 \rightarrow 01\ 01\ 01\ 11\ 01, \ldots \\ s_1 \rightarrow 01\ 01\ 01, s_2 \rightarrow 01\ 01\ 11, \ldots \end{array} \qquad (1.9.22)$$

These labels of characters and states are in a sense arbitrary, since the transition table is what gives them meaning.

We also encode the movement commands. The movement commands are not arbitrary, since the UTM must know how to interpret them. We have allowed the TM to displace more than one character, so we must encode a set of movements such as R_1, L_1, R_2, L_2, and H. These correspond respectively to moving one character right, one character left, two characters right, two characters left, and entering the halt state. Because the UTM must understand the move that is to be made, we must agree once and for all on a coding of these movements. We use the lowest order bit as a direction bit (1 Right, 0 Left) and the rest of the bits as the number of displacements in binary

$$\begin{aligned} R_1 &\rightarrow 011, R_2 \rightarrow 101, \ldots, \\ L_1 &\rightarrow 010, L_2 \rightarrow 100, \ldots, \\ H &\rightarrow 000 \text{ or } 001 \end{aligned} \tag{1.9.23}$$

The doubled binary representation is as before: each bit becomes two bits with the low order bit a 1,

$$\begin{aligned} R_1 &\rightarrow 01\ 11\ 11\ , R_2 \rightarrow 11\ 01\ 11\ , \ldots, \\ L_1 &\rightarrow 01\ 11\ 01\ , L_2 \rightarrow 11\ 01\ 01\ , \ldots, \\ H &\rightarrow 01\ 01\ 01 \text{ or } 01\ 01\ 11 \end{aligned} \tag{1.9.24}$$

Care is necessary in the UTM design because we do not know in advance how many types of TM moves are possible. We also don't know how many characters or internal states the TM has. This means that we don't know the length of their binary representations.

We need a number of markers that indicate to the UTM the beginning and end of encoded characters, states and movements described above. We also need markers to distinguish different regions of the tape. A sufficient set of markers are:

M_1—the beginning of a TM character,

M_2—the beginning of a TM internal state,

M_3—the beginning of a TM table entry, which is also the beginning of a movement command,

M_4—a separator between the TM table and the workspace,

M_5—a separator between the workspace and the TM tape,

M_6—the beginning of the current TM character (the location of the TM head),

M_7—the identified TM table entry to be used in the current step, and

B—the blank, which we include among the markers.

Depending on the design of the UTM, these markers need not all be distinct. In any case, we encode them also in binary

$$B \rightarrow 000, M_1 \rightarrow 001, M_2 \rightarrow 010, \ldots \tag{1.9.25}$$

and then doubled binary form where the second character is now zero:

$$B \rightarrow 00\ 00\ 00, M_1 \rightarrow 00\ 00\ 10, M_2 \rightarrow 00\ 10\ 00, \ldots \tag{1.9.26}$$

We are now in a position to encode both the tape and table of the TM on the tape of the UTM. The representation of the table consists of a sequence of representations of the lines of the table, L_1L_2..., where each line is represented by the doubled binary representation of

$$M_3\, \mu\, M_2\, s'\, M_1\, a'\, M_2\, s\, M_1 a \tag{1.9.27}$$

The markers are definite but the characters and states and movements correspond to those in a particular line in the table. The UTM representation of the tape of the TM, $a_1a_2 \ldots$, is a doubled binary representation of

$$\mathrm{M}_1\, a_1\, \mathrm{M}_1\, a_2\, \mathrm{M}_1\, a_3 \ldots \tag{1.9.28}$$

The workspace starts with the character M_4 and ends with the character M_5. There is room enough for the representation of the current TM machine state,the current tape character and the movement command to be executed. At a particular time in execution it appears as:

$$\mathrm{M}_4\, \mu\, \mathrm{M}_2\, s\, \mathrm{M}_1\, a\, \mathrm{M}_5 \tag{1.9.29}$$

We describe in general terms the operation of the UTM using this representation of a TM. Before execution we must indicate the starting location of the TM head and its initial state. This is done by changing the corresponding marker M_1 to M_6 (at the UTM tape location to the left of the character corresponding to the initial location of the TM), and the initial state of the TM is encoded in the workspace after M_2.

The UTM starts from the leftmost nonblank character of its tape. It moves to the right until it encounters M_6. It then copies the character after M_6 into the work area after M_1. It compares the values of (s,a) in the work area with all of the possible (s,a) pairs in the transition table pairs until it finds the same pair. It marks this table entry with M_7. The corresponding s' from the table is copied into the work area after M_2. The corresponding a' is copied to the tape after M_6. The corresponding movement command μ is copied to the work area after M_4. If the movement command is H the TM halts. Otherwise, the marker M_6 is moved according to the value of μ. It is moved one step at a time (i.e.,the marker M_6 is switched with the adjacent M_1) while decrementing the value of the digits of μ (except the rightmost bit) and in the direction specified by the rightmost bit. When the movement command is decremented to zero, the TM begins the cycle again by copying the character after M_6 into the work area.

There is one detail we have overlooked: the TM can write to the left of its nonblank characters. This would cause problems for the UTM we have designed,since to the left of the TM tape representation is the workspace and TM table. There are various ways to overcome this difficulty. One is to represent the TM tape by folding it upon itself and interleaving the characters.Starting from an arbitrary location on the TM tape we write all characters on the UTM tape to the right of M_5 , so that odd characters are the TM tape to the right, and even ones are the TM tape to the left. Movements of the M_6 marker are doubled, and it is reflected (bounces) when it encounters M_5.

A TM is a dynamic system. We can reformulate Turing's model of computation in the form of a cellular automaton (Section 1.5) in a way that will shed some light on the dynamics that are being discussed. The most direct way to do this is to make an automaton with two adjacent tapes. The only information in the second strip is a single nonblank character at the location of the head that represents its internal state. The TM update is entirely contained within the update rule of the automaton. This update rule may be constructed so that it acts at every point in the space, but is enabled by the nonblank character in the adjacent square on the second tape. When the dynamics reaches a steady state (it is enough that two successive states of the automaton are the same),the computation is completed. If desired we could reduce this CA to one tape by placing each pair of squares in the two tapes adjacent to each other, interleaving the two tapes. While a TM can be represented as a CA,any CA with only a finite number of active cells can be updated by a Turing machine program (it is computable). There are many other CA that can be programmed by their initial state to perform computations. These can be much simpler than using the TM model as a starting point. One example is Conway's Game of Life, discussed in Section 1.5.Like a UTM, this CA is a universal computer—any computation can be performed by starting from some initial state and looking at the final steady state for the result.

When we consider the relationship of computation theory to dynamic systems, there are some intentional restrictions in the theory that should be recognized. The conventional theory of computation describes a single computational unit operating on a character string formed from a finite alphabet of characters. Thus, computation theory does not describe a continuum in space,an infinite array of processors, or real numbers. Computer operations only mimic approximately the formal definition of real numbers. Since an arbitrary real number requires infinitely many digits to specify, computations upon them in finite time are impossible. The rejection by computation theory of operations upon real numbers is not a trivial one. It is rooted in fundamental results of computation theory regarding limits to what is inherently possible in any computation.

This model of computation as dynamics can be summarized by saying that a computation is the steady-state result of a deterministic CA with a finite alphabet (finite number of characters at each site) and finite domain update rule.One of the characters (the blank or vacuum) must be such that it is unchanged when the system is filled with these characters. The space is infinite but the conditions are such that all space except for a finite region must be filled with the blank character.

1.9.5 *Computability and the halting problem*

The construction of a UTM guarantees that if we know how to perform a particular operation on numbers, we can program a UTM to perform this computation. However, if someone gives you such a program—can you determine what it will compute? This seemingly simple question turns out to be at the core of a central problem of logic theory. It turns out that it is not only difficult to determine what it will compute,it is,in a formal sense that will be described below, impossible to figure out if it will compute anything at all. The requirement that it will compute something is that eventually it will halt. By halting, it declares its computation completed and the answer given. Instead of halting, it might loop forever or it might continue to write on ever larger regions of tape. To say that we can determine whether it will compute something is equivalent to saying that it will eventually halt. This is called the halting problem. How could we determine if it would halt? We have seen above how to represent an arbitrary TM on the tape of a particular TM. Consistent with computation theory, the halting problem is to construct a special TM, T_H, whose input is a description of a TM and whose output is a single bit that specifies whether or not the

TM will halt. In order for this to make sense,the program T_H must itself halt. We can prove by contradiction that this is not possible in general, and therefore we say that the halting problem is not computable. The proof is based on constructing a paradoxical logical statement of the form “This statement is false.”

A proof starts by assuming we have a TM called T_H that accepts as input a tape representing a TM Y and its tape y. The output, which can be represented in functional form as $T_H(Y,y)$, is always well-defined and is either 1 or 0 representing the statement that the TM Y halts on y or doesn’t halt on y respectively. We now construct a logical contradiction by constructing an additional TM based on T_H. First we consider $T_H(Y,Y)$, which asks whether Y halts when acting on a tape representing itself. We design a new TM T_{H1} that takes only Y as input, copies it and then acts in the same way as T_H. So we have

$$T_{H1}(Y) = T_H(Y,Y) \tag{1.9.30}$$

We now define a TM T_{H2} that is based on T_{H1} but whenever T_{H1} gives the answer 0 it gives the answer 1,and whenever T_{H1} gives the answer 1 it enters a loop and computes forever. A moment’s meditation shows that this is possible if we have T_{H1}. Applying T_{H2} to itself then gives us the contradiction, since $T_{H2}(T_{H2})$ gives 1 if

$$T_{H1}(T_{H2}) = T_H(T_{H2},T_{H2}) = 0 \tag{1.9.31}$$

By definition of T_H this means that $T_{H2}(T_{H2})$ does not halt, which is a contradiction. Alternatively, $T_{H2}(T_{H2})$ computes forever if

$$T_{H1}(T_{H2}) = T_H(T_{H2},T_{H2}) = 1$$

by definition of T_H this means that $T_{H2}(T_{H2})$ halts, which is a contradiction.

The noncomputability of the halting problem is similar to Gödel’s theorem and other results denying the completeness of logic, in the sense that we can ask a question about a logical construction that cannot be answered by it.Gödel’s theorem may be paraphrased as: In any axiomatic formulation of number theory (i.e.,integers),it is possible to write a statement that cannot be proven T or F. There has been a lot of discussion about the philosophical significance of these theorems.A basic conclusion that may be reached is that they describe something about the relationship of the finite and infinite. Turing machines can be represented,as we have seen, by a finite set of characters. This means that we can enumerate them, and they correspond one-to-one to the integers. Like the integers, there are (countably) infinitely many of them. Gödel’s theorem is part of our understanding of how an infinite set of numbers must be described. It tells us that we cannot describe their properties using a finite set of statements. This is appealing from the point of view of information theory since an arbitrary integer contains an arbitrarily large amount of information. The noncomputability of the halting problem tells us more specifically that we can ask a question about a system that is described by a finite amount of information whose answer (in the sense of computation) is not contained within it. We have thus made a vague connection between computation and information theory. We take this connection one step further in the following section.

1.9.6 *Computation and information in brief*

One of our objectives will be to relate computation and information. We therefore ask, Can a calculation produce information? Let us think about the results of a TM calculation which is a string of characters—the nonblank characters on the output tape. How much information is necessary to describe it? We could describe it directly, or use a Markov model as in Section 1.8. However, we could also give the input of the TM and the TM description, and this would be enough information to enable us to obtain the output by computation. This description might contain more or fewer characters than the direct description of the output. We now return to the problem of defining the information content of a string of characters. Utilizing the full power of computation, we can define this as the length of the shortest possible input tape for a UTM that gives the desired character string as its output. This is called the algorithmic (or Kolmogorov) complexity of a character string. We have to be careful with the definition, since there are many different possible UTM. We will discuss this in greater detail in Chapter 8. However, this discussion does imply that a calculation cannot produce information. The information present at the beginning is sufficient to obtain the result of the computation. It should be understood, however, that the information that seems to us to be present in a result may be larger than the original information unless we are able to reconstruct the starting point and the TM used for the computation.

1.9.7 *Logic, computation and human thought*

Both logic and computation theory are designed to capture aspects of human thought. A fundamental question is whether they capture enough of this process—are human beings equivalent to glorified Turing machines? We will ask this question in several ways throughout the text and arrive at various conclusions,some of which support this identification and some that oppose it.One way to understand the question is as one of progressive approximation. Logic was originally designed to model human thought. Computation theory, which generalizes logic, includes additional features not represented in logic. Computers as we have defined them are instruments of computation. They are given input (information) specifying both program and data and provide well-defined output an indefinite time later. One of the features that is missing from this kind of machine is the continuous input-output interaction with the world characteristic of a sensory-motor system. An appropriate generalization of the Turing machine would be a robot. As it is conceived and sometimes realized,a robot has both sensory and motor capabilities and an embedded computer. Thus it has more of the features characteristic of a human being. Is this sufficient, or have we missed additional features?

Logic and computation are often contrasted with the concept of creativity. One of the central questions about computers is whether they are able to simulate creativity. In Chapter 3 we will produce a model of creativity that appears to be possible to simulate on a computer. Hidden in this model, however, is a need to use random numbers. This might seem to be a minor problem, since we often use computers to generate random numbers. However, computers do not actually generate randomness,they generate pseudo-random numbers. If we recall that randomness is the same as information, by the discussion in the previous section,a computer cannot generate true randomness.A Turing machine cannot generate a result that has more information than it is given in its initial data. Thus creativity appears to be tied at least in part to randomness, which has often been suggested, and this may be a problem for conventional computers. Conceptually, this problem can be readily resolved by adding to the description of the Turing machine an infinite random tape in addition to the infinite blank tape. This new system appears quite similar to the original TM specification.A reasonable question would ask whether it is really inherently different. The main difference that we can ascertain at this time is that the new system would be capable of generating results with arbitrarily large information content, while the original TM could not. This is not an unreasonable distinction to make between a creative and a logical system. There are still key problems with understanding the practical implications of this distinction.

The subtlety of this discussion increases when we consider that one branch of theoretical computer science is based on the commonly believed assumption that there exist functions that are inherently difficult to invert—they can only be inverted in a time that grows exponentially with the length of the nonblank part of the tape. For all practical purposes, they cannot be inverted, because the estimated lifetime of the universe is insufficient to invert such functions. While their existence is not proven, it has been proven that if they do exist, then such a function can be used to generate a string of characters that, while not random, cannot be distinguished from a random string in less than exponential time. This would suggest that there can be no practical difference between a TM with a random tape,and one without. Thus,the possibility of the existence of noninvertible functions is intimately tied to questions about the relationship between TM, randomness and human thought.

1.9.8 *Using computation and information to describe the real world*

In this section we review the fundamental relevance of the theories of computation and information in the real world. This relevance ultimately arises from the properties of observations and measurements.

In our observations of the world, we find that quantities we measure vary. Indeed, without variation there would be no such thing as an observation. There are variations over time as well as over space. Our intellectual effort is dedicated to classifying or understanding this variation. To concretize the discussion, we consider observations of a variable s which could be as a function of time $s(t)$ or of space $s(x)$. Even though x or t may appear continuous, our observations may often be described as a finite discrete set $\{s_i\}$. One of the central (meta)observations about the variation in value of $\{s_i\}$ is that sometimes the value of the variable s_i can be inferred from, is correlated with, or is not independent from its value or values at some other time or position s_j.

These concepts have to do with the relatedness of s_i to s_j . Why is this important? The reason is that we would like to know the value of s_i without having to observe it.

We can understand this as a problem in prediction—to anticipate events that will occur. We would also like to know what is located at unobserved positions in space;e.g., around the corner. And even if we have observed something, we do not want to have to remember all observations we make. We could argue more fundamentally that knowledge/information is important only if prediction is possible. There would be no reason to remember past observations if they were uncorrelated with anything in the future. If correlations enable prediction,then it is helpful to store information about the past. We want to store as little as possible in order to make the prediction. Why? Because storage is limited, or because accessing the right information requires a search that takes time. If a search takes more time than we have till the event we want to predict, then the information is not useful. As a corollary (from a simplified utilitarian point of view), we would like to retain only information that gives us the best, most rapid prediction, under the most circumstances, for the least storage.

Inference is the process of logic or computation. To be able to infer the state of a variable s_i means that we have a definite formula $f(s_j)$ that will give us the value of s_i with complete certainty from a knowledge of s_j. The theory of computation describes what functions f are possible. If the index i corresponds to a later time than j we say that we can predict its value. In addition to the value of s_j we need to know the function f in order to predict the value of s_i. This relationship need not be from a single value s_j to a single value s_i. We might need to know a collection of values $\{s_j\}$ in order to obtain the value of s_i from $f(\{s_j\})$.

As part of our experience of the world, we have learned that observations at a particular time are more closely related to observations at a previous time than observations at different nearby locations. This has been summarized by the principle of causality. Causality is the ability to determine what happens at one time from what happened at a previous time. This is more explicitly stated as microcausality—what happens at a particular time and place is related to what happened at a previous time in its immediate vicinity. Causality is the principle behind the notion of determinism, which suggests that what occurs is determined by prior conditions. One of the ways that we express the relationship between system observations over time is by conservation laws. Conservation laws are the simplest form of a causal relationship.

Correlation is a looser relationship than inference. The statement that values s_i and s_j are correlated implies that even if we cannot tell exactly what the value s_i is from a knowledge of s_j, we can describe it at least partially. This partial knowledge may also be inherently statistical in the context of an ensemble of values as discussed below. Correlation often describes a condition where the values s_i and s_j are similar. If they are opposite, we might say they are anticorrelated. However, we sometimes use the term "correlated" more generally. In this case, to say that s_i and s_j are correlated would mean that we can construct a function $f(s_j)$ which is close to the value of s_i but not exactly the same. The degree of correlation would tell us how close we expect them to be. While correlations in time appear to be more central than correlations in space, systems with interactions have correlations in both space and time.

Concepts of relatedness are inherently of an ensemble nature. This means that they do not refer to a particular value s_i or a pair of values (s_i, s_j) but rather to a collection of such values or pairs. The ensemble nature of relationships is often more explicit for correlations, but it also applies to inference. This ensemble nature is hidden by functional terminology that describes a relationship between particular values. For example, when we say that the temperature at 1:00 P.M. is correlated with the temperature at 12:00 P.M., we are describing a relationship between two temperature values. Implicitly, we are describing the collection of all pairs of temperatures on different days or at different locations. The set of such pairs are analogs. The concept of inference also generally makes sense only in reference to an ensemble. Let us assume for the moment that we are discussing only a single value s_i. The statement of inference would imply that we can obtain s_i as the value $f(s_j)$. For a single value,the easiest way (requiring the smallest amount of information) to specify $f(s_j)$ would be to specify s_i. We do not gain by using inference for this single case. However, we can gain if we know that, for example,the velocity of an object will remain the same if there are no forces upon it. This describes the velocity $v(t)$ in terms of $v(t')$ of any one object out of an ensemble of objects. We can also gain from inference if the function $f(s_j)$ gives a string of more than one s_i.

The notion of independence is the opposite of inference or correlation. Two values s_i and s_j are independent if there is no way that we can infer the value of one from the other, and if they are not correlated. Randomness is similar to independence. The word "independent" is used when there is no correlation between two observations. The word "random" is stronger, since it means that there is no correlation between an observed value and anything else. A random process,like a sequence of coin tosses,is a sequence where each value is independent of the others. We have seen in Section 1.8 that randomness is intimately related with information. Random processes are unpredictable,therefore it makes no sense for us to try to accumulate information that will help predict it. In this sense, a random process is simple to describe. However, once a random process has occurred,other events may depend upon it. For example, someone who wins a lottery will be significantly affected by an event presumed to be random. Thus we may want to remember the results of the random process after it occurs. In this case we must remember each value. We might ask, Once the random process has occurred, can we summarize it in some way? The answer is that we cannot. Indeed, this property has been used to define randomness.

We can abstract the problem of prediction and description of observations to the problem of data compression. Assume there are a set of observations $\{s_i\}$ for which we would like to obtain the shortest possible description from which we can reconstruct the complete set of observations. If we can infer one value from another, then the set might be compressed by eliminating the inferable values. However, we must make sure that the added information necessary to describe how the inference is to be done is less than the information in the eliminated values. Correlations also enable compression. For example,let us assume that the values are biased ON with a probability $P(1) = .999$ and OFF with a probability $P(-1) = 0.001$. This means that one in a thousand values is OFF and the others are ON. In this case we can remember which ones are OFF rather than keeping a list of all of the values. We would say they are ON except for numbers 3, 2000,2403,5428, etc. This is one way of coding the information. This method of encoding has a problem in that the numbers representing the locations of the OFF values may become large. They will be correlated because the first few digits of successive locations will be the same (...,431236,432112,434329,...). We can further reduce the list if we are willing to do some more processing, by giving the intervals between successive OFF values rather than the absolute numbers of their location.

Ultimately, when we have reached the limits of our ability to infer one observation from another, the rest is information that we need. For example, differential equations are based on the presumption that boundary conditions (initial conditions in time,and boundary conditions in space) are sufficient to predict the behavior of a system. The values of the initial conditions and the boundary conditions are the information we need. This simple model of a system, where information is clearly and simply separated from the problem of computation, is not always applicable.

Let us assume that we have made extensive observations and have separated from these observations a minimal set that then can be used to infer all the rest.A minimal set of information would have the property that no one piece of information in it could be obtained from other pieces of information. Thus,as far as the set itself is concerned, the information appears to be random. Of course we would not be satisfied with any random set; it would have to be this one in particular, because we want to use this information to tell us about all of the actual observations.

One of the difficulties with random numbers is that it is inherently difficult to prove that numbers are random. We may simply not have thought of the right function f that can predict the value of the next number in a sequence from the previous numbers. We could argue that this is one of the reasons that gambling is so attractive to people because of the use of "lucky numbers" that are expected by the individual to have a better-than-random chance of success. Indeed,it is the success of science to have shown that apparently uncorrelated events may be related. For example, the falling of a ball and the motion of the planets. At the same time, science provides a framework in which noncausal correlations, otherwise called superstitions, are rejected.

We have argued that the purpose of knowledge is to succinctly summarize information that can be used for prediction. Thus,in its most abstract form, the problem of deduction or prediction is a problem in data compression. It can thus be argued that science is an exercise in data compression. This is the essence of the principle of Occam's razor and the importance of simplicity and universality in science. The more universal and the more general a law is,and the simpler it is,then the more data compression has been achieved. Often this is considered to relate to how valuable is the contribution of the law to science. Of course, even if the equations are general and simple,if we cannot solve them then they are not particularly useful from a practical point of view. The concept of simplicity has always been poorly defined. While science seeks to discover correlations and simplifications in observations of the universe around us,ultimately the minimum description of a system (i.e.,the universe) is given by the number of independent pieces of information required to describe it.

Our understanding of information and computation enters also into a discussion of our models of systems discussed in previous sections. In many of these models, we assumed the existence of random variables, or random processes. This randomness represents either unknown or complex phenomena. It is important to recognize that this represents an assumption about the nature of correlations between different aspects of the problem that we are modeling. It assumes that the random process is independent of (uncorrelated with) the aspects of the system we are explicitly studying. When we model the random process on a computer by a pseudo-random number generator, we are assuming that the computations in the pseudo-random number generator are also uncorrelated with the system we are studying. These assumptions may or may not be valid, and tests of them are not generally easy to perform.

1.10 Fractals, Scaling and Renormalization

The physics of Newton and the related concepts of calculus, which have dominated scientific thinking for three hundred years,are based upon the understanding that at smaller and smaller scales—both in space and in time—physical systems become simple,smooth and without detail.A more careful articulation of these ideas would note that the fine scale structure of planets, materials and atoms is not without detail. However, for many problems, such detail becomes irrelevant at the larger scale. Since the details are irrelevant, formulating theories in a way that assumes that the detail does not exist yields the same results as a more exact description.

In the treatment of complex systems, including various physical and biological systems,there has been a recognition that the concept of progressive smoothness on finer scales is not always a useful mathematical starting point. This recognition is an important fundamental change in perspective whose consequences are still being explored.

We have already discussed in Section 1.1 the subject of chaos in iterative maps. In chaotic maps, the smoothness of dynamic behavior is violated. It is violated because fine scale details matter. In this section we describe fractals,mathematical models of the spatial structure of systems that have increasing detail on finer scales.Geometric fractals have a self-similar structure, so that the structure on the coarsest scale is repeated on finer length scales. A more general framework in which we can articulate questions about systems with behavior on all scales is that of scaling theory introduced in Section 1.10.3.One of the most powerful analytic tools for studying systems that have scaling properties is the renormalization group. We apply it to the Ising model in Section 1.10.4, and then return full cycle by applying the renormalization group to chaos in Section 1.10.5.A computational technique,the multigrid method, that enables the description of problems on multiple scales is discussed in Section 1.10.6. Finally, we discuss briefly the relevance of these concepts to the study of complex systems in Section 1.10.7.

1.10.1 *Fractals*

Traditional geometry is the study of the properties of spaces or objects that have integral dimensions. This can be generalized to allow effective fractional dimensions of objects, called fractals, that are embedded in an integral dimension space. In recent years the recognition that fractals can play an important role in modeling natural phenomena has fueled a whole area of research investigating the occurrence and properties of fractal objects in physical and biological systems.

Fractals are often defined as geometric objects whose spatial structure is self-similar. This means that by magnifying one part of the object, we find the same structure as of the original object. The object is characteristically formed out of a collection of elements: points, line segments, planar sections or volume elements. These elements exist in a space of the same or higher dimension to the elements themselves. For example, line segments are one-dimensional objects that can be found on a line, plane, volume or higher dimensional space. We might begin to describe a fractal by the objects of which it is formed. However, geometric fractals are often described by a procedure (algorithm) that creates them in an explicitly self-similar manner.

One of the simplest examples of a fractal object is the Cantor set (Fig. 1.10.1). This set is formed by a procedure that starts from a single line segment. We remove the middle third from the segment. There are then two line segments left. We then remove the middle third from both of these segments, leaving four line segments. Continuing iteratively, at the kth iteration there are 2^k segments. The Cantor set, which is the limiting set of points obtained from this process,has no line segments in it. It is self-similar by direct construction,since the left and right third of the original line segment can be expanded by a factor of three to appear as the original set.

An analog of the Cantor set in two dimensions is the Sierpinski gasket (Fig. 1.10.2). It is constructed from an equilateral triangle by removing an internal triangle which is half of the size of the original triangle. This procedure is then iterated for all of the smaller triangles that result. We can see that there are no areas that are left in this shape. It is self-similar, since each of the three corner triangles can be expanded by a factor of two to appear as the original set.

For self-similar objects, we can obtain the effective fractal dimension directly by considering their composition from parts. We do this by analogy with conventional

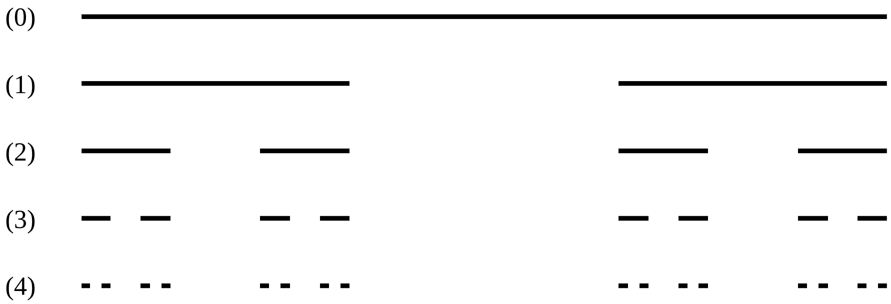


Figure 1.10.1 Illustration of the construction of the Cantor set, one of the best-known fractals. The Cantor set is formed by iteratively removing the middle third from a line segment, then the middle third from the two remaining line segments, and so on. Four iterations of the procedure are shown starting from the complete line segment at the top. ∎

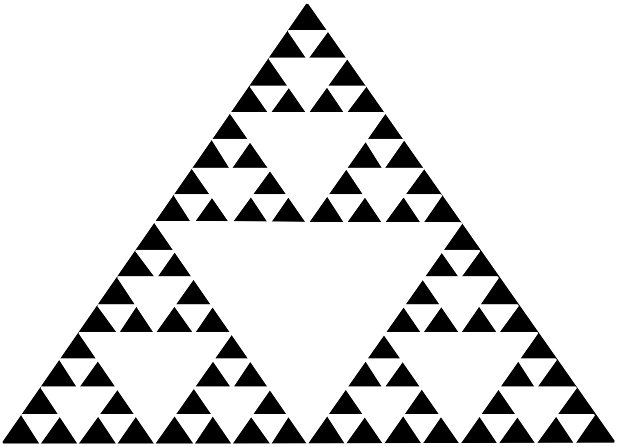


Figure 1.10.2 The Sierpinski gasket is formed in a similar manner to the Cantor set. Starting from an equilateral triangle, a similar triangle one half the size is removed from the middle leaving three triangles at the corners. The procedure is then iteratively applied to the remaining triangles. The figure shows the set that results after four iterations of the procedure. ▮

geometric objects which are also self-similar. For example,a line segment,a square, or a cube can be formed from smaller objects of the same type. In general, for a d-dimensional cube, we can form the cube out of smaller cubes. If the size of the smaller cubes is reduced from that of the large cube by a factor of η, where η is inversely proportional to their diameter, $\eta \quad 1/R$, then the number of smaller cubes necessary to form the original is $N = \eta^d$. Thus we could obtain the dimension as:

$$d = \ln(N) / \ln(\eta) \qquad (1.10.1)$$

For self-similar fractals we can do the same, where N is the number of parts that make up the whole.Each of the parts is assumed to have the same shape, but reduced in size by a factor of η from the original object.

We can generalize the definition of fractal dimension so that we can use it to characterize geometric objects that are not strictly self-similar. There is more than one way to generalize the definition. We will adopt an intuitive definition of fractal dimension which is closely related to Eq. (1.10.1). If the object is embedded in d-dimensions, we cover the object with d-dimensional disks. This is illustrated in Fig. 1.10.3 for a line segment and a rectangle in a two-dimensional space. If we cover the object with two-dimensional disks of a fixed radius, R, using the minimal number of disks possible, the number of these disks changes with the radius of the disks according to the power law:

$$N(R) \quad R^{-d} \qquad (1.10.2)$$

where d is defined as the fractal dimension. We note that the use of disks is only illustrative. We could use squares and the result can be proven to be equivalent.

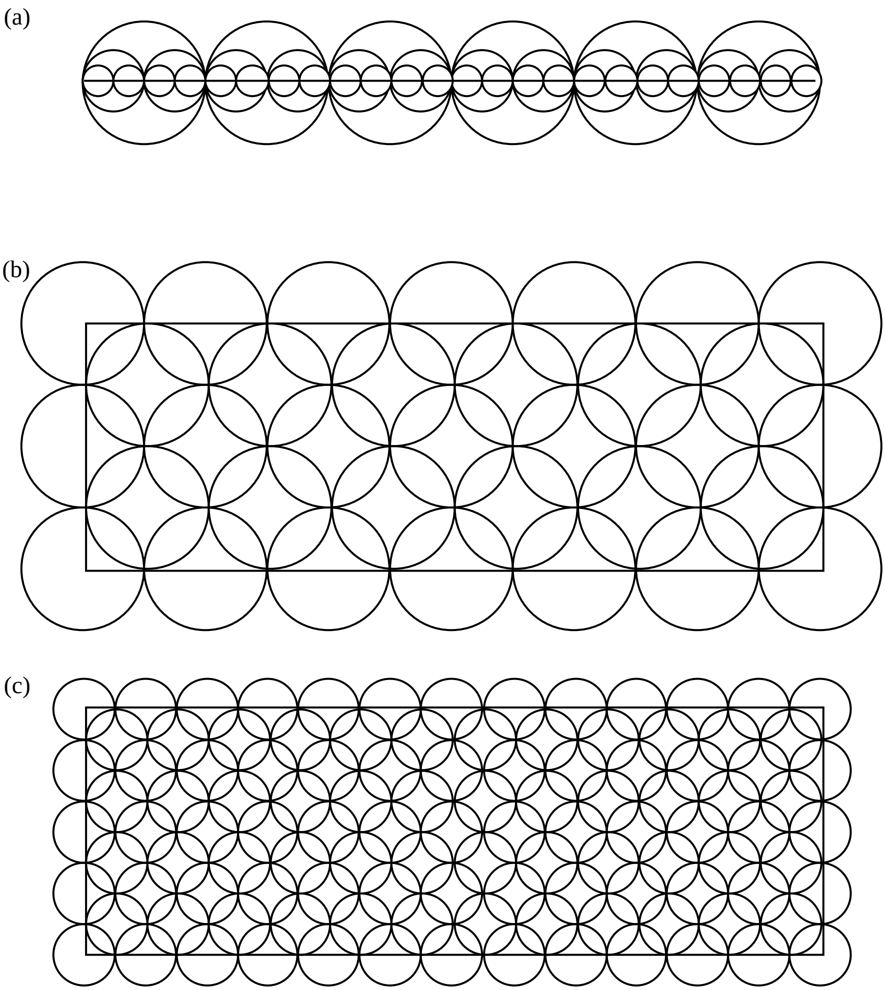


Figure 1.10.3 In order to define the dimension of a fractal object, we consider the problem of covering a set with a minimal number of disks of radius R. (a) shows a line segment with three different coverings superimposed. (b) and (c) show a rectangle with two different coverings respectively. As the size of the disks decreases the number of disks necessary to cover the shape grows as R^{-d}. This behavior becomes exact only in the limit $R \to 0$. The fractal dimension defined in this way is sometimes called the box-counting dimension, because d-dimensional boxes are often used rather than disks. ∎

We can use either Eq. (1.10.1) or Eq. (1.10.2) to calculate the dimension of the Cantor set and the Sierpinski gasket. We illustrate the use of Eq. (1.10.2). For the Cantor set, by construction, 2^k disks (or line segments) of radius $1/3^k$ will cover the set. Thus we can write:

$$N(R / 3^k) = 2^k N(R) \tag{1.10.3}$$

Using Eq. (1.10.2) this is:

$$(R / 3^k)^{-\mathrm{d}} = 2^k R^{-d} \tag{1.10.4}$$

or:

$$3^d = 2 \tag{1.10.5}$$

which is:

$$d = \ln(2) / \ln(3) \quad 0.631 \tag{1.10.6}$$

We would arrive at the same result more directly from Eq. (1.10.1).

For the Sierpinski gasket, we similarly recognize that the set can be covered by three disks of radius 1/2, nine disks of radius 1/4,and more generally 3^k disks of radius $1/2^k$. This gives a dimension of:

$$d = \ln(3) / \ln(2) \quad 1.585 \tag{1.10.7}$$

For these fractals there is a deterministic algorithm that is used to generate them. We can also consider a kind of stochastic fractal generated in a similar way, however, at each level the algorithm involves choices made from a probability distribution. The simplest modification of the sets is to assume that at each level a choice is made with equal probability from several possibilities.For example,in the Cantor set,rather than removing the middle third from each of the line segments, we could choose at random which of the three thirds to remove. Similarly for the Sierpinski gasket, we could choose which of the four triangles to remove at each stage. These would be stochastic fractals,since they are not described by a deterministic self-similarity but by a statistical self-similarity. Nevertheless, they would have the same fractal dimension as the deterministic fractals.

Question 1.10.1 How does the dimension of a fractal,as defined by Eq. (1.10.2), depend on the dimension of the space in which it is embedded?

Solution 1.10.1 The dimension of a fractal is independent of the dimension of the space in which it is embedded. For example, we might start with a d-dimensional space and increase the dimension of the space to $d + 1$ dimensions. To show that Eq. (1.10.2) is not changed, we form a covering of the fractal by $d + 1$ dimensional spheres whose intersection with the d-dimensional space is the same as the covering we used for the analysis in d dimensions. ▮

Question 1.10.2 Prove that the fractal dimension does not change if we use squares or circles for covering an object.

Solution 1.10.2 Assume that we have minimal coverings of a shape using $N_1(\mathbf{R}) = c_1 R^{-d_1}$ squares, and minimal coverings by $N_2(R) = c_2 R^{-d_2}$ circles, with $d_1 \neq d_2$. The squares are characterized using R as the length of their side, while the circles are characterized using R as their radius. If d_1 is less than d_2, then for smaller and smaller R the number of disks becomes arbitrarily smaller than the number of squares. However, we can cover the same shape using squares that circumscribe the disks. The number of these squares is $N'_1(R) = c_1(R / 2)^{-d_1}$. This is impossible, because for small enough R, $N'_1(R)$ will be smaller than $N_1(R)$, which violates the assumption that the latter is a minimal covering. Similarly, if d is greater than d', we use disks circumscribed around the squares to arrive at a contradiction. ▮

Question 1.10.3 Calculate the fractal dimension of the Koch curve given in Fig. 1.10.4.

Solution 1.10.3 The Koch curve is composed out of four Koch curves reduced in size from the original by a factor of 3. Thus, the fractal dimension is $d = \ln(4) / \ln(3) \approx 1.2619$. ▮

Question 1.10.4 Show that the length of the Koch curve is infinite.

Solution 1.10.4 The Koch curve can be constructed by taking out the middle third of a line segment and inserting two segments equivalent to the one that was removed. They are inserted so as to make an equilateral triangle with the removed segment. Thus, at every iteration of the construction procedure, the length of the perimeter is multiplied by 4/3, which means that it diverges to infinity. It can be proven more generally that any fractal of dimension $2 > d > 1$ must have an infinite length and zero area,since these measures of size are for one-dimensional and two-dimensional objects respectively. ▮

Eq. (1.10.2) neglects the jumps in $N(R)$ that arise as we vary the radius R. Since $N(R)$ can only have integral values,as we lower R and add additional disks there are discrete jumps in its value. It is conventional to define the fractal dimension by taking the limit of Eq.(1.10.2) as $R \rightarrow 0$, where this problem disappears. This approach,however, is linked philosophically to the assumption that systems simplify in the limit of small length scales. The assumption here is not that the system becomes smooth and featureless, but rather that the fractal properties will continue to all finer scales and remain ideal. In a physical system,the fractal dimension cannot be taken in this limit. Thus, we should allow the definition to be applied over a limited domain of length scales as is appropriate for the problem. As long as the domain of length scales is large, we can use this definition. We then solve the problem of discrete jumps by treating the leading behavior of the function $N(R)$ over this domain.

The problem of treating distinct dimensions at different length scales is only one of the difficulties that we face in discussing fractal systems. Another problem is inhomogeneity. In the following section we discuss objects that are inherently inhomogeneous but for which an alternate natural definition of dimension can be devised to describe their structure on all scales.

1.10.2 *Trees*

Iterative procedures like those used to make fractals can also be used to make geometric objects called trees. An example of a geometric tree, which bears vague

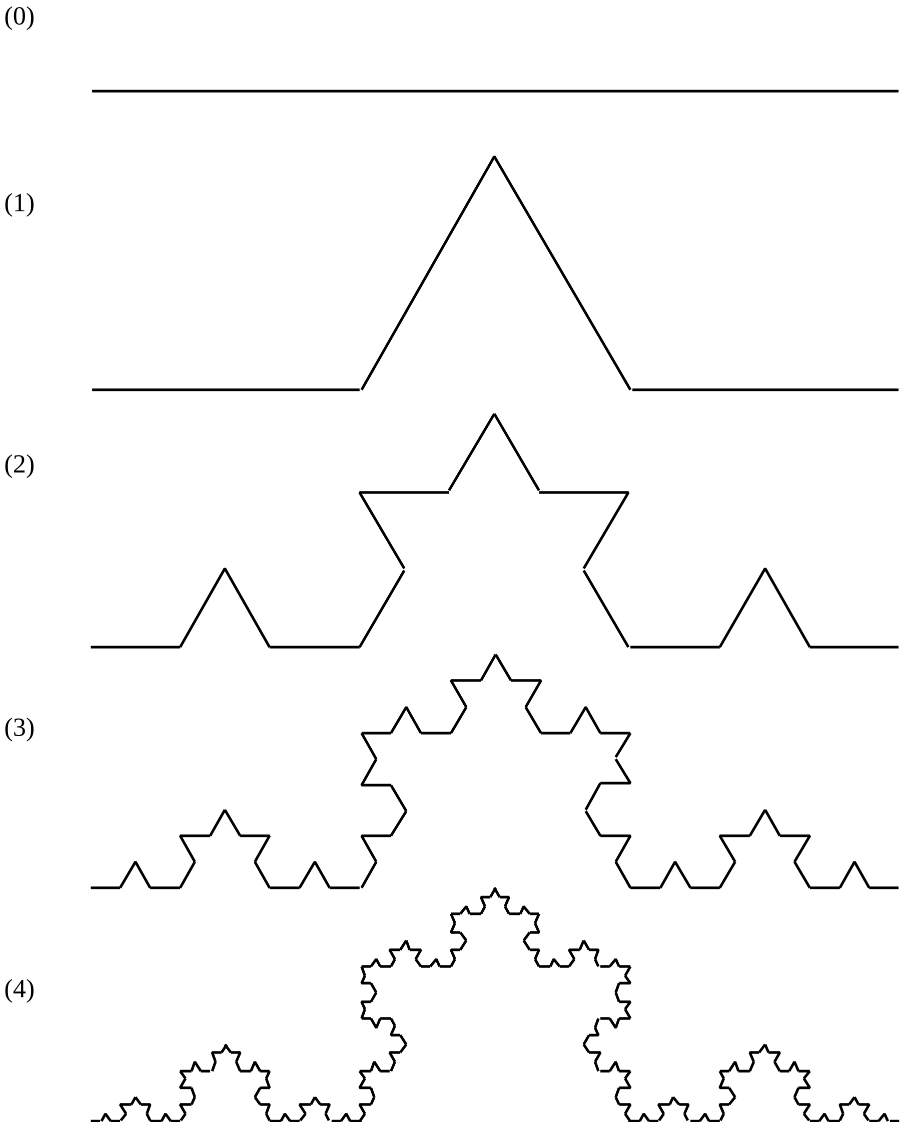


Figure 1.10.4 Illustration of the starting line segment and four successive stages in the formation of the Koch curve. For further discussion see Questions 1.10.3 and 1.10.4. ∎

resemblance to physical trees, is shown in Fig. 1.10.5. The tree is formed by starting with a single object (a line segment), scaling it by a factor of 1/2, duplicating it two times and attaching the parts to the original object at its boundary. This process is then iterated for each of the resulting parts. The iterations create structure on finer and finer scales.

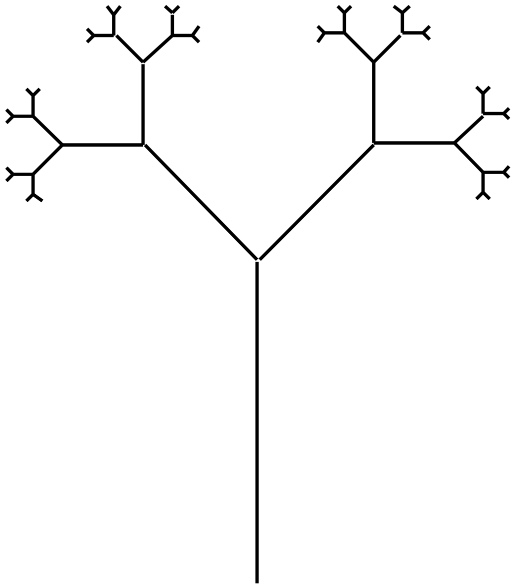


Figure 1.10.5 A geometric tree formed by an iterative algorithm similar to those used in forming fractals. This tree can be formed starting from a single line segment. Two copies of it are then reduced by a factor of 2, rotated by 45° left and right and attached at one end. The procedure is repeated for each of the resulting line segments. Unlike a fractal, a tree is not solely composed out of parts that are self-similar. It is formed out of self-similar parts, along with the original shape — its trunk. ▮

We can generalize the definition of a tree to be a set formed by iteratively adding to an object copies of itself. At iteration *t*, the added objects are reduced in size by a factor η^t and duplicated N^t times, the duplicated versions being rotated and then shifted by vectors whose lengths converge to zero as a function of *t*. A tree is different from a fractal because the smaller versions of the original object, are not contained within the original object.

The fractal dimension of trees is not as straightforward as it is for self-similar fractals. The effective fractal dimension can be calculated; however, it gives results that are not intuitively related to the tree structure. We can see why this is a problem in Fig. 1.10.6. The dimension of the region of the tree which is above the size *R* is that of the embedded entity (line segments), while the fractal dimension of the region which is less than the size *R* is determined by the spatial structure of the tree. Because of the changing value of *R* in the scaling relation, an intermediate value for the fractal dimension would typically be found by a direct calculation (Question 1.10.5).

It is reasonable to avoid this problem by classifying trees in a different category than fractals. We can define the tree dimension by considering the self-similarity of the tree structure using the same formula as Eq. (1.10.1), but now applying the definition to the number *N* and scaling η of the displaced parts of the generating structure, rather than the embedded parts as in the fractal. In Section 1.10.7 we will encounter a treelike structure; however, it will be more useful to describe it rather than to give a dimension that might characterize it.

Question 1.10.5 A simple version of a tree can be constructed as a set of points $\{1/k\}$ where *k* takes all positive integer values. The tree dimension

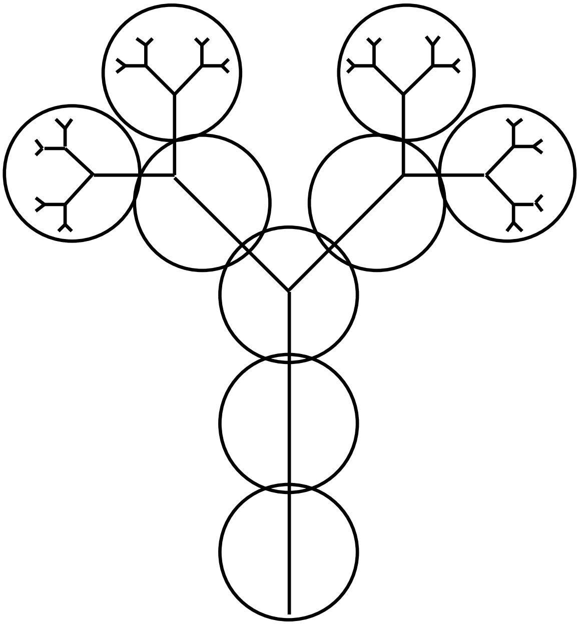


Figure 1.10.6 Illustration of the covering of a geometric tree by disks. The covering shows that the larger-scale structures of the tree (the trunk and first branches in this case) have an effective dimension given by the dimension of their components. The smaller scale structures have a dimension that is determined by the algorithm used to make the tree. This inhomogeneity implies that the fractal dimension is not always the natural way to describe the tree. ▮

of this set is zero because it can be formed from a point which is duplicated and then displaced by progressively smaller vectors. Calculate the fractal dimension of this set.

Solution 1.10.5 We construct a covering of scale *R* from line segments of this length. The covering that we construct will be formed out of two parts. One part is constructed from segments placed side by side. This part starts from zero and covers infinitely many points of the set. The other part is constructed from segments that are placed on individual points. The crossing point between the two sets can be calculated as the value of *k* where the difference between successive points is *R*. For *k* below this value,it is not possible to include more than one point in one line segment. For *k* above this value, there are two or more points per line segment. The critical value of *k* is found by setting:

$$\frac{1}{k_c} - \frac{1}{k_c + 1} = \frac{1}{k_c(k_c + 1)} \approx \frac{1}{k_c^2} = R \tag{1.10.8}$$

or $k_c = R^{-1/2}$. This means that the number of segments needed to cover individual points is given by this value. Also, the number of segments that are placed side by side must be enough to go up to this point, which has the value $1/k_c$. This number of segments is given by

$$\frac{1/k_c}{R} = R^{-1/2} \approx k_c \tag{1.10.9}$$

Thus we must cover the line segment up to the point $R^{1/2}$ with $R^{-1/2}$ line segments, and use an additional $R^{-1/2}$ line segments to cover the rest of the points. This gives a total number of line segments in a covering of $2R^{-1/2}$. The fractal dimension is thus $d = 1/2$.

We could have used fewer line segments in the covering by covering pairs of points and triples of points rather than covering the whole line segment below $1/k_c$. However, each partial covering of the set that is concerned with pairs, triples and so on consists of a number of segments that grows as $R^{-1/2}$. Thus our conclusion remains unchanged by this correction. ■

Trees illustrate only one example of how system properties may exist on many scales, but are not readily described as fractals in the conventional sense. In order to generalize our concepts to enable the discussion of such properties, we will introduce the concept of scaling.

1.10.3 *Scaling*

Geometric fractals suggest that systems may have a self-similar structure on all length scales. This is in contrast with the more typical approach of science, where there is a specific scale at which a phenomenon appears. We can think about the problem of describing the behavior of a system on multiple length scales in an abstract manner. A phenomenon (e.g., a measurable quantity) may be described by some function of scale, $f(x)$. Here x represents the characteristic scale rather than the position. When there is a well-defined length scale at which a particular effect occurs, for longer length scales the function would typically decay exponentially:

$$f(x) \quad e^{-x/\lambda} \tag{1.10.10}$$

This functional dependence implies that the characteristic scale at which this property disappears is given by λ.

In order for a system property to be relevant over a large range of length scales,it must vary more gradually than exponentially. In such cases, typically, the leading behavior is a power law:

$$f(x) \quad x^{\alpha} \tag{1.10.11}$$

A function that follows such power-law behavior can also be characterized by the scaling rule:

$$f(ax) = a^{\alpha} f(x) \tag{1.10.12}$$

This means that if we characterize the system on one scale, then on a scale that is larger by the factor a it has a similar appearance, but scaled by the factor a^{α}. α is called the scaling exponent. In contrast to the behavior of an exponential, for a power law there is no particular length at which the property disappears. Thus, it may extend over a wide range of length scales. When the scaling exponent is not an integer, the function $f(x)$ is nonanalytic. Non-analyticity is often indicative of a property that cannot be treated by assuming that it becomes smooth on small or large scales. However, fractional scaling exponents are not necessary in order for power-law scaling to be applicable.

Even when a system property follows power-law scaling, the same behavior cannot continue over arbitrarily many length scales. The disappearance of a certain power law may occur because of the appearance of a new behavior on a longer scale. This change is characterized by a crossover in the scaling properties of $f(x)$. An example of crossover occurs when we have a quantity whose scaling behavior is

$$f(x) \sim A_1 x^{\alpha_1} + A_2 x^{\alpha_2} \tag{1.10.13}$$

If $A_1 > A_2$ and $\alpha_1 < \alpha_2$ then the first term will dominate at smaller length scales, and the second at larger length scales. Alternatively, the power-law behavior may eventually succumb to exponential decay at some length scale.

There are three related approaches to applying the concept of scaling in model or physical systems. The first approach is to consider the scale x to be the physical size of the system, or the amount of matter it contains. The quantity $f(x)$ is then a property of the system measured as the size of the system changes. The second approach is to keep the system the same, but vary the scale of our observation. We assume that our ability to observe the system has a limited degree of discernment of fine details—a finest scale of observation. Finer details are to be averaged over or disregarded. By moving toward or away from the system, we change the physical scale at which our observation can no longer discern details. x then represents the smallest scale at which we can observe variation in the system structure. Finally, in the third approach we consider the relationship between a property measured at one location in the system and the same property measured at another location separated by the distance x. The function $f(x)$ is a correlation of the system measurements as a function of the distance between regions that are being considered.

Examples of quantities that follow scaling relations as a function of system size are the extensive properties of thermodynamic systems (Section 1.3) such as the energy, entropy, free energy, volume, number of particles and magnetization:

$$U(ax) = a^d U(x) \tag{1.10.14}$$

These properties measure quantities of the whole system as a function of system size. All have the same scaling exponent—the dimension of space. Intrinsic thermodynamic quantities are independent of system size and therefore also follow a scaling behavior where the scaling exponent is zero.

Another example of scaling can be found in the random walk (Section 1.2). We can generalize the discussion in Section 1.2 to allow a walk in d dimensions by choosing steps which are ±1 in each dimension independently. A random walk of N steps in three dimensions can be thought of as a simple model of a molecule formed as a chain of molecular units—a polymer. If we measure the average distance between the ends of the chain as a function of the number of steps $R(N)$, we have the scaling relation:

$$R(aN) = a^{1/2} R(N) \tag{1.10.15}$$

This scaling of distance traveled in a random walk with the number of steps taken is independent of dimension. We will consider random walks and other models of polymers in Chapter 5.

Often our interest is in knowing how different parts of the system affect each other. Direct interactions do not always reflect the degree of influence. In complex systems, in which many elements are interacting with each other, there are indirect means of interacting that transfer influence between one part of a system and another. The simplest example is the Ising model, where even short-range interactions can lead to longer-range correlations in the magnetization. The correlation function introduced in Section 1.6.5 measures the correlations between different locations. These correlations show the degree to which the interactions couple the behavior of different parts of the system. Correlations of behavior occur in both space and time. As we mentioned in Section 1.3.4, near a second-order phase transition, there are correlations between different places and times on every length and time scale, because they follow a power-law behavior. This example will be discussed in greater detail in the following section.

Our discussion of scaling also finds application in the theory of computation (Section 1.9) and the practical aspects of simulation (Section 1.7). In addition to the question of computability discussed in Section 1.9, we can also ask how hard it is to compute something. Such questions are generally formulated by describing a class of problems that can be ordered by a parameter N that describes the size of the problem. The objective of the theory of computational complexity is to determine how the number of operations necessary to solve a problem grows with N. A scaling analysis can also be used to compare different algorithms that may solve the same problem. We are often primarily concerned with the scaling behavior (exponential, power law and the value of the scaling exponent) rather than the coefficients of the scaling behavior, because in the comparison of the difficulty of solving different problems or different methodologies this is often, though not always, the most important issue.

1.10.4 *Renormalization group*

General method The renormalization group is a formalism for studying the scaling properties of a system. It starts by assuming a set of equations that describe the behavior of a system. We then change the length scale at which we are describing the system. In effect, we assume that we have a finite ability to see details. By moving away from a system, we lose some of the detail. At the new scale we assume that the same set of equations can be applied, but possibly with different coefficients. The objective is to relate the set of equations at one scale to the set of equations at the other scale. Once this is achieved, the scale-dependent properties of the system can be inferred.

Applications of the renormalization group method have been largely to the study of equilibrium systems,particularly near second-order phase transitions where mean field approaches break down (Section 1.6). The premise of the renormalization group is that exactly at a second-order phase transition,the equations describing the system are independent of scale. In recent years, dynamic renormalization theory has been developed to describe systems that evolve in time. In this section we will describe the more conventional renormalization group for thermodynamic systems.

We illustrate the concepts of renormalization using the Ising model. The Ising model,discussed in Section 1.6, describes the interactions of spins on a lattice. It is a first model of any system that exhibits simple cooperative behavior, such as a magnet.

In order to appreciate the concept of renormalization, it is useful to recognize that the Ising model is not a true microscopic theory of the behavior of a magnet. It might seem that there is a well-defined way to identify an individual spin with a single electron at the atomic level. However, this is far from apparent when equations that describe quantum mechanics at the atomic level are considered. Since the relationship between the microscopic system and the spin model is not manifest, it is clear that our description of the magnet using the Ising model relies upon the macroscopic properties of the model rather than its microscopic nature. Statistical mechanics does not generally attempt to derive macroscopic properties directly from microscopic reality. Instead, it attempts to describe the macroscopic phenomena from simple models. We might not give up hope of identifying a specific microscopic relationship between a particular material and the Ising model,however, the use of the model does not rely upon this identification.

Essential to this approach is that many of the details of the atomic regime are somehow irrelevant at longer length scales. We will return later to discuss the relevance or irrelevance of microscopic details. However, our first question is: What is a single spin variable? A spin variable represents the effective magnetic behavior of a region of the material. There is no particular reason that we should imagine an individual spin variable as representing a small or a large region of the material. Sometimes it might be possible to consider the whole magnet as a single spin in an external field. Identifying the spin with a region of the material of a particular size is an assignment of the length scale at which the model is being applied.

What is the difference between an Ising model describing the system at one length scale and the Ising model describing it on another? The essential point is that the interactions between spins will be different depending on the length scale at which we choose to model the system. The renormalization group takes this discussion one step further by explicitly relating the models at different scales.

In Fig. 1.10.7 we illustrate an Ising model in two dimensions. There is a second Ising model that is used to describe this same system but on a length scale that is twice as big. The first Ising model is described by the energy function (Hamiltonian):

$$E[\{s_i\}] = -c\sum_i 1 - h\sum_i s_i - J\sum_{<ij>} s_i s_j \qquad (1.10.16)$$

For convenience, in what follows we have included a constant energy term $-cN = -c\Sigma 1$. This term does not affect the behavior of the system,however, its variation from scale to scale should be included. The second Ising model is described by the Hamiltonian

$$E'[\{s'_i\}] = -c'\sum_i 1 - h'\sum_i s'_i - J'\sum_{<ij>} s'_i s'_j \qquad (1.10.17)$$

where both the variables and the coefficients have primes. While the first model has N spins, the second model has N' spins. Our objective is to relate these two models. The general process is called renormalization. When we go from the fine scale to the coarse scale by eliminating spins, the process is called decimation.

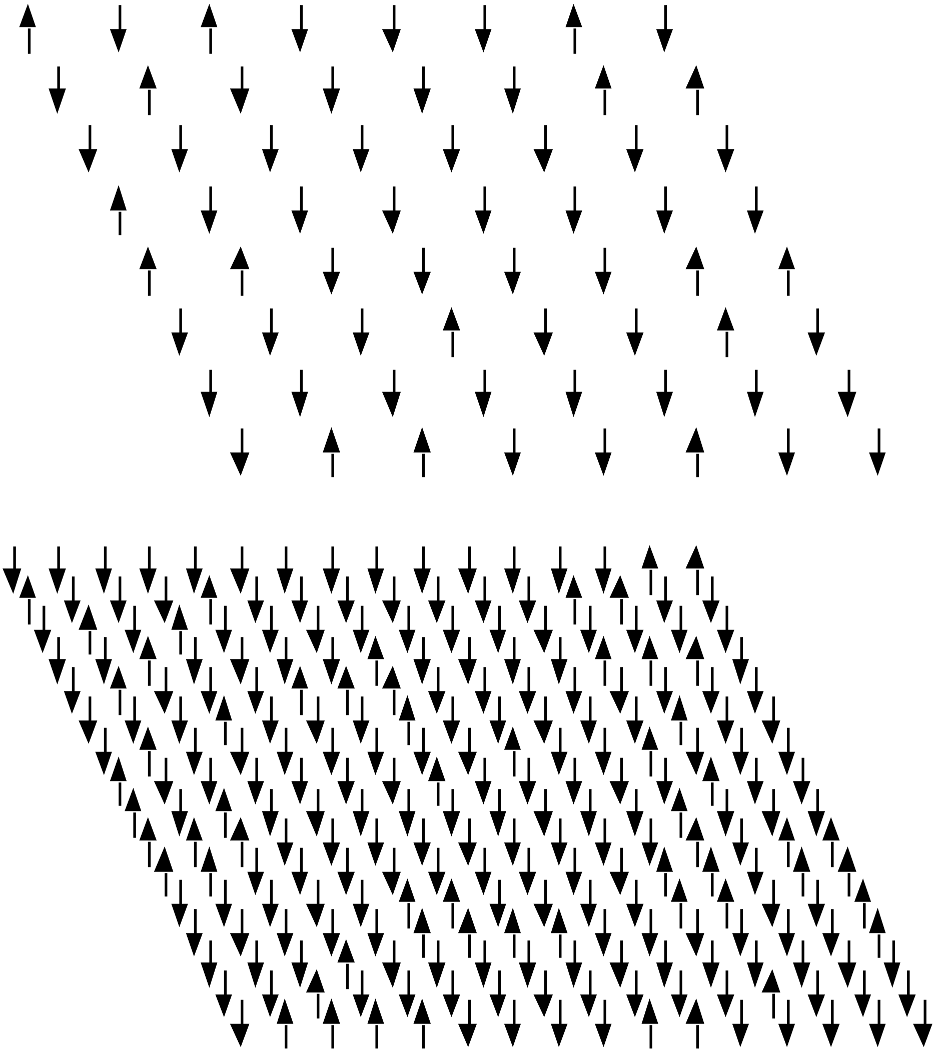


Figure 1.10.7 Schematic illustration of two Ising models in two dimensions. The spins are indicated by arrows that can be UP or DOWN. These Ising models illustrate the modeling of a system with different levels of detail. In the upper model there are one-fourth as many spins as in the lower model. In a renormalization group treatment the parameters of the lower model are related to the parameters of the upper model so that the same system can be described by both. Each of the spins in the upper model, in effect, represents four spins in the lower model. The interactions between adjacent spins in the upper model represent the net effect of the interactions between groups of four spins in the lower model. ▮

There are a variety of methods used for relating models at different scales. Each of them provides a distinct conceptual and practical approach. While in principle they should provide the same answer, they are typically approximated at some stage of the calculation and therefore the answers need not be the same. All the approaches we describe rely upon the partition function to enable direct connection from the microscopic statistical treatment to the macroscopic thermodynamic quantities. For a particular system, the partition function can be written so that it has the same value, independent of which representation is used:

$$Z = \sum_{\{s_i\}} e^{-\beta E[\{s_i\}]} = \sum_{\{s_i'\}} e^{-\beta E'[\{s_i'\}]} \tag{1.10.18}$$

It is conventional and convenient when performing renormalization transformations to set $\beta = 1/kT = 1$. Since β multiplies each of the parameters of the energy function, it is a redundant parameter. It can be reinserted at the end of the calculations.

The different approaches to renormalization are useful for various models that can be studied. We will describe three of them in the following paragraphs because of the importance of the different conceptual treatments. The three approaches are (1) summing over values of a subset of the spins, (2) averaging over a local combination of the spins, and (3) summing over the short wavelength degrees of freedom in a Fourier space representation.

1. Summing over values of a subset of the spins. In the first approach we consider the spins on the larger scale to be a subset of the spins on the finer scale. To find the energy of interaction between the spins on the larger scale we need to eliminate (decimate) some of the spins and replace them by new interactions between the spins that are left. Specifically, we identify the larger scale spins as corresponding to a subset $\{s_i\}_A$ of the smaller scale spins. The rest of the spins $\{s_i\}_B$ must be eliminated from the fine scale model to obtain the coarse scale model. We can implement this directly by using the partition function:

$$e^{-E'[\{s_i'\}]} = \sum_{\{s_i\}_B} e^{-E[\{s_i'\}_A,\{s_i\}_B]} = \sum_{\{s_i\}} e^{-E[\{s_i\}]} \prod_{i\ A} \delta_{s_i',s_i} \tag{1.10.19}$$

   In this equation we have identified the spins on the larger scale as a subset of the finer scale spins and have summed over the finer scale spins to obtain the effective energy for the larger scale spins.

2. Averaging over a local combination of the spins. We need not identify a particular spin of the finer scale with a particular spin of the coarser scale. We can choose to identify some function of the finer scale spins with the coarse scale spin. For example, we can identify the majority rule of a certain number of fine scale spins with the coarse scale spins:

$$e^{-E[\{s_i'\}]} = \sum_{\{s_i\}} \prod_{i\ A} \delta_{s_i',\mathrm{sign}(\sum s_i)} e^{-E[\{s_i\}]} \tag{1.10.20}$$

This is easier to think about when an odd number of spins are being renormalized to become a single spin. Note that this is quite similar to the concept of defining a collective coordinate that we used in Section 1.4 in discussing the two-state system. The difference here is that we are defining a collective coordinate out of only a few original coordinates, so that the reduction in the number of degrees of freedom is comparatively small. Note also that by convention we continue to use the term "energy," rather than "free energy," for the collective coordinates.

3. Summing over the short wavelength degrees of freedom in a Fourier space representation. Rather than performing the elimination of spins directly, we may recognize that our procedure is having the effect of removing the fine scale variation in the problem. It is natural then to consider a Fourier space representation where we can remove the rapid changes in the spin values by eliminating the higher Fourier components. To do this we need to represent the energy function in terms of the Fourier transform of the spin variables:

$$s_k = \sum_i e^{ikx_i} s_i \tag{1.10.21}$$

Writing the Hamiltonian in terms of the Fourier transformed variables, we then sum over the values of the high frequency terms:

$$e^{-E[\{s_k\}]} = \sum_{\{s_k\}_{k<k_0}} e^{-E[\{s_k\}]} \tag{1.10.22}$$

The remaining coordinates s_k have $k > k_0$.

All of the approaches described above typically require some approximation in order to perform the analysis. In general there is a conservation of effort in that the same difficulties tend to arise in each approach, but with different manifestation. Part of the reason for the difficulties is that the Hamiltonian we use for the Ising model is not really complete. This means that there can be other parameters that should be included to describe the behavior of the system. We will see this by direct application in the following examples.

Ising model in one dimension We illustrate the basic concepts by applying the renormalization group to a one-dimensional Ising model where the procedure can be done exactly. It is convenient to use the first approach (number 1 above) of identifying a subset of the fine scale spins with the larger scale model. We start with the case where there is an interaction between neighboring spins, but no magnetic field:

$$E[\{s_i\}] = -c\sum_i 1 - J\sum_{<ij>} s_i s_j \tag{1.10.23}$$

We sum the partition function over the odd spins to obtain

$$Z = \sum_{\{s_i\}_{\text{even}}} \sum_{\{s_i\}_{\text{odd}}} e^{c\sum\limits_i 1 + J\sum\limits_i s_i s_{i+1}} = \sum_{\{s_i\}_{\text{even}}} \prod_{i\,\text{even}} 2\cosh(J(s_i + s_{i+2}))e^{2c} \tag{1.10.24}$$

We equate this to the energy for the even spins by themselves, but with primed quantities:

$$Z = \sum_{\{s_i\}_{\text{even}}} e^{\sum_i c' + J' \sum_i s_i s_{i+2}} = \sum_{\{s_i\}_{\text{even}}} \prod_{i\,\text{even}} 2\cosh(J(s_i + s_{i+2}))e^{2c} \qquad (1.10.25)$$

This gives:

$$e^{c' + J' s_i s_{i+2}} = 2\cosh(J(s_i + s_{i+2}))e^{2c} \qquad (1.10.26)$$

or

$$c' + J' s_i s_{i+2} = \ln(2\cosh(J(s_i + s_{i+2}))) + 2c \qquad (1.10.27)$$

Inserting the two distinct combinations of values of s_i and s_{i+2} ($s_i = s_{i+2}$ and $s_i = -s_{i+2}$), we have:

$$\begin{aligned} c' + J' &= \ln(2\cosh(2J)) + 2c \\ c' - J' &= \ln(2\cosh(0)) + 2c = \ln(2) + 2c \end{aligned} \qquad (1.10.28)$$

Solving these equations gives the primed quantities for the larger scale model as:

$$\begin{aligned} J' &= (1/2)\ln(\cosh(2,J)) \\ c' &= 2c + (1/2)\ln(4\cosh(2J)) \end{aligned} \qquad (1.10.29)$$

This is the renormalization group relationship that we have been looking for. It relates the values of the parameters in the two different energy functions at the different scales.

While it may not be obvious by inspection, this iterative map always causes J to decrease. We can see this more easily if we transform the relationship of J to J' to the equivalent form:

$$\tanh(J') = \tanh(J)^2 \qquad (1.10.30)$$

This means that on longer and longer scales the effective interaction between neighboring spins becomes smaller and smaller. Eventually the system on long scales behaves as a string of decoupled spins.

The analysis of the one-dimensional Ising model can be extended to include a magnetic field. The decimation step becomes:

$$Z = \sum_{\{s_i\}_{\text{even}}} \sum_{\{s_i\}_{\text{odd}}} e^{c\sum_i 1 + h\sum_i s_i + J\sum_i s_i s_{i+1}} = \sum_{\{s_i\}_{\text{even}}} \prod_{i\,\text{odd}} 2\cosh(h + J(s_i + s_{i+2}))e^{2c} \qquad (1.10.31)$$

We equate this to the coarse scale partition function:

$$Z = \sum_{\{s_i\}_{\text{odd}}} e^{c' + h'\sum_i s_i' + J'\sum_i s_i' s_{i+1}'} = \sum_{\{s_i\}_{\text{odd}}} \prod_{i\,\text{odd}} 2\cosh(h + J(s_i + s_{i+2}))e^{2c} \qquad (1.10.32)$$

which requires that:

$$\begin{aligned} c' + h' + J' &= h + \ln(2\cosh(h + 2J)) + 2c \\ c' - J' &= \ln(2\cosh(h)) + 2c \qquad (1.10.33) \\ c' - h' + J' &= -h + \ln(2\cosh(h - 2J)) + 2c \end{aligned}$$

We solve these equations to obtain:

$$\begin{aligned} c' &= 2c + (1/4)\ln(16\cosh(h + 2J)\cosh(h - 2J)\cosh(h)^2) \\ J' &= (1/4)\ln(\cosh(h + 2J)\cosh(h - 2J)/\cosh(h)^2) \qquad (1.10.34) \\ h' &= h + (1/2)\ln(\cosh(h + 2J)/\cosh(h - 2J)) \end{aligned}$$

which is the desired renormalization group transformation. The renormalization transformation is an iterative map in the parameter space (c, h, J).

We can show what happens in this iterative map using a plot of changes in the values of J and h at a particular value of these parameters. Such a diagram of flows in the parameter space is illustrated in Fig. 1.10.8. We can see from the figure or from Eq. (1.10.34) that there is a line of fixed points of the iterative map at $J = 0$ with arbitrary

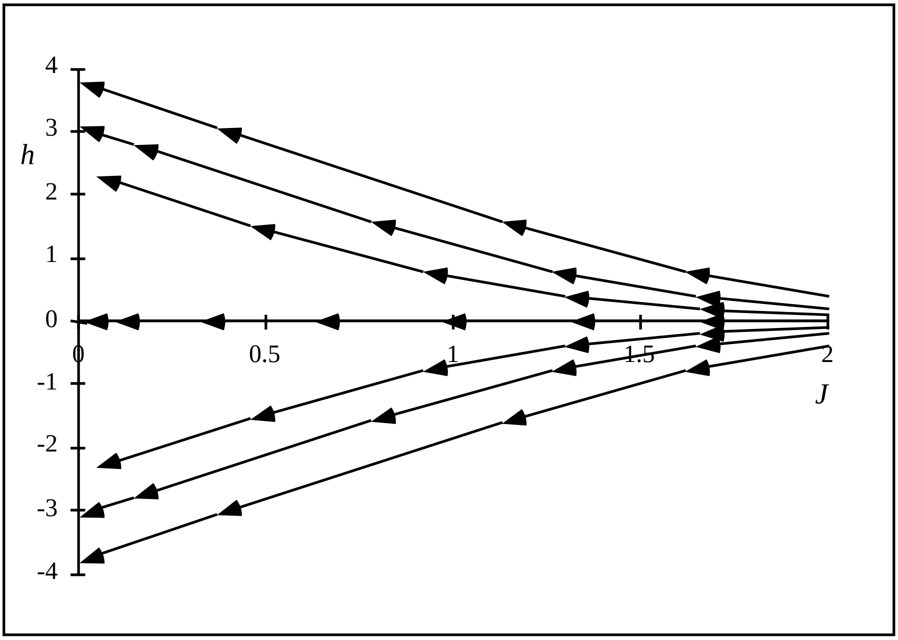


Figure 1.10.8 The renormalization transformation for the one-dimensional Ising model is illustrated as an iterative flow diagram in the two-dimensional (h,J) parameter space. Each of the arrows represents the effect of decimating half of the spins. We see that after a few iterations the value of J becomes very small. This indicates that the spins become decoupled from each other on a larger scale. The absence of any interaction on this scale means that there is no phase transition in the one-dimensional Ising model. ∎

value of *h*. This simply means that the spins are decoupled. For $J = 0$ on any scale,the behavior of the spins is determined by the value of the external field.

The line of fixed points at $J = 0$ is a stable (attracting) set of fixed points. The flow lines of the iterative map take us to these fixed points on the attractor line. In addition, there is an unstable fixed point at $J = \infty$. This would correspond to a strongly coupled line of spins, but since this fixed point is unstable it does not describe the large scale behavior of the model. For any finite value of *J*, changing the scale rapidly causes the value of *J* to become small. This means that the large scale behavior is always that of a system with $J = 0$.

Ising model in two dimensions In the one-dimensional case treated in the previous section, the renormalization group works perfectly and is also, from the point of view of studying phase transitions,uninteresting. We will now look at two dimensions, where the renormalization group must be approximated and where there is also a phase transition.

We can simplify our task in two dimensions by eliminating half of the spins (Fig. 1.10.9) instead of three out of four spins as illustrated previously in Fig. 1.10.7. Eliminating half of the spins causes the square cell to be rotated by 45°,but this should not cause any problems. Labeling the spins as in Fig. 1.10.9 we write the decimation step for a Hamiltonian with $h = 0$:

$$\begin{aligned} Z &= \sum_{\{s_i\}_A} \sum_{\{s_i\}_B} e^{c\sum_i 1 + J\sum_i s_0(s_1+s_2+s_3+s_4)} \\ &= \sum_{\{s_i\}_A} \prod_{i\ B} 2\cosh(J(s_1+s_2+s_3+s_4))e^{c} \\ &= \sum_{\{s_i\}_A} \prod_{i\ B} e^{c'+(J'/2)(s_1s_2+s_2s_3+s_3s_4+s_4s_1)} \end{aligned} \qquad (1.10.35)$$

In the last expression we take into consideration that each bond of the form s_1s_2 appears in two squares and each spin appears in four squares.

In order to solve Eq. (1.10.35) for the values of $c'$and J' we must insert all possible values of the spins (s_1,s_2,s_3,s_4). However, this leads to a serious problem. There are four distinct equations that arise from the different values of the spins. This is reduced from $2^4 = 8$ because, by symmetry, inverting all of the spins gives the same answer. The problem is that while there are four equations, there are only two unknowns to solve for, c' and J'. The problem can be illustrated by recognizing that there are two distinct ways to have two spins UP and two spins DOWN. One way is to have the spins that are the same be adjacent to each other, and the other way is to have them be opposite each other across a diagonal. The two ways give the same result for the value of $(s_1 + s_2 + s_3 + s_4)$ but different results for $(s_1s_2 + s_2s_3 + s_3s_4 + s_4s_1)$.

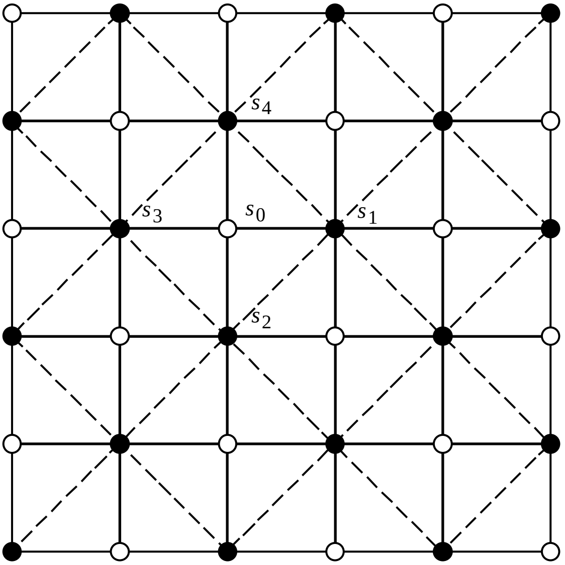


Figure 1.10.9 In a renormalization treatment of the two-dimensional Ising model it is possible to decimate one out of two spins as illustrated in this figure. The black dots represent spins that remain in the larger-scale model, and the white dots represent spins that are decimated. The nearest-neighbor interactions in the larger-scale model are shown by dashed lines. As discussed in the text, the process of decimation introduces new interactions between spins across the diagonal, and four spin interactions between spins around a square. ▮

In order to solve this problem, we must introduce additional parameters which correspond to other interactions in the Hamiltonian. To be explicit, we would make a table of symmetry-related combinations of the four spins as follows:

(s_1,s_2,s_3,s_4)	(1,1,1,1)	(1,1,1,−1)	(1,1,−1,−1)	(1,−1,1,−1)
1	1	1	1	1
$(s_1+s_2+s_3+s_4)$	4	2	0	0
$(s_1s_2+s_2s_3+s_3s_4+s_4s_1)$	4	0	0	−4
$(s_1s_3+s_2s_4)$	2	0	−2	2
$s_1s_2s_3s_4$	1	−1	1	1

(1.10.36)

In order to make use of these to resolve the problems with Eq.(1.10.35), we must introduce new interactions in the Hamiltonian and new parameters that multiply them. This leads to second-neighbor interactions (across a cell diagonal),and four spin interactions around a square:

$$E[\{s_i\}] = -c'\sum_i 1 - J'\sum_{<ij>} s_i s_j - K'\sum_{<<ij>>} s_i s_j - L'\sum_{<ijkl>} s_i s_j s_k s_l \qquad (1.10.37)$$

where the notation $<<ij>>$ indicates second-neighbor spins across a square diagonal, and $<ijkl>$ indicates spins around a square. This might seem to solve our problem. However, we started out from a Hamiltonian with only two parameters,and now we are switching to a Hamiltonian with four parameters. To be self-consistent, we should start from the same set of parameters we end up with. When we start with the additional parameters K and L this will,however, lead to still further terms that should be included.

Relevant and irrelevant parameters In general,as we eliminate spins by renormalization, we introduce interactions between spins that might not have been included in the original model. We will have interactions between second or third neighbors or between more than two spins at a time. In principle, by using a complete set of parameters that describe the system we can perform the renormalization transformation and obtain the flows in the parameter space. These flows tell us about the scale-dependent properties of the system.

We can characterize the flows by focusing on the fixed points of the iterative map. These fixed points may be stable or unstable. When a fixed point is unstable, renormalization takes us away from the fixed point so that on a larger scale the properties of the system are found to be different from the values at the unstable fixed point. Significantly, it is the unstable fixed points that represent the second-order phase transitions. This is because deviating from the fixed point in one direction causes the parameters to flow in one direction, while deviating from the fixed point in another direction causes the parameters to flow in a different direction. Thus,the macroscopic properties of the system depend on the direction microscopic parameters deviate from the fixed point—a succinct characterization of the nature of a phase transition.

Using this characterization of fixed points, we can now distinguish between different types of parameters in the model. This includes all of the additional parameters that might be introduced in order to achieve a self-consistent renormalization transformation. There are two major categories of parameters: relevant or irrelevant. Starting near a particular fixed point, changes in a relevant parameter grow under renormalization. Changes in an irrelevant parameter shrink. Because renormalization indicates the values of system parameters on a larger scale,this tells us which microscopic parameters are important to the macroscopic scale. When observed on the macroscopic scale, relevant parameters change at the phase transition, while irrelevant parameters do not. A relevant parameter should be included in the Hamiltonian because its value affects the macroscopic behavior. An irrelevant parameter may often be included in the model in a more approximate way. Marginal parameters are the borderline cases that neither grow nor shrink at the fixed point.

Even when we are not solely interested in the behavior of a system at a phase transition, but rather are concerned with its macroscopic properties in general,the definition of "relevant" and "irrelevant" continues to make sense. If we start from a particular microscopic description of the system, we can ask which parameters are relevant for the macroscopic behavior. The relevant parameters are the ones that can affect the macroscopic behavior of the system. Thus, a change in a relevant microscopic parameter changes the macroscopic behavior. In terms of renormalization, changes in relevant parameters do not disappear as a result of renormalization.

We see that the use of any model, such as the Ising model, to model a physical system assumes that all of the parameters that are essential in describing the system have been included. When this is true, the results are universal in the sense that all microscopic Hamiltonians will give rise to the same behavior. Additional terms in the Hamiltonian cannot affect the macroscopic behavior. We know that the microscopic behavior of the physical system is not really described by the Ising model or any other simple model. Thus, in creating models we always rely upon the concept, if not the process, of renormalization to make many of the microscopic details disappear, enabling our simple models to describe the macroscopic behavior of the physical system.

In the Ising model, in addition to longer range and multiple spin interactions, there is another set of parameters that may be relevant. These parameters are related to the use of binary variables to describe the magnetization of a region of the material. It makes sense that the process of renormalization should cause the model to have additional spin values that are intermediate between fully magnetized UP and fully magnetized DOWN. In order to accommodate this, we might introduce a continuum of possible magnetizations.Once we do this,the amplitude of the magnetization has a probability distribution that will be controlled by additional parameters in the Hamiltonian. These parameters may also be relevant or irrelevant. When they are irrelevant,the Ising model can be used without them. However, when they are relevant, a more complete model should be used.

The parameters that are relevant generally depend on the dimensionality of space. From our analysis of the behavior of the one-dimensional Ising model,the parameter *J* is irrelevant. It is clearly irrelevant because not only variations in *J* but *J* itself disappears as the scale increases. However, in two dimensions this is not true.

For our purposes we will be satisfied by simplifying the renormalization treatment of the two-dimensional Ising model so that no additional parameters are introduced. This can be done by a fourth renormalization group technique which has some conceptual as well as practical advantages over the others. However, it does hide the importance of determining the relevant parameters.

Bond shifting We simplify our analysis of the two-dimensional Ising model by making use of the Migdal-Kadanoff transformation. This renormalization group technique is based on the recognition that the correlation between adjacent spins can enable us to, in effect, substitute the role of one spin for another. As far as the coarser scale model is concerned, the function of the finer scale spins is to mediate the interaction between the coarser scale spins. Because one spin is correlated to the behavior of its neighbor, we can shift the responsibility for this interaction to a neighbor, and use this shift to simplify elimination of the spins.

To apply these ideas to the two-dimensional Ising model, we move some of the interactions (bonds) between spins, as shown in Fig. 1.10.10. We note that the distance over which the bonds act is preserved. The net result of the bond shifting is that we form short linear chains that can be renormalized just like a one-dimensional chain. The renormalization group transformation is thus done in two steps.First we shift the bonds, then we decimate. Once the bonds are moved, we write the renormalization of the partition function as:

$$\begin{aligned}
Z &= \sum_{\{s_i\}_A} \sum_{\{s_i\}_B} \sum_{\{s_i\}_C} e^{c \sum_i 1 + 2J \sum_i s_0(s_1 + s_2)} \\
&= \sum_{\{s_i\}} \prod_{i\ A} 2\cosh(2J(s_1 + s_2))e^{4c} \qquad (1.10.38) \\
&= \sum_{\{s_i\}} \prod_{i\ A} e^{c' + J'(s_1 s_2)}
\end{aligned}$$

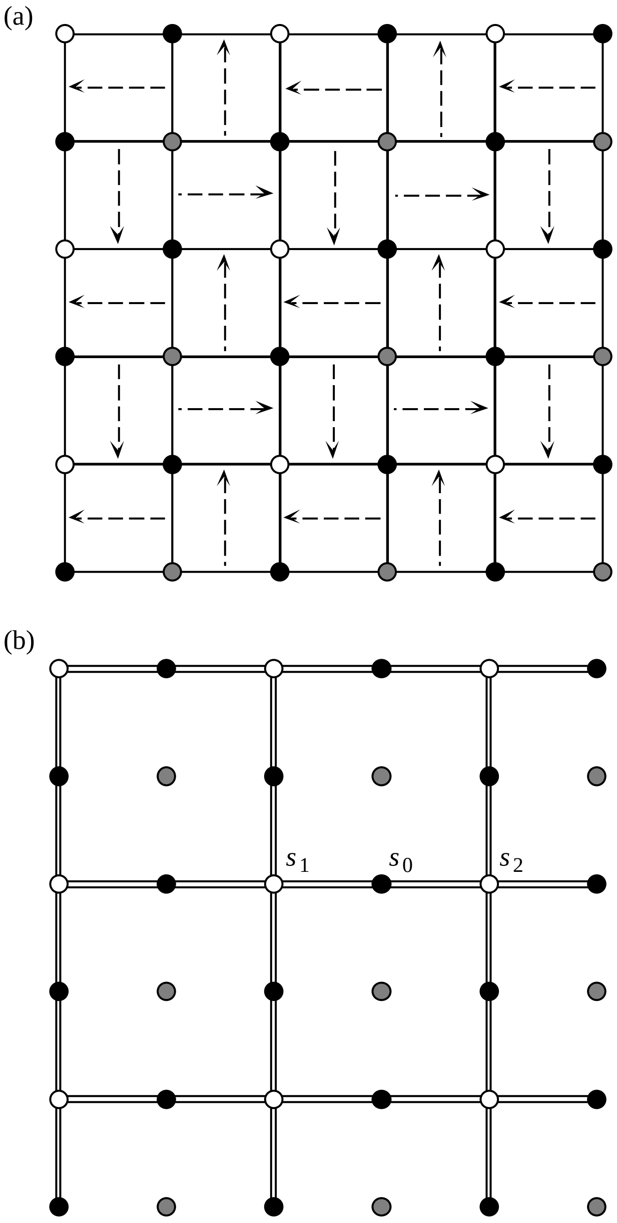


Figure 1.10.10 Illustration of the Migdal-Kadanoff renormalization transformation that enables us to bypass the formation of additional interactions. In this approach some of the interactions between spins are moved to other spins. If all the spins are aligned (at low temperature or high J), then shifting bonds doesn't affect the spin alignment. At high temperature, when the spins are uncorrelated, the interactions are not important anyway. Near the phase transition, when the spins are highly correlated, shifting bonds still makes sense. A pattern of bond movement is illustrated in (a) that gives rise to the pattern of doubled bonds in (b). Note that we are illustrating only part of a periodic lattice, so that bonds are moved into and out of the region illustrated. Using the exact renormalization of one-dimensional chains, the gray spins and the black spins can be decimated to leave only the white spins. ∎

The spin labels s_0, s_1, s_2 are assigned along each doubled bond, as indicated in Fig. 1.10.10. The three types of spin A, B and C correspond to the white, black and gray dots in the figure. The resulting equation is the same as the one we found when performing the one-dimensional renormalization group transformation with the exception of factors of two. It gives the result:

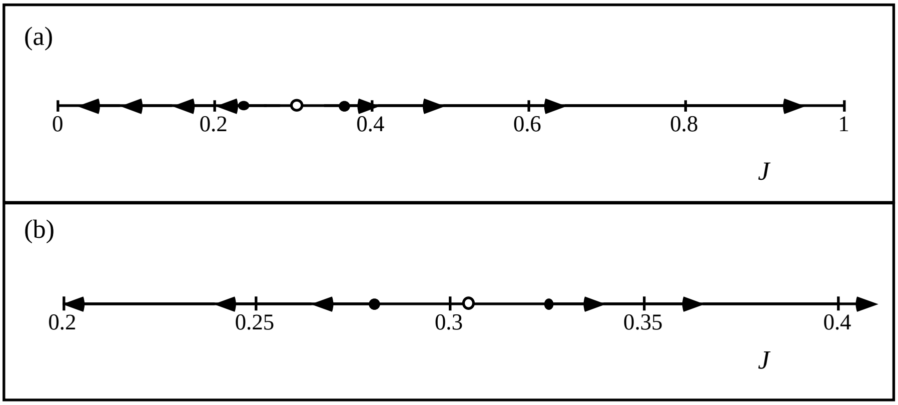


Figure 1.10.11 The two-dimensional Ising model renormalization group transformation obtained from the Migdal-Kadanoff transformation is illustrated as a flow diagram in the one-dimensional parameter space (J). The arrows show the effect of successive iterations starting from the black dots. The white dot indicates the position of the unstable fixed point, J_c, which is the phase transition in this model. Starting from values of J slightly below J_c, iteration results in the model on a large scale becoming decoupled with no interactions between spins ($J \rightarrow 0$). This is the high-temperature phase of the material. However, starting from values of J slightly above J_c iteration results in the model on the large scale becoming strongly coupled ($J \rightarrow \infty$) and spins are aligned. (a) shows only the range of values from 0 to 1, though the value of J can be arbitrarily large. (b) shows an enlargement of the region around the fixed point. ∎

$$\begin{aligned} J' &= (1/2)\ln(\cosh(4J)) \\ c' &= 4c + (1/2)\ln(4\cosh(4J)) \end{aligned} \tag{1.10.39}$$

The renormalization of J in the two-dimensional Ising model turns out to behave qualitatively different from the one-dimensional case. Its behavior is plotted in Fig. 1.10.11 using a flow diagram. There is an unstable fixed point of the iterative map at $J \approx .305$. This nonzero and noninfinite fixed point indicates that we have a phase transition. Reinserting the temperature, we see that the phase transition occurs at $\beta J = .305$ which is significantly larger than the mean field result $\beta zJ = 1$ or $\beta J = .25$ found in Section 1.6. The exact value for the phase transition for this lattice, $\beta J \approx .441$, which can be obtained analytically by other techniques, is even larger.

It turns out that there is a trick that can give us the exact transition point using a similar renormalization transformation. This trick begins by recognizing that we could have moved bonds in a larger square. For a square with b cells on a side, we would end up with each bond on the perimeter being replaced by a bond of strength b. Using Eq.(1.10.30) we can infer that a chain of b bonds of strength bJ gives rise to an effective interaction whose strength is

$$J'(b) = \tanh^{-1}(\tanh(bJ)^b) \tag{1.10.40}$$

The trick is to take the limit $b \rightarrow 1$, because in this limit we are left with the original Ising model. Extending b to nonintegral values by analytic continuation may seem a little strange, but it does make a kind of sense. We want to look at the incremental change in J as a result of renormalization, with b incrementally different from 1. This can be most easily found by taking the hyperbolic tangent of both sides of Eq. (1.10.40), and then taking the derivative with respect to b. The result is:

$$\left.\frac{dJ'(b)}{db}\right|_{b=1} = J + \sinh(J)\cosh(J)\ln(\tanh(J)) \qquad (1.10.41)$$

Setting this equal to zero to find the fixed point of the transformation actually gives the exact result for the phase transition.

The renormalization group gives us more information than just the location of the phase transition. Fig. 1.10.11 shows changes that occur in the parameters as the length scale is varied. We can use this picture to understand the behavior of the Ising model in some detail. It shows what happens on longer length scales by the direction of the arrows. If the flow goes toward a particular point,then we can tell that on the longest (thermodynamic) length scale the behavior will be characterized by the behavior of the model at that point. By knowing how close we are to the original phase transition, we can also learn from the renormalization group what is the length scale at which the behavior characteristic of the phase transition will disappear. This is the length scale at which the iterative map leaves the region of the repelling fixed point and moves to the attracting one.

We can also characterize the relationship between systems at different values of the parameters: temperatures or magnetic fields. Renormalization takes us from a system at one value of βJ to another. Thus, we can relate the behavior of a system at one temperature to another by performing the renormalization for both systems and stopping both at a particular value of βJ. At this point we can directly relate properties of the two systems, such as their free energies. Different numbers of renormalization steps in the two cases mean that we are relating the two systems at different scales. Such descriptions of relationships of the properties of one system at one scale with another system at a different scale are known as scaling functions because they describe how the properties of the system change with scale.

The renormalization group was developed as an analytic tool for studying the scaling properties of systems with spatially arrayed interacting parts. We will study another use of renormalization in Section 1.10.5. Then in Section 1.10.6 we will introduce a computational approach—the multigrid method.

Question 1.10.6 In this section we displayed our iterative maps graphically as flow diagrams, because in renormalization group transformations we are often interested in maps that involve more than one variable. Make a diagram like Fig. 1.1.1 for the single variable J showing the iterative renormalization group transformation for the two-dimensional Ising model as given in Eq. (1.10.39).

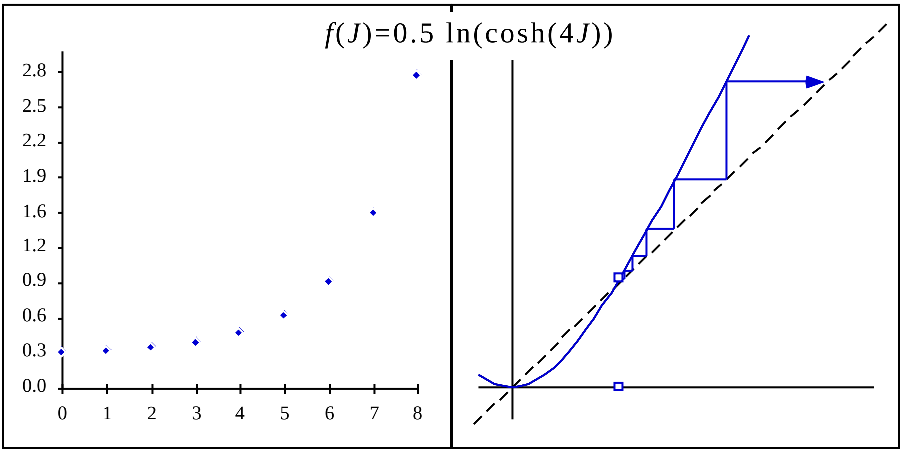


Figure 1.10.12 The iterative map shown as a flow diagram in Fig. 1.10.11 is shown here in the same manner as the iterative maps in Section 1.1. On the left are shown the successive values of J as iteration proceeds. Each iteration should be understood as a loss of detail in the model and hence an observation of the system on a larger scale. Since in general our observations of the system are macroscopic, we typically observe the limiting behavior as the iterations go to ∞. This is similar to considering the limiting behavior of a standard iterative map. On the right is the graphical method of determining the iterations as discussed in Section 1.1. The fixed points are visible as intersections of the iterating function with the diagonal line. ∎

Solution 1.10.6 See Fig. 1.10.12. The fixed point and the iterative behavior are readily apparent. ∎

1.10.5 *Renormalization and chaos*

Our final example of renormalization brings us back to Section 1.1, where we studied the properties of iterative maps and the bifurcation route to chaos. According to our discussion, cycles of length 2^k, k = 0,1,2,..., appeared as the parameter a was varied from 0 to a_c = 3.56994567, at which point chaotic behavior appeared.Fig. 1.1.3 summarizes the bifurcation route to chaos.A schematic of the bifurcation part of this diagram is reproduced in Fig. 1.10.13. A brief review of Section 1.1 may be useful for the following discussion.

The process of bifurcation appears to be a self-similar process in the sense that the appearance of a 2-cycle for $f(s)$ is repeated in the appearance of a 2-cycle for $f^2(s)$, but over a smaller range of a. The idea of self-similarity seems manifest in Fig. 1.10.13, where we would only have to change the scale of magnification in the s and a directions in order to map one bifurcation point onto the next one. While this mapping might not work perfectly for smaller cycles, it becomes a better and better

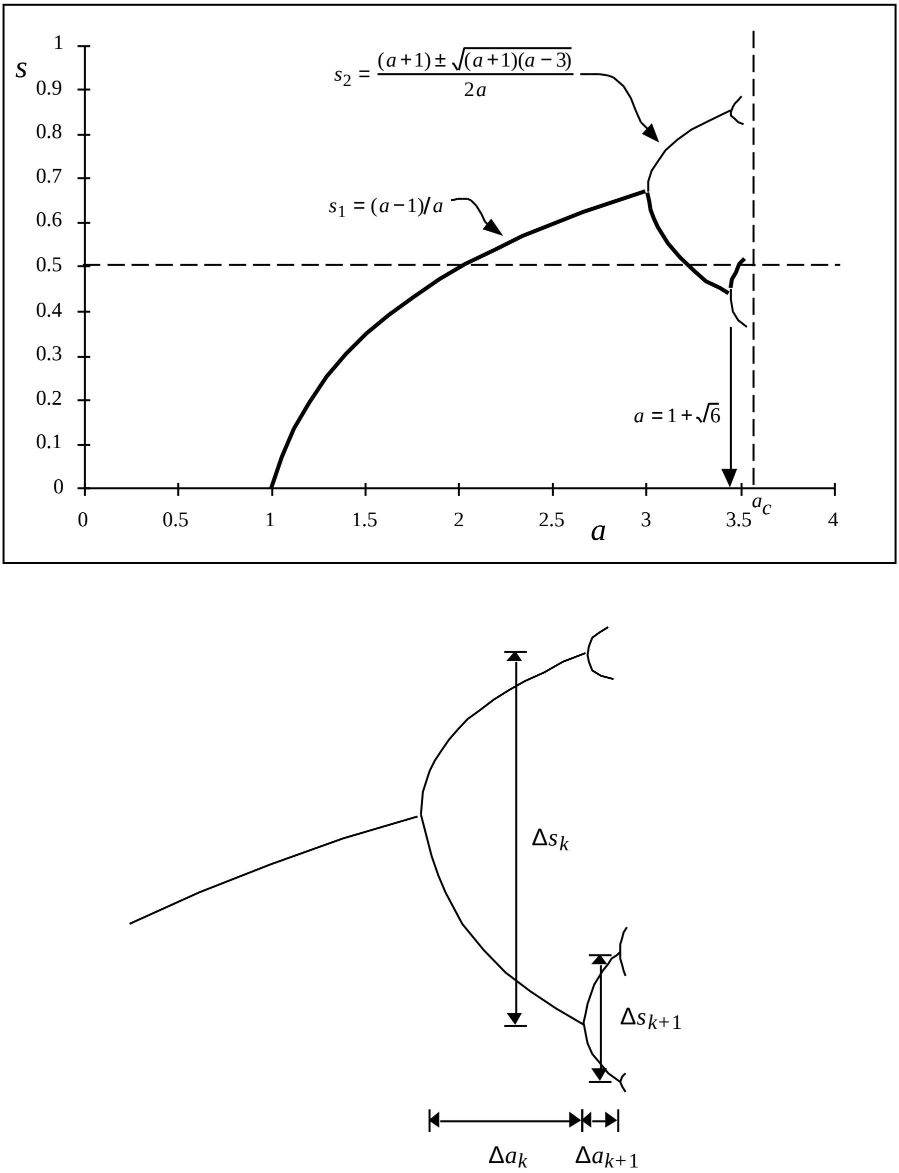


Figure 1.10.13 Schematic reproduction of Fig. 1.1.4, which shows the bifurcation route to chaos. Successive branchings are approximately self-similar. The bottom figure shows the definition of the scaling factors that relate the successive branchings. The horizontal rescaling of the branches, δ, is given by the ratio of Δa_k to Δa_{k+1}. The vertical rescaling of the branches, α, is given by the ratio of Δs_k to Δs_{k+1}. The top figure shows the values from which we can obtain a first approximation to the values of α and δ, by taking the ratios from the zeroth, first and second bifurcations. The zeroth bifurcation point is actually the point $a = 1$. The first bifurcation point occurs at $a = 3$. the second occurs at $a = 1 + \sqrt{6}$. The values of s at the bifurcation points were obtained in Section 1.1, and formulas are indicated on the figure. When the scaling behavior of the tree is analyzed using a renormalization group treatment, we focus on the tree branches that cross $s = 1/2$. These are indicated by bold lines in the top figure. ∎

approximation as the number of cycles increases. The bifurcation diagram is thus a treelike object. This means that the sequence of bifurcation points forms a geometrically converging sequence, and the width of the branches is also geometrically converging. However, the distances in the s and a directions are scaled by different factors. The factors that govern the tree rescaling at each level are δ and α, as shown in Fig. 1.10.13 (b):

$$\begin{aligned}\delta &= \lim_{k\to\infty} \frac{\Delta a_k}{\Delta a_{k+1}} \\ \alpha &= \lim_{k\to\infty} \frac{\Delta s_k}{\Delta s_{k+1}}\end{aligned} \tag{1.10.42}$$

By this convention,the magnitude of both α and δ is greater than one. α is defined to be negative because the longer branch flips up to down at every branching. The values are to be obtained by taking the limit as $k \to \infty$ where these scale factors have well-defined limits.

We can find a first approximation to these scaling factors by using the values at the first and second bifurcations that we calculated in Section 1.1. These values, given in Fig. 1.10.13, yield:

$$\delta \approx (3-1)/(1+\sqrt{6}-3) = 4.449 \tag{1.10.43}$$

$$\alpha \approx \frac{2s_1\big|_{a=3}}{\left(s_2^+ - s_2^-\right)\big|_{a=1+\sqrt{6}}} = \frac{4}{3}\frac{a}{\sqrt{(a+1)(a-3)}}\Bigg|_{a=1+\sqrt{6}} = 3.252 \tag{1.10.44}$$

Numerically, the asymptotic value of δ for large k is found to be 4.6692016. This differs from our first estimate by only 5%. The numerical value for α is 2.50290787, which differs from our first estimate by a larger margin of 30%.

We can determine these constants with greater accuracy by studying directly the properties of the functions $f, f^2, \ldots f^{2^k} \ldots$ that are involved in the formation of 2^k cycles. In order to do this we modify our notation to explicitly include the dependence of the function on the parameter a. $f(s,a), f^2(s,a)$, etc. Note that iteration of the function f only applies to the first argument.

The tree is formed out of curves $s_{2^k}(a)$ that are obtained by solving the fixed point equation:

$$s_{2^k}(a) = f^{2^k}(s_{2^k}(a), a) \tag{1.10.45}$$

We are interested in mapping a segment of this curve between the values of s where

$$\frac{df^{2^k}(s,a)}{ds} = 1 \tag{1.10.46}$$

and

$$\frac{df^{2^k}(s,a)}{ds} = -1 \tag{1.10.47}$$

onto the next function, where k is replaced everywhere by $k + 1$. This mapping is a kind of renormalization process similar to that we discussed in the previous section. In order to do this it makes sense to expand this function in a power series around an intermediate point, which is the point where these derivatives are zero. This is known as the superstable point of the iterative map. The superstable point is very convenient for study, because for any value of k there is a superstable point at $s = 1/2$. This follows because $f(s,a)$ has its maximum at $s = 1/2$, and so its derivative is zero. By the chain rule,the derivative of $f^{2^k}(s,a)$,is also zero. As illustrated in Fig. 1.10.13,the line at $s = 1/2$ intersects the bifurcation tree at every level of the hierarchy at an intermediate point between bifurcation points. These intersection points must be superstable.

It is convenient to displace the origin of s to be at $s = 1/2$,and the origin of a to be at the convergence point of the bifurcations.We thus define a function g which represents the structure of the tree. It is approximately given by:

$$g(s,a) \approx f(s + 1/2,a + a_c) - 1/2 \tag{1.10.48}$$

However, we would like to represent the idealized tree rather than the real tree. The idealized tree would satisfy the scaling relation exactly at all values of a. Thus g should be the analog of the function f which would give us an ideal tree. To find this function we need to expand the region near $a = a_c$ by the appropriate scaling factors. Specifically we define:

$$g(s,a) = \lim_{k \to \infty} \alpha^k f^{2^k} (s / \alpha^k + 1/2, a / \delta^k + a_c) - 1/2 \tag{1.10.49}$$

The easiest way to think about the function $g(s,a)$ is that it is quite similar to the quadratic function $f(s,a)$ but it has the form necessary to cause the bifurcation tree to have the ideal scaling behavior at every branching. We note that $g(s,a)$ depends on the behavior of $f(s,a)$ only very near to the point $s = 1/2$. This is apparent in Eq. (1.10.49) since the region near $s = 1/2$ is expanded by a factor of α^k.

We note that $g(s,a)$ has its maximum at s = 0. This is a consequence of the shift in origin that we chose to make in defining it.

Our objective is to find the form of $g(s,a)$ and,with this form,the values of α and δ. The trick is to recognize that what we need to know can be obtained directly from its scaling properties. To write the scaling properties we look at the relationship between successive iterations of the map and write:

$$g(s,a) = \alpha g^2(s/\alpha,a/\delta) \tag{1.10.50}$$

This follows either from our discussion and definition of the scaling parameters α and δ or directly from Eq. (1.10.49).

For convenience, we analyze Eq. (1.10.50) first in the limit $a \to 0$. This corresponds to looking at the function $g(s,a)$ as a function of s at the limit of the bifurcation sequence. This function (Fig. 1.10.14) still looks quite similar to our original function $f(s)$, but its specific form is different. It satisfies the relationship:

$$g(s,0) = g(s) = \alpha g^2(s/\alpha) \tag{1.10.51}$$

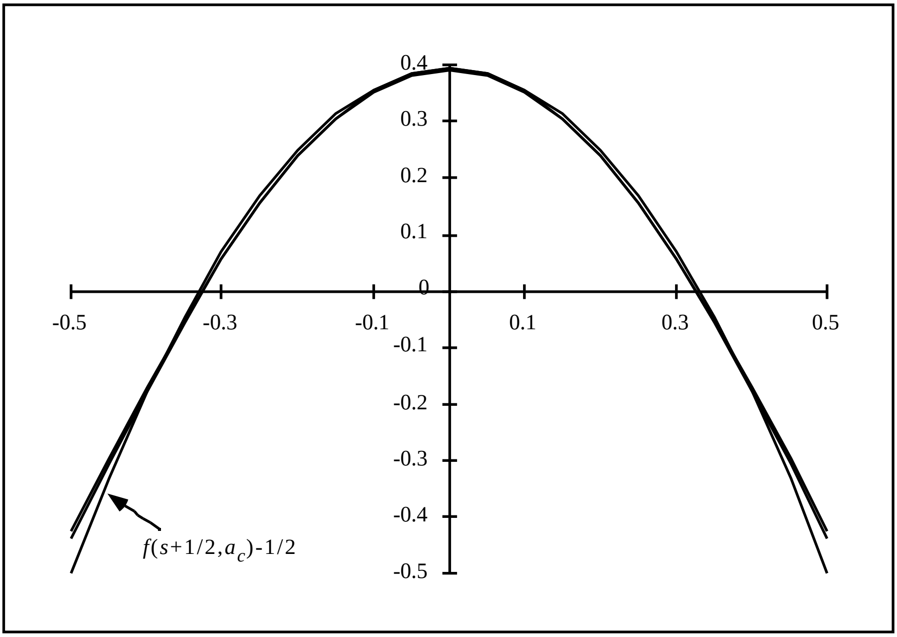


Figure 1.10.14 Three functions are plotted that are successive approximations to $g(s) = g(s, 0)$. This function is essentially the limiting behavior of the quadratic iterative map $f(s)$ at the end of the bifurcation tree a_c. The functions plotted are the first three k values inserted in Eq. (1.10.49): $f(s + 1/2, a + a_c) - 1/2$, $af^2(s/\alpha + 1/2, a_c) - 1/2$ and $a^2f^4(s/\alpha^2 + 1/2, a_c) - 1/2$. The latter two are almost indistinguishable, indicating that the sequence of functions converges rapidly. ▮

We approximate this function by a quadratic with no linear term because $g(s)$ has its maximum at $s = 0$:

$$g(s) \approx g_0 + cs^2 \tag{1.10.52}$$

Inserting into Eq. (1.10.51) we obtain:

$$g_0 + cs^2 \approx \alpha(g_0 + c(g_0 + c(s / \alpha)^2)^2) \tag{1.10.53}$$

Equating separately the coefficients of the first and second terms in the expansion gives the solution:

$$\begin{gathered}\alpha = 1 / (1 + cg_0) \\ \alpha = 2cg_0\end{gathered} \tag{1.10.54}$$

We see that c and g_0 only appear in the combination cg_0, which means that there is one parameter that is not determined by the scaling relationship. However, this does not prevent us from determining α. Eq. (1.10.54) can be solved to obtain:

$$\begin{aligned} cg_0 &= (-1 \pm \sqrt{3})/2 = -1.3660254 \\ \alpha &= (-1 \pm \sqrt{3}) = -2.73205081 \end{aligned} \qquad (1.10.55)$$

We have chosen the negative solutions because the value of α and the value of cg_0 must be negative.

We return to consider the dependence of $g(s,a,)$ on a to obtain a new estimate for δ. Using a first-order linear dependence on a we have:

$$g(s,a,) \approx g_0 + cs^2 + ba \qquad (1.10.56)$$

Inserting into Eq. (1.10.50) we have:

$$g_0 + cs^2 + ba \approx \alpha(g_0 + c(g_0 + c(s / \alpha)^2 + ba / \delta)^2 + ba / \delta) \qquad (1.10.57)$$

Taking only the first order term from this equation in a we have:

$$\delta = 2\alpha cg_0 + \alpha = 4.73205 \qquad (1.10.58)$$

Eq. (1.10.55) and Eq.(1.10.58) are a significant improvement over Eq.(1.10.44) and Eq.(1.10.43). The new value of α is less than 10% from the exact value. The new value of δ is less than 1.5% from the exact value. To improve the accuracy of the results, we need only add additional terms to the expansion of $g(s,a)$ in s. The first-order term in a is always sufficient to obtain the corresponding value of δ.

It is important, and actually central to the argument in this section, that the explicit form of $f(s,a)$ never entered into our discussion. The only assumption was that the functional behavior near the maximum is quadratic. The rest of the argument follows independent of the form of $f(s,a)$ because we are looking at its properties after many iterations. These properties depend only on the region right in the vicinity of the maximum of the function. Thus only the first-order term in the expansion of the original function $f(s,a)$ matters. This illustrates the notion of universality so essential to the concept of renormalization—the behavior is controlled by very few parameters. All other parameters are irrelevant—changing their values in the original iterative map is irrelevant to the behavior after many iterations (many renormalizations) of the iterative map. This is similar to the study of renormalization in models like the Ising model, where most of the details of the behavior at small scales no longer matter on the largest scales.

1.10.6 *Multigrid*

The multigrid technique is designed for the solution of computational problems that benefit from a description on multiple scales. Unlike renormalization, which is largely an analytic tool,the multigrid method is designed specifically as a computational tool. It works well when a problem can be approximated using a description on a coarse lattice, but becomes more and more accurate as the finer-scale details on finer-scale lattices are included. The idea of the method is not just to solve an equation on finer and finer levels of description, but also to correct the coarser-scale equations based on the finer-scale results. In this way the methodology creates an improved description of the problem on the coarser-scale.

The multigrid approach relies upon iterative refinement of the solution. Solutions on coarser scales are used to approximate the solutions on finer scales. The finer-scale solutions are then iteratively refined. However, by correcting the coarser-scale equations,it is possible to perform most of the iterative refinement of the fine-scale solution on the coarser scales. Thus the iterative refinement of the solution is based both upon correction of the solution and correction of the equation. The idea of correcting the equation is similar in many ways to the renormalization group approach. However, in this case it is a particular solution, which may be spatially dependent, rather than an ensemble averaging process, which provides the correction.

We explain the multigrid approach using a conventional problem, which is the solution of a differential equation. To solve the differential equation we will find an approximate solution on a grid of points.Our ultimate objective is to find a solution on a fine enough grid so that the solution is within a prespecified accuracy of the exact answer. However, we will start with a much coarser grid solution and progressively refine it to obtain more accurate results. Typically the multigrid method is applied in two or three dimensions, where it has greater advantages than in one dimension. However, we will describe the concepts in one dimension and leave out many of the subtleties.

For concreteness we will assume a differential equation which is:

$$\frac{d^2 f(x)}{dx^2} = g(x) \tag{1.10.59}$$

where $g(x)$ is specified. The domain of the equation is specified, and boundary conditions are provided for $f(x)$ and its derivative.On a grid of equally spaced points we might represent this equation as:

$$\frac{1}{d^2}(f(i+1) + f(i-1) - 2f(i)) = g(i) \tag{1.10.60}$$

This can be written as a matrix equation:

$$\sum_j A(i,j)\, f(j) = g(i) \tag{1.10.61}$$

The matrix equation can be solved for the values of $f(i)$ by matrix inversion (using matrix diagonalization). However, diagonalization is very costly when the matrix is large, i.e., when there are many points in the grid.

A multigrid approach to solving this equation starts by defining a set of lattices (grids), $L_j, j \quad \{0,\ldots,q\}$, where each successive lattice has twice as many points as the previous one (Fig. 1.10.15). To explain the procedure it is simplest to assume that we start with a good approximation for the solution on grid L_{j-1} and we are looking for a solution on the grid L_j. The steps taken are then:

1. Interpolate to find $f_0^j(i)$, an approximate value of the function on the finer grid L_j.

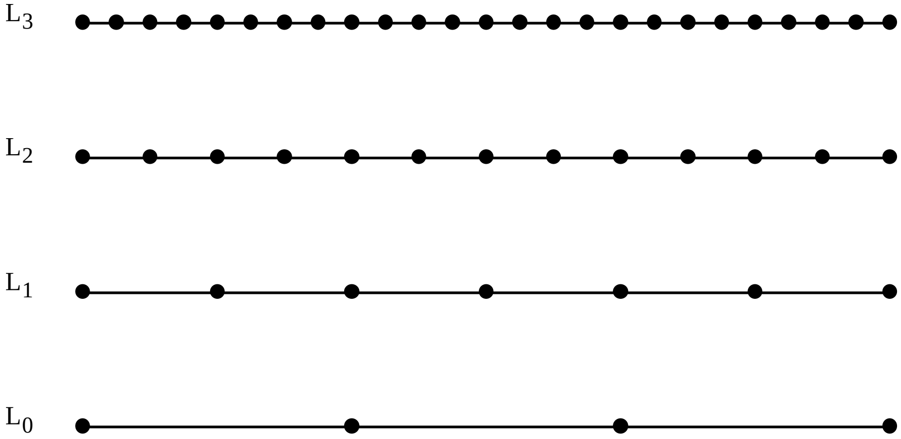


Figure 1.10.15 Illustration of four grids for a one-dimensional application of the multigrid technique to a differential equation by the procedure illustrated in Fig. 1.10.16. ∎

2. Perform an iterative improvement (relaxation) of the solution on the finer grid. This iteration involves calculating the error

$$\sum_{i'} A(i,i') f_0^j(i') - g(i) = r^j(i) \qquad (1.10.62)$$

where all indices refer to the grid L_j. This error is used to improve the solution on the finer grid, as in the minimization procedures discussed in Section 1.7.5:

$$f_1^j(i) = f_0^j(i) - cr^j(i) \qquad (1.10.63)$$

The scalar c is generally replaced by an approximate inverse of the matrix $A(i,j)$ as discussed in Section 1.7.5. This iteration captures much of the correction of the solution at the fine-scale level; however, there are resulting corrections at coarser levels that are not captured. Rather than continuing to iteratively improve the solution at this fine-scale level, we move the iteration to the next coarser level.

3. Recalculate the value of the function on the coarse grid L_{j-1} to obtain $f_1^{j-1}(i)$. This might be just a restriction from the fine-grid points to the coarse-grid points. However, it often involves some more sophisticated smoothing. Ideally, it should be such that interpolation will invert this process to obtain the values that were found on the finer grid. The correction for the difference between the interpolated and exact fine-scale results are retained.

4. Correct the value of $g(i)$ on the coarse grid using the values of $r^j(i)$ restricted to the coarser grid. We do this so that the coarse-grid equation has an exact solution that is consistent with the fine-grid equation. From Eq. (1.10.62) this essentially means adding $r^j(i)$ to $g(i)$. The resulting corrected values we call $g_1^{j-1}(i)$.

5. Relax the solution $f_1^{j-1}(i)$ on the coarse grid to obtain a new approximation to the function on the coarse grid $f_2^{j-1}(i)$. This is done using the same procedure for relaxation described in step 3; however $g(i)$ is replaced by $g_1^{j-1}(i)$.

The procedure of going to coarser grids in steps 3 through 5 is repeated for all grids $L_{j-2}, L_{j-3},\ldots$ till the coarsest grid, L_0. The values of the function $g(i)$ are progressively corrected by the finer-scale errors. Step 5 on the coarsest grid is replaced by exact solution using matrix diagonalization. The subsequent steps are designed to bring all of the iterative refinements to the finest-scale solution.

6. Interpolate the coarse-grid solution L_0 to the finer-grid L_1.
7. Add the correction that was previously saved when going from the fine to the coarse grid.

Steps 6–7 are then repeated to take us to progressively finer-scale grids all the way back to L_j.

This procedure is called a V-cycle since it appears as a V in a schematic that shows the progressive movement between levels. A V-cycle starts from a relaxed solution on grid L_{j-1} and results in a relaxed solution on the grid L_j. A full multigrid procedure involves starting with an exact solution at the coarsest scale L_0 and then performing V-cycles for progressively finer grids. Such a multigrid procedure is graphically illustrated in Fig. 1.10.16.

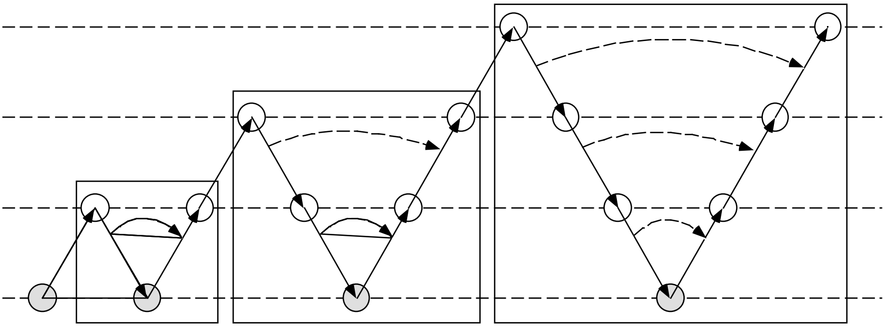


Figure 1.10.16 The multigrid algorithm used to obtain the solution to a differential equation on the finest grid is described schematically by this sequence of operations. The operation sequence is to be read from left to right. The different grids that are being used are indicated by successive horizontal lines with the coarsest grid on the bottom and the finest grid on the top. The sequence of operations starts by solving a differential equation on the coarsest grid by exact matrix diagonalization (shaded circle). Then iterative refinement of the equations is performed on finer grids. When the finer-grid solutions are calculated, the coarse-grid equations are corrected so that the iterative refinement of the fine-scale solution can be performed on the coarse grids. This involves a V-cycle as indicated in the figure by the boxes. The horizontal curved arrows indicate the retention of the difference between coarse- and fine-scale solutions so that subsequent refinements can be performed. ▮

There are several advantages of the multigrid methodology for the solution of differential equations over more traditional integration methods that use a single-grid representation. With careful implementation, the increasing cost of finer-scale grids grows slowly with the number of grid points, scaling as $N \ln(N)$. The solution of multiple problems of similar type can be even more efficient,since the corrections of the coarse-scale equations can often be carried over to similar problems, accelerating the iterative refinement. This is in the spirit of developing universal coarse-scale representations as discussed earlier. Finally, it is natural in this method to obtain estimates of the remaining error due to limited grid density, which is important to achieving a controlled error in the solution.

1.10.7 *Levels of description, emergence of simplicity and complexity*

In our explorations of the world we have often discovered that the natural world may be described in terms of underlying simple objects, concepts, and laws of behavior (mechanics) and interactions. When we look still closer we see that these simple objects are composite objects whose internal structure may be complex and have a wealth of possible behavior. Somehow, the wealth of behavior is not relevant at the larger scale. Similarly, when we look at longer length scales than our senses normally are attuned to, we discover that the behavior at these length scales is not affected by objects and events that appear important to us.

Examples are found from the behavior of galaxies to elementary particles: galaxies are composed of suns and interstellar gases, suns are formed of complex plasmas and are orbited by planets, planets are formed from a diversity of materials and even life, materials and living organisms are formed of atoms,atoms are composed of nuclei and electrons, nuclei are composed of protons and neutrons (nucleons),and nucleons appear to be composed of quarks.

Each of these represents what we may call a level of description of the world. A level is an internally consistent picture of the behavior of interacting elements that are simple. When taken together, many such elements may or may not have a simple behavior, but the rules that give rise to their collective behavior are simple. We note that the interplay between levels is not always just a self-contained description of one level by the level immediately below. At times we have to look at more than one level in order to describe the behavior we are interested in.

The existence of these levels of description has led science to develop the notion of fundamental law and unified theories of matter and nature. Such theories are the self-consistent descriptions of the simple laws governing the behavior and interplay of the entities on a particular level. The laws at a particular level then give rise to the larger-scale behavior.

The existence of simplicity in the description of underlying fundamental laws is not the only way that simplicity arises in science. The existence of multiple levels implies that simplicity can also be an emergent property. This means that the collective behavior of many elementary parts can behave simply on a much larger scale.

The study of complex systems focuses on understanding the relationship between simplicity and complexity. This requires both an understanding of the emergence of complex behavior from simple elements and laws, as well as the emergence of simplicity from simple or complex elements that allow a simple larger-scale description to exist.

Much of our discussion in this section was based upon the understanding that macroscopic behavior of physical systems can be described or determined by only a few relevant parameters. These parameters arise from the underlying microscopic description. However, many of the aspects of the microscopic description are irrelevant. Different microscopic models can be used to describe the same macroscopic phenomenon. The approach of scaling and renormalization does not assume that all the details of the microscopic description become irrelevant, however, it tries to determine self-consistently which of the microscopic parameters are relevant to the macroscopic behavior in order to enable us to simplify our analysis and come to a better understanding.

Whenever we are describing a simple macroscopic behavior, it is natural that the number of microscopic parameters relevant to model this behavior must be small. This follows directly from the simplicity of the macroscopic behavior. On the other hand,if we describe a complex macroscopic behavior, the number of microscopic parameters that are relevant must be large.

Nevertheless, we know that the renormalization group approach has some validity even for complex systems. At long length scales, all of the details that occur on the smallest length scale are not relevant. The vibrations of an individual atom are not generally relevant to the behavior of a complex biological organism. Indeed, there is a pattern of levels of description in the structure of complex systems. For biological organisms, composed out of atoms, there are additional levels of description that are intermediate between atoms and the organism: molecules, cells, tissues, organs and systems. The existence of these levels implies that many of the details of the atomic behavior are not relevant at the macroscopic level. This should also be understood from the perspective of the multi-grid approach. In this picture, when we are describing the behavior of a complex system, we have the possibility of describing it at a very coarse level or a finer and yet finer level. The number of levels that are necessary depends on the level of precision or level of detail we wish to achieve in our description. It is not always necessary to describe the behavior in terms of the finest scale. It is essential, however, to identify properly a model that can capture the essential underlying parameters in order to discuss the behavior of any system.

Like biological organisms, man-made constructs are also built from levels of structure. This method of organization is used to simplify the design and enable us to understand and work with our own creations. For example, we can consider the construction of a factory from machines and computers,machines constructed from individual moving parts, computers constructed from various components including computer chips, chips constructed from semiconductor devices, semiconductor devices composed out of regions of semiconductor and metal. Both biology and engineering face problems of design for function or purpose. They both make use of interacting building blocks to engineer desired behavior and therefore construct the complex out of the simple. The existence of these building blocks is related to the existence of levels of description for both natural and artificial systems.

Our discussion thus brings us to recognize the importance of studying the properties of substructure and its relationship to function in complex systems. This relationship will be considered in Chapter 2 in the context of our study of neural networks.