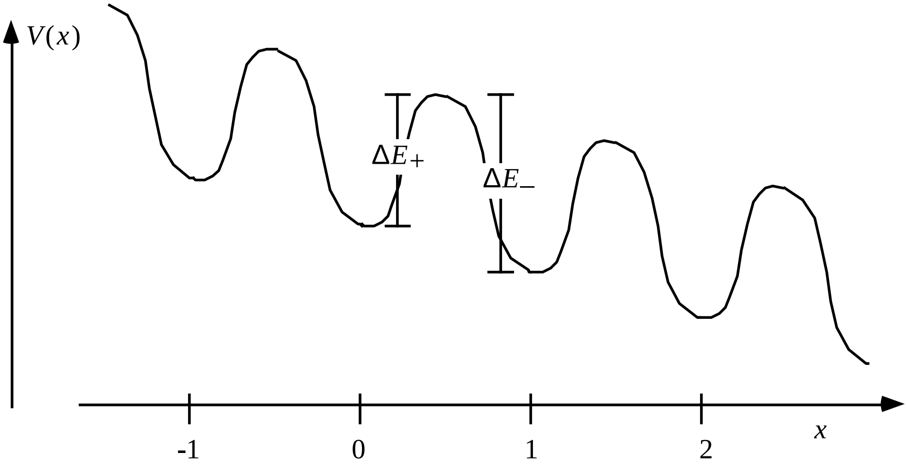


Figure 1.4.5 The biased random walk is also found in a multiple-well system when the illustrated washboard potential is used. The velocity of the system is given by the difference in hopping rates to the right and to the left. ▮

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.

$$s(i, j, k;t) = s(x_i, y_j, z_k;t) \qquad (1.5.1)$$

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:

$$s(i, j, k;t) = R(\{s(i' - i, j' - j, k' - k;t - 1)\}_{i',j',k' \quad Z}) \qquad (1.5.2)$$

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:

$$s(i, j, k;t) = R(s(i \pm 1{,}0, j \pm 1, 0, k \pm 1, 0;t - 1)) \qquad (1.5.3)$$

*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:

$$s(i;t + 1) = \mathrm{sign}(s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad (1.5.4)$$

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:

$$s(i;t + 1) = \mathrm{mod}_2\,(s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad (1.5.5)$$

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. ∎

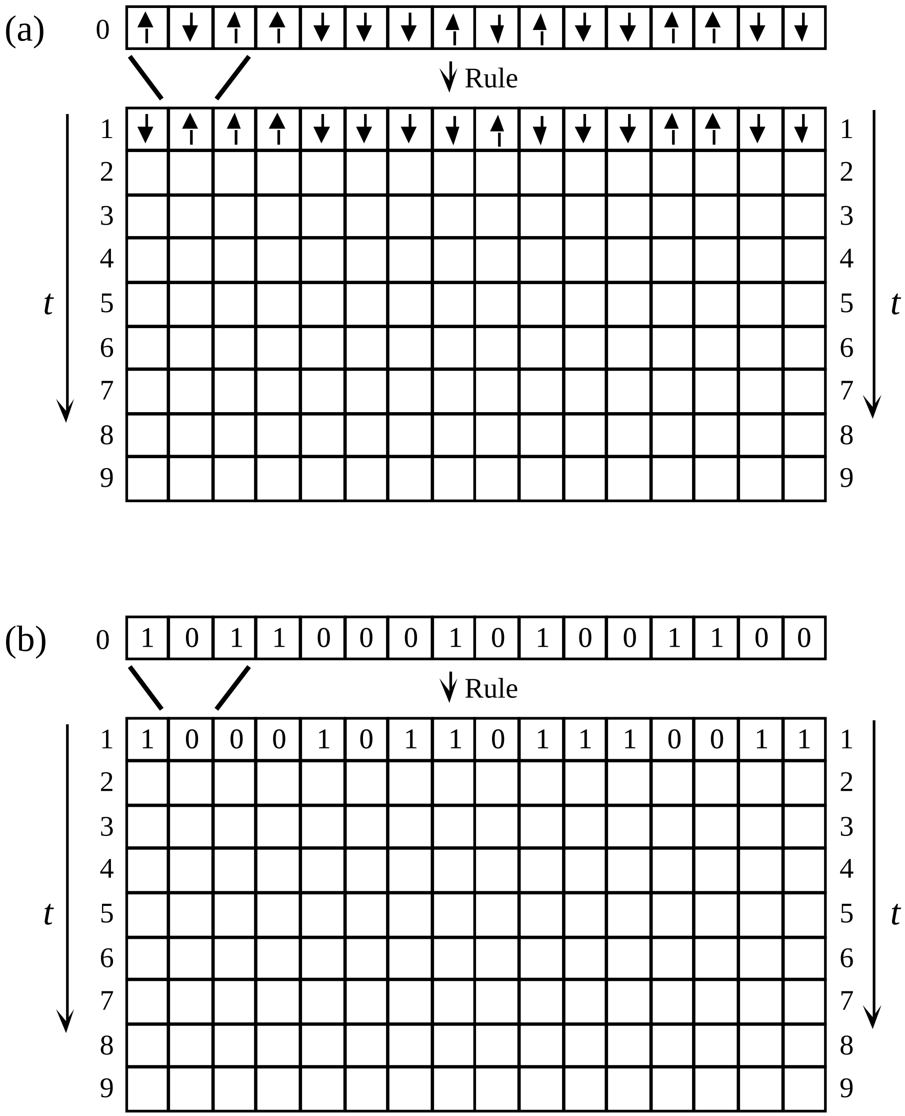


Figure 1.5.1 Two examples of one dimensional (1-d) cellular automata. The top row in each case gives the initial conditions. The value of a cell at a particular time is given by a rule that depends on the values of the cells in its neighborhood at the previous time. For these rules the neighborhood consists 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.

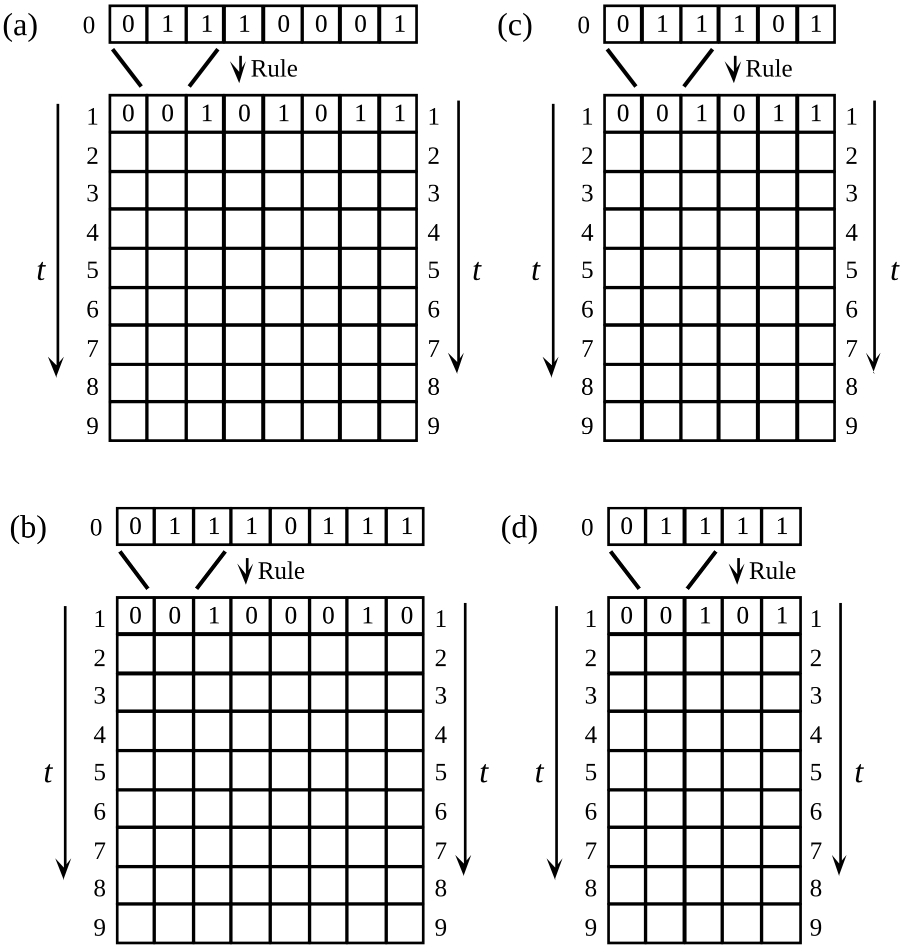


Figure 1.5.2 Four additional examples for the mod2 rule that have different initial conditions with specific periodicity: (a) is periodic in 8 cells, (b) is periodic in 4 cells, though it is shown embedded in a space of periodicity 8, (c) is periodic in 6 cells, (d) is periodic in 5 cells. By filling 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:

$$s_1 = s_2 \qquad s_1 + s = s_2 + s \qquad \text{mod}_2 \qquad (1.5.6)$$

as well as the special property:

$$2s = 0 \qquad \text{mod}_2 \qquad (1.5.7)$$

Together these imply that variables may be moved from one side of the equality to the other:

$$s_1 + s = s_2 \qquad s_1 = s_2 + s \qquad \text{mod}_2 \qquad (1.5.8)$$

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))

$$s(i;t + 1) = (s(i - 1;t) + s(i;t) + s(i + 1;t)) \qquad \text{mod}_2 \qquad (1.5.9)$$

can be rewritten to give us the value of a cell in a layer in terms of the next layer and its own neighbors:

$$s(i - 1;t) = s(i;t + 1) + s(i;t) + s(i + 1;t) \qquad \text{mod}_2 \qquad (1.5.10)$$

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:

$$s(i - 1;t) = s(i;t + 1) + s(i + 1;t + 1) + s(i + 2;t) \qquad \text{mod}_2 \qquad (1.5.12)$$

It is convenient to rewrite this with one higher index:

$$s(i;t) = s(i + 1;t + 1) + s(i + 2;t + 1) + s(i + 3;t) \qquad \text{mod}_2 \qquad (1.5.13)$$

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:

$$\begin{aligned} s(i;t) &= s(i+1;t+1) + s(i+2;t+1) \\ &+ s(i+4;t+1) + s(i+5;t+1) + s(i+6;t) \end{aligned} \quad \text{mod}_2 \qquad (1.5.14)$$

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

$$\begin{aligned} s(i;t) &= s(i+1;t+1) + s(i+2;t+1) \\ &+ s(i+4;t+1) + s(i+5;t+1) + s(i+6;t) \\ &+ \ldots \\ &+ s(i+N-2;t+1) + s(i+N-1;t+1) + s(i;\, t) \end{aligned} \quad \text{mod}_2 \qquad (1.5.15)$$

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.

$$\begin{aligned} 0 &= s(i+1;\, t+1) + s(i+2;\, t+1) \\ &+ s(i+4;\, t+1) + s(i+5;\, t+1) + s(i+6;\, t) \\ &+ \ldots \\ &+ s(i+N-2;\, t+1) + s(i+N-1;\, t+1) \end{aligned} \quad \text{mod}_2 \qquad (1.5.16)$$

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:

$$\begin{aligned}&1 - s(i;t+1) = 1 - (s(i-1;t) + s(i;t) + s(i+1;t))\\&= (1 - s(i-1;t)) + (1 - s(i;t)) + (1 - s(i+1;t))) - 2 \qquad \mathrm{mod}_2 \qquad (1.5.17)\\&= (1 - s(i-1;t)) + (1 - s(i;t)) + (1 - s(i+1;t)))\end{aligned}$$

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:

$$s(i;t) = s(i;t+1) + \sum_{j=1}^{(N-1)/3} (s(i+3j-2;t+1) + s(i+3j;t+1)) \qquad \mathrm{mod}_2 \qquad (1.5.18)$$

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.

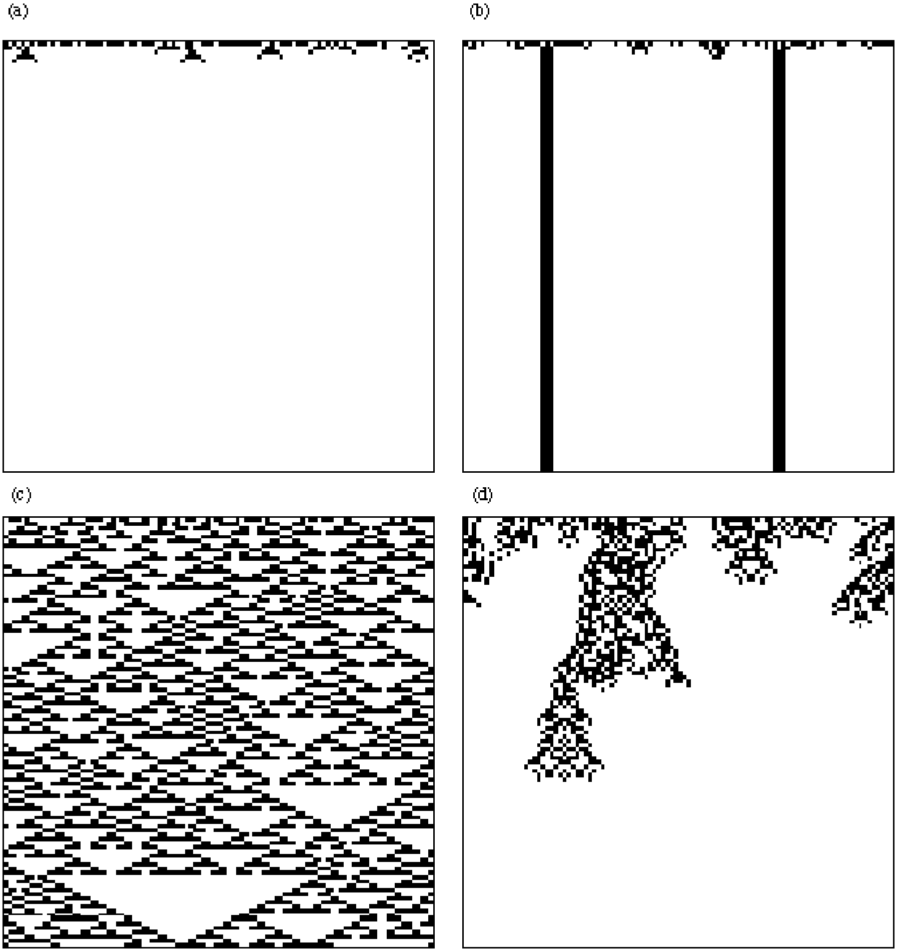


Figure 1.5.3 Illustration of four CA update rules with random initial conditions that are in a periodic space with a period of 100 cells. The initial conditions are shown at the top and time proceeds downward. Each is updated for 100 steps. ON cells are indicated as filled squares. OFF cells are not 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

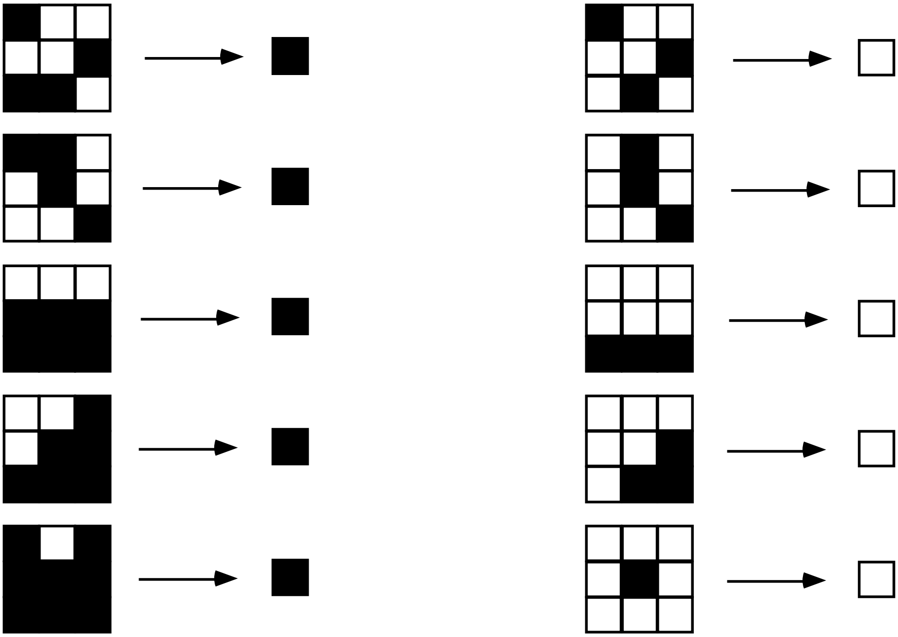


Figure 1.5.4 Illustration of a 2-d CA that may be thought of as a simple model of droplet condensation. The rule sets a cell to be ON (condensed) if four or more of its neighbors are condensed in the previous time, and OFF (uncondensed) otherwise. There are a total of 2^9=512 possible initial configurations; 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.

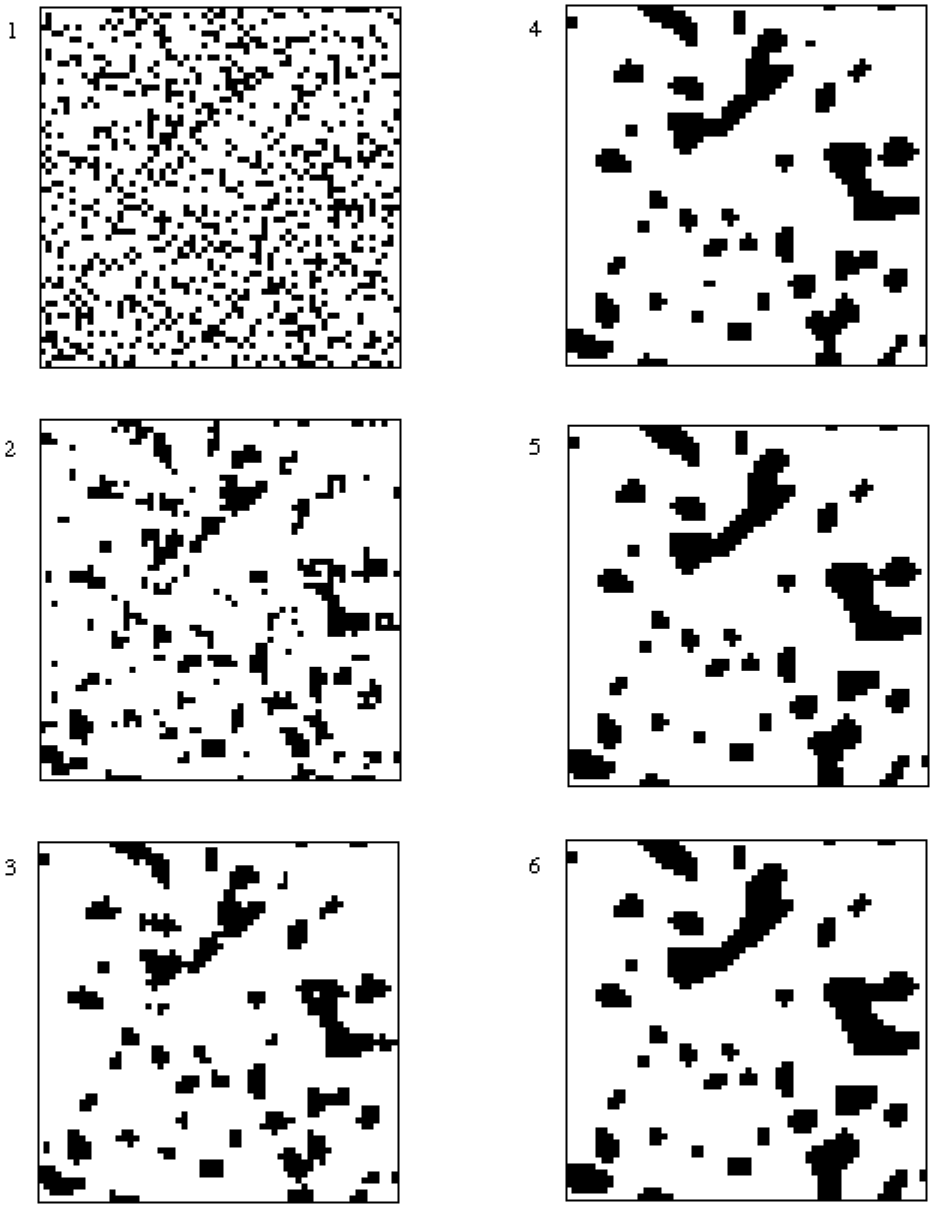


Figure 1.5.5 Simulation of the condensation CA described in Fig. 1.5.4. The initial conditions are chosen by setting randomly each site ON with a probability of 1 in 4. The initial few steps result in isolated ON sites disappearing and small ragged droplets of ON sites forming in higher-density regions. 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.

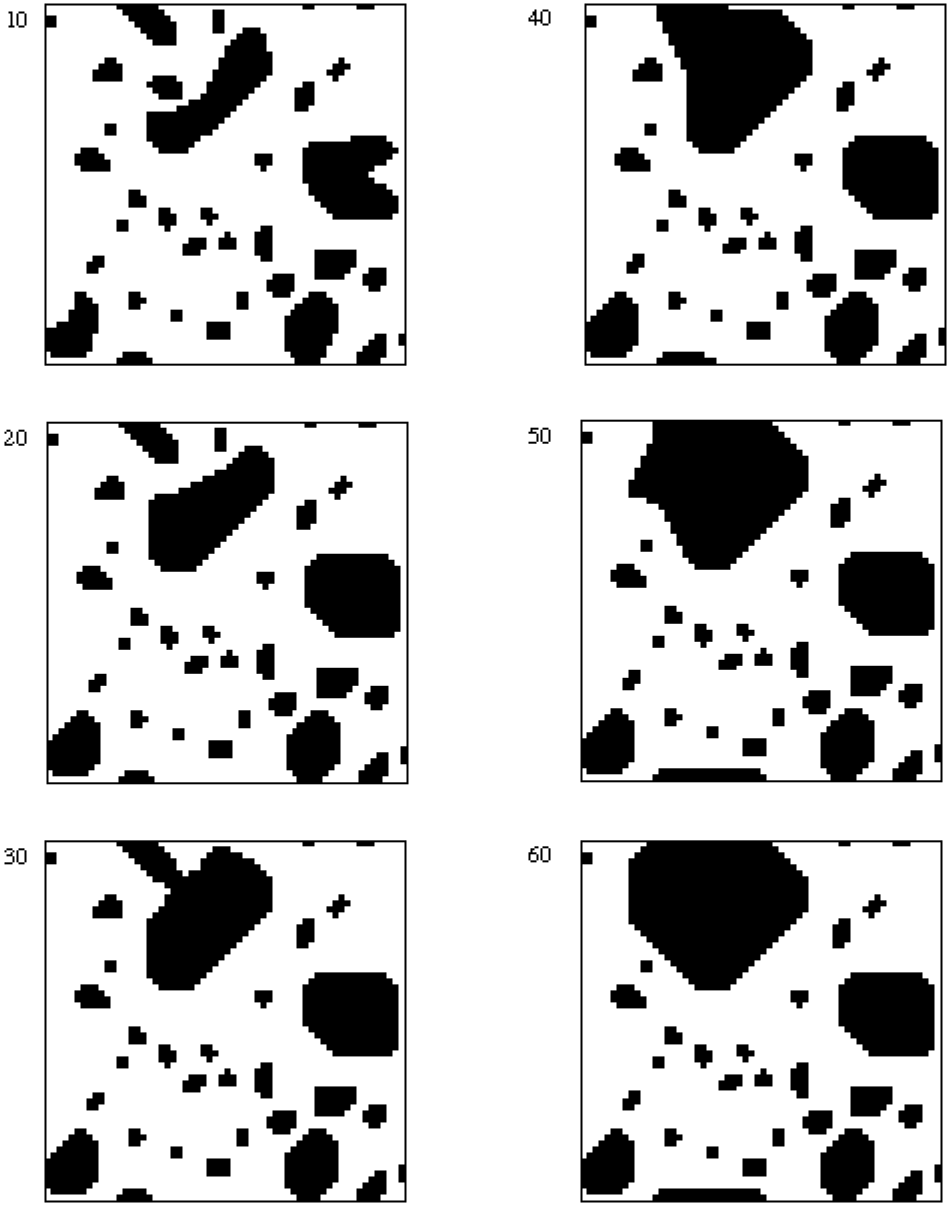


Figure 1.5.5 *Continued.* The initial occupation probability of 1 in 4 is near a phase transition in the kinetics of this model for a space of this size. For slightly higher densities the final configuration consists of a droplet covering the whole space. For slightly lower densities the final configuration is of isolated droplets. At a probability of 1 in 4 either may occur depending on the specific initial state. ■

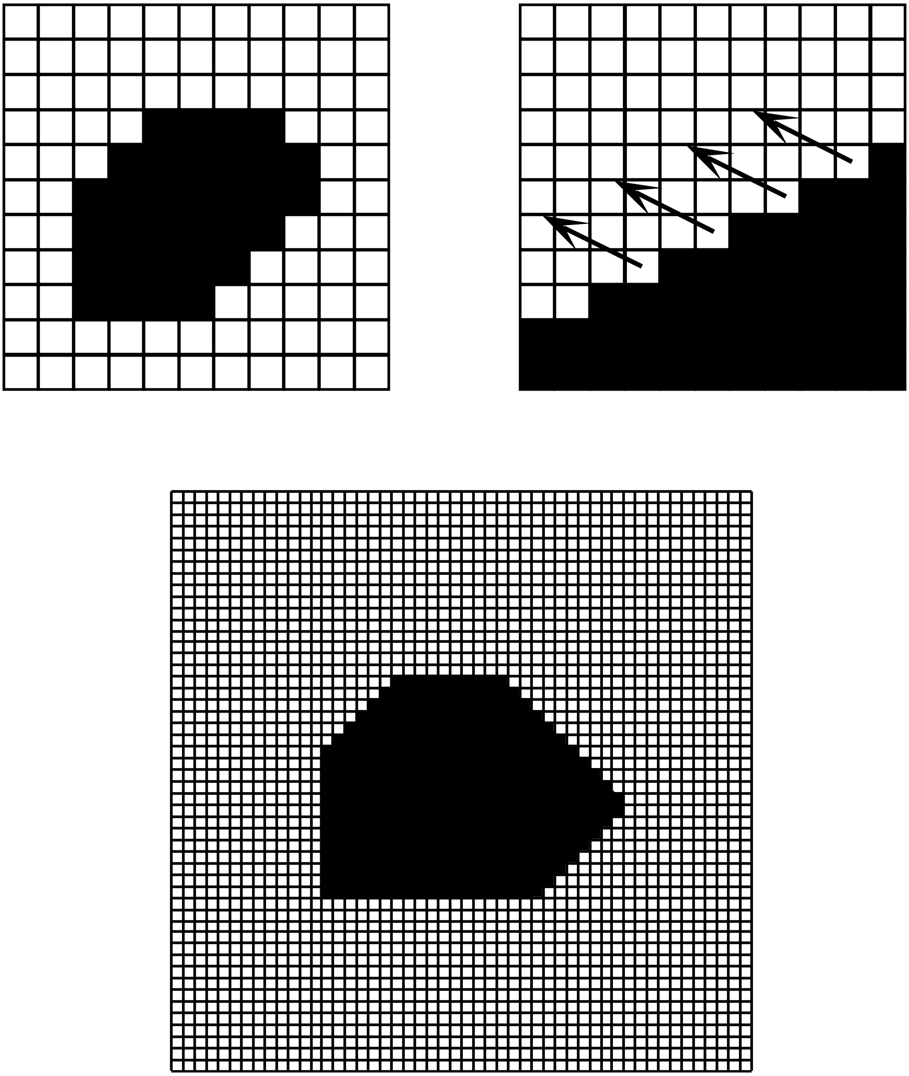


Figure 1.5.6 The droplet condensation model of Fig. 1.5.4 may be understood by noting that certain boundaries between condensed and uncondensed regions are stable. A completely stable shape is illustrated in the upper left. It is composed of boundaries that are horizontal, vertical or diagonal at 45°. 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. ∎

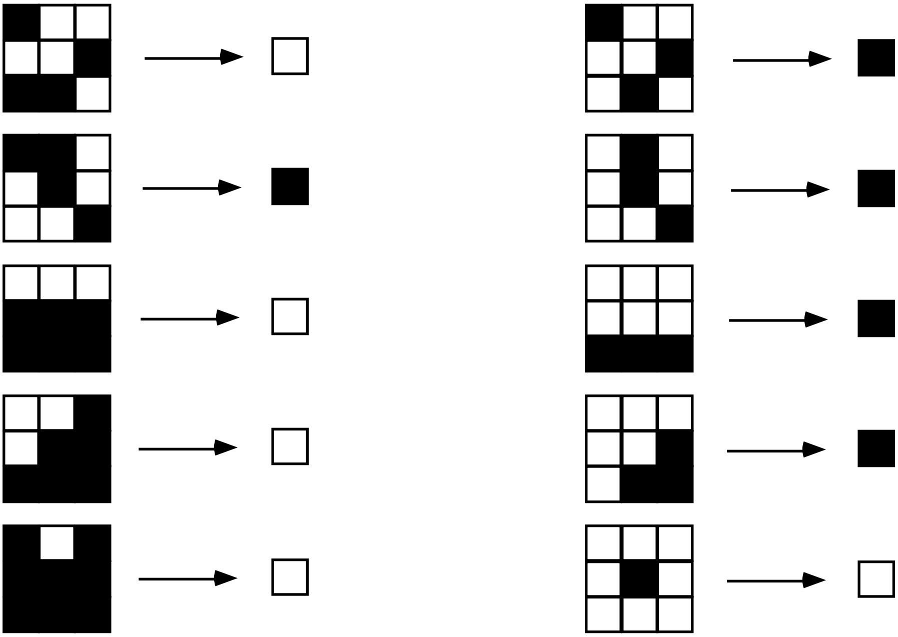


Figure 1.5.7 The CA rule Conway's Game of Life is illustrated for a few cases. When there are fewer than three or more than four neighbors in the 3 × 3 region the central cell is OFF in the next step. When there are three neighbors the central cell is ON in the next step. When there are four neighbors 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

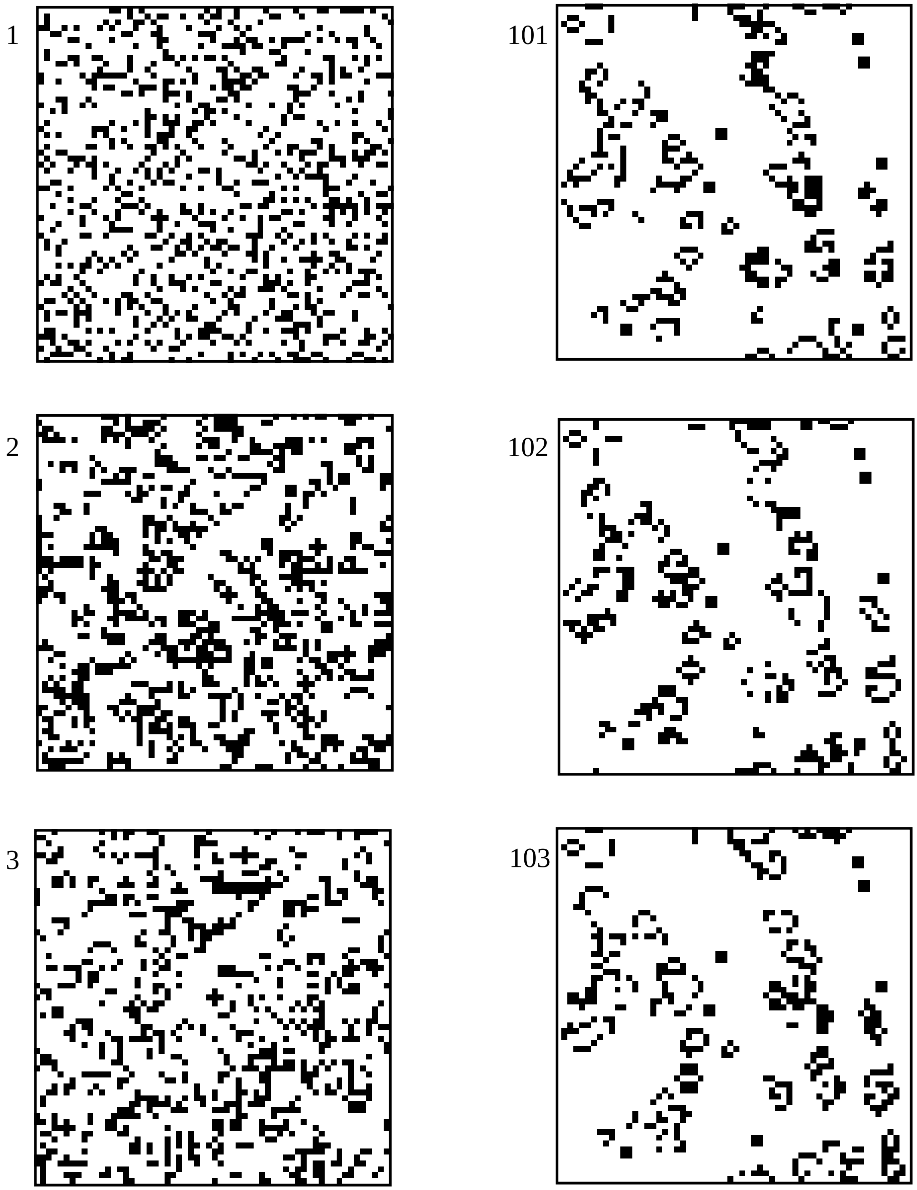


Figure 1.5.8 Simulation of Conway's Game of Life starting from the same initial conditions as used in Fig. 1.5.6 for the condensation rule where 1 in 4 cells are ON. Unlike the condensation rule there remains an active step-by-step evolution of the population of ON cells for many cycles. Illustrated are the three initial steps, and three successive steps each starting at steps 100, 200 and 300.

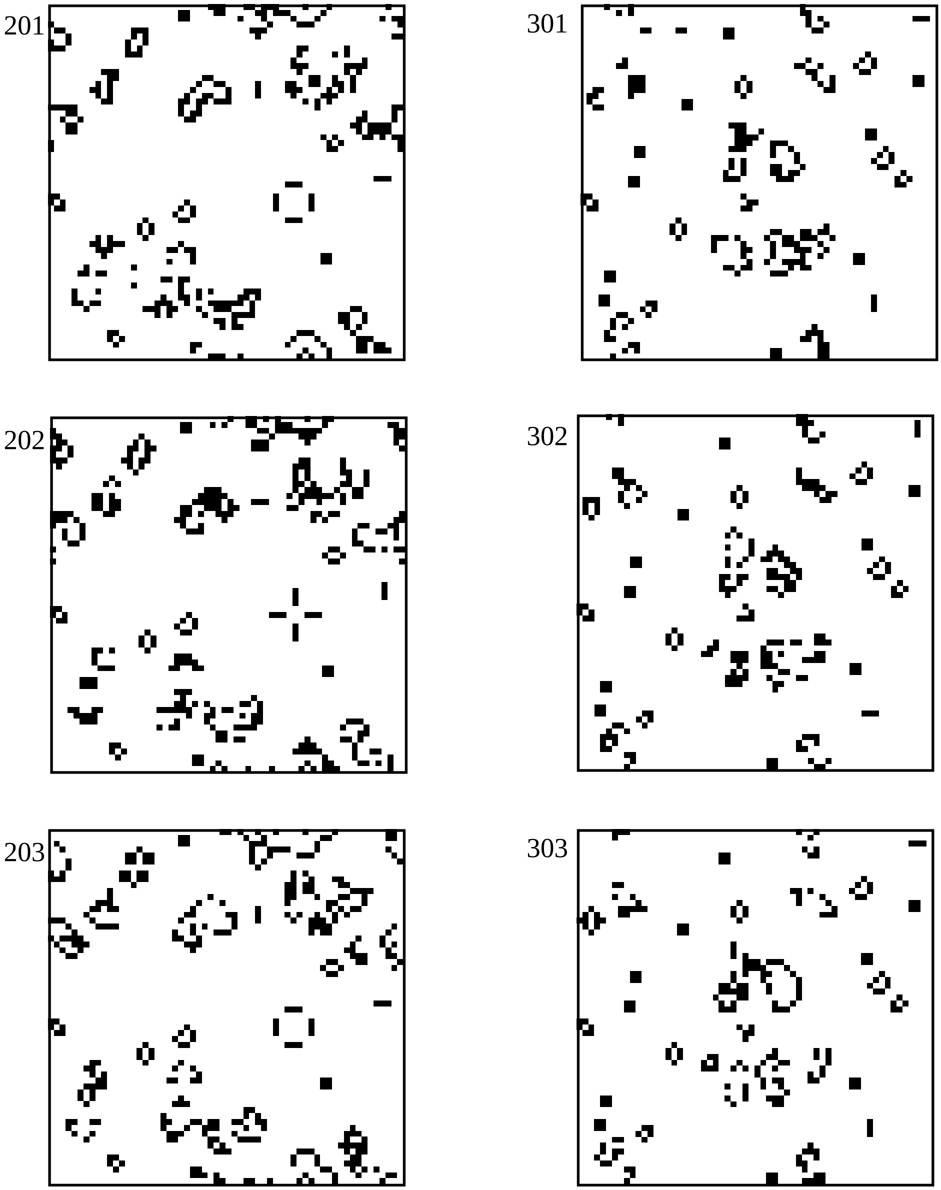


Figure 1.5.8 *Continued.* After the initial activity that occurs everywhere, the pattern of activity consists of regions that are active and regions that are static or have short cyclical activity. However, the active regions move over time around the whole space leading to changes everywhere. Eventually, 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. ▮

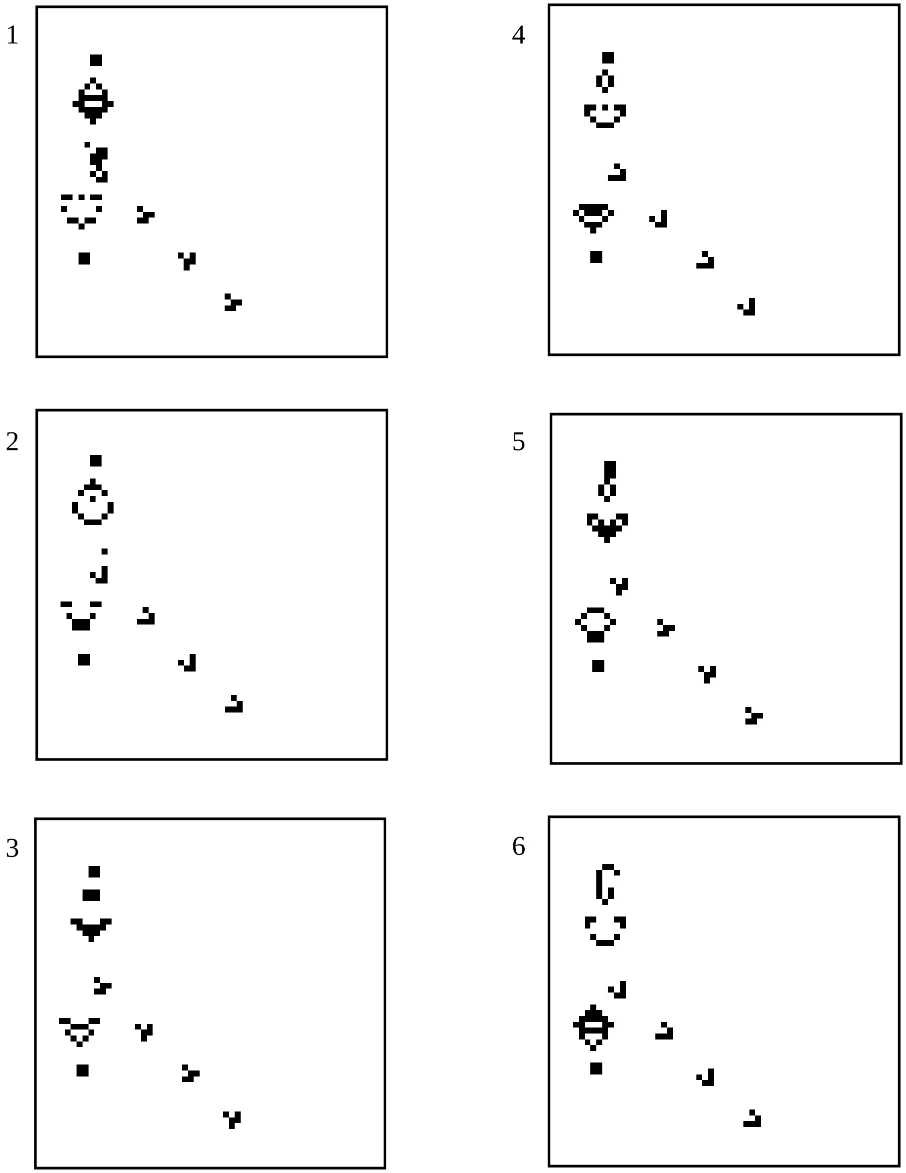


Figure 1.5.9 Special initial conditions simulated using Conway's Game of Life result in structures of ON cells called gliders that travel in space while progressing cyclically through a set of configurations. Several of the simplest type of gliders are shown moving toward the lower right. The more complex 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:

$$P(\{s(i,j,k;\ t)\}|\{s(i,j,k;\ t-1)\}) \qquad (1.5.19)$$

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.

$$P(\{s(i,j,k;t)\}\,|\,\{s(i,j,k;t-1)\}) = \prod_{i,j,k} P_0(s(i,j,k;t)\,|\,N(i,j,k;t-1)) \qquad (1.5.20)$$

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

$$P(\{s(i,j,k,t)\}) = \prod_{t}\ \prod_{i,j,k} P_0(s(i,j,k;t)\,|\,N(i,j,k;t-1))P(\{s(i,j,k;0)\}) \qquad (1.5.21)$$

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.

$$P(\{s(i,j,k,t)\}) = e^{-E(\{s(i,j,k,t)\})/kT} \qquad (1.5.22)$$

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:

$$E(\{s(i,j,k,t)\}) = kT\ln(P(\{s(i,j,k,t)\})) \qquad (1.5.23)$$

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

$$s(i, j, k; t) = \mathrm{mod}_2(R(N(i, j, k; t-1)) + s(i, j, k; t-2)) \qquad (1.5.24)$$

The inverse of the update rule is immediately constructed using the properties of addition modulo 2 (Eq. (1.5.8)) as:

$$s(i, j, k; t-2) = \mathrm{mod}_2(R(N(i, j, k; t-1)) + s(i, j, k; t)) \qquad (1.5.25)$$

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-

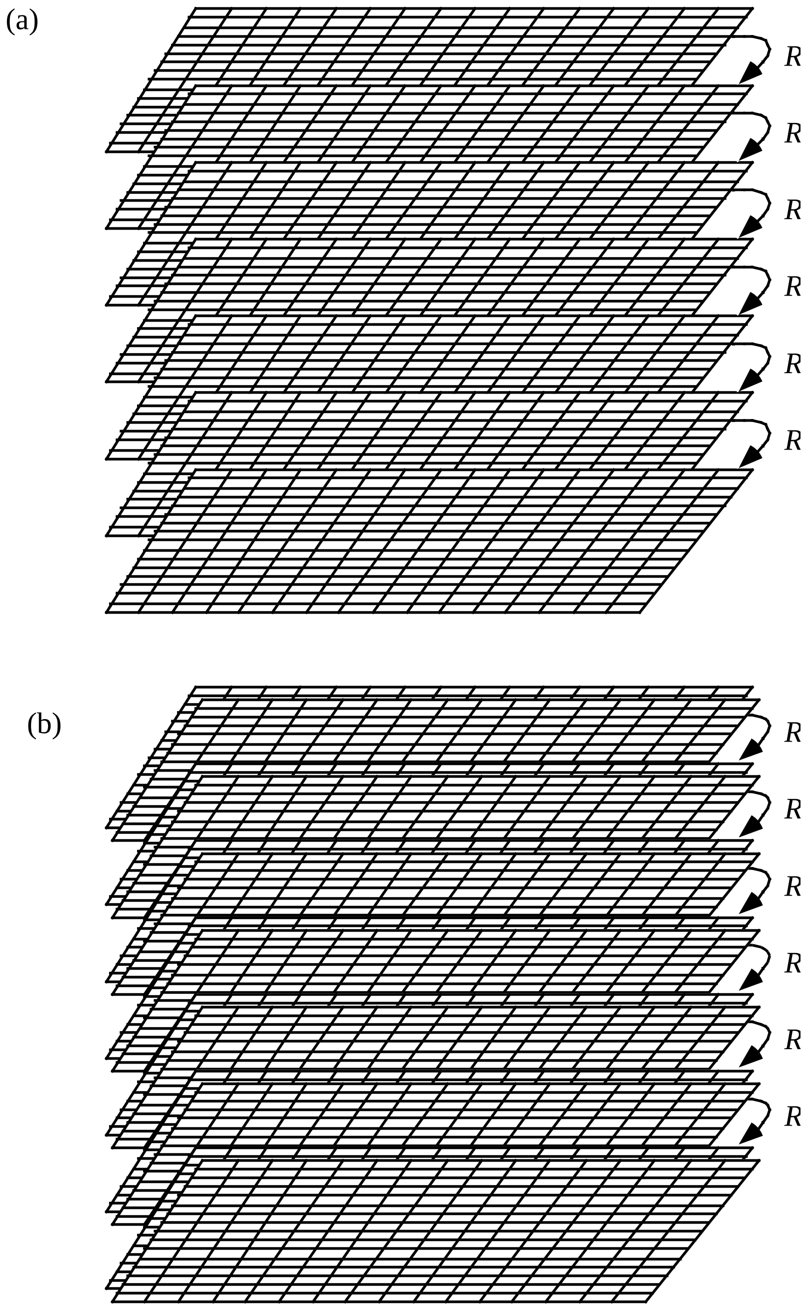
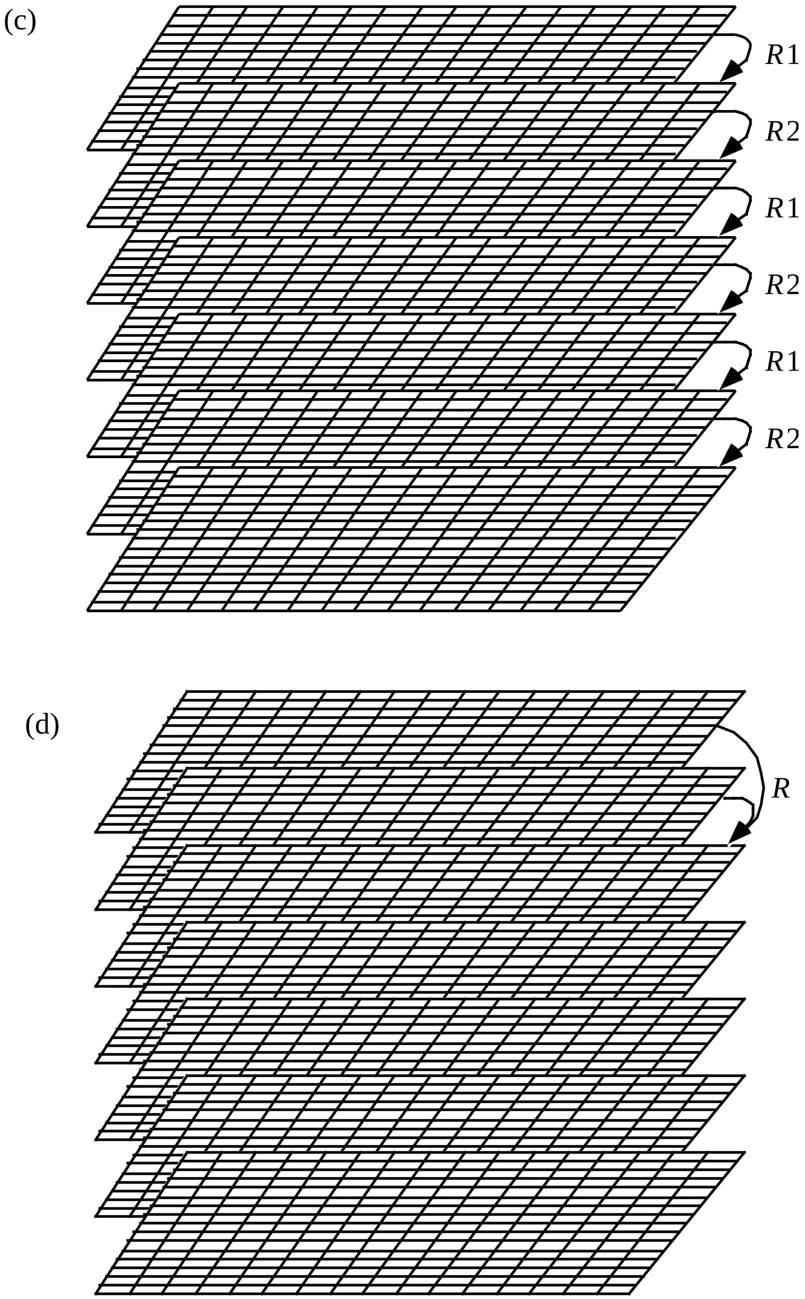
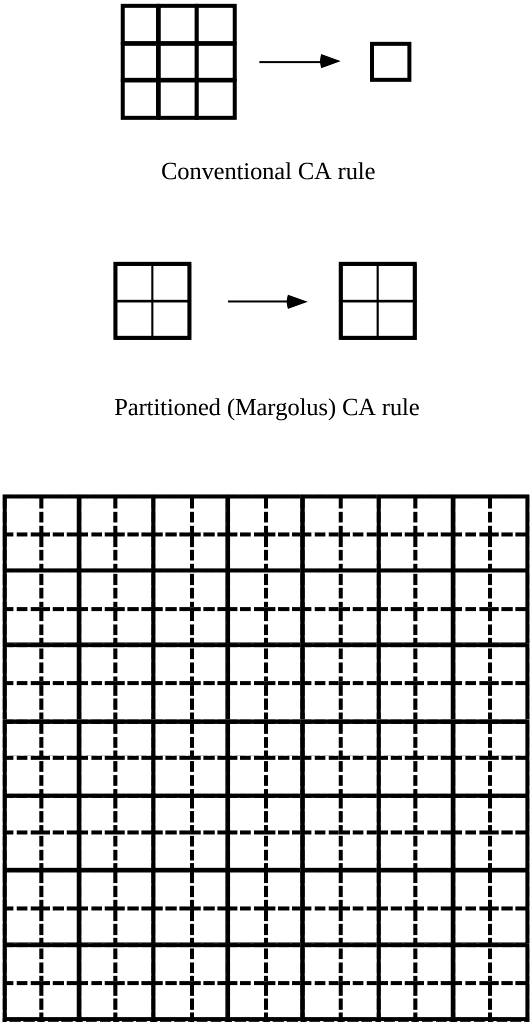


Figure 1.5.10 Schematic illustrations of several modifications of the simplest CA rule. The basic CA rule updates a set of spatially arrayed cell variables shown in (a). The first modification uses more than one variable in each cell. Conceptually this may be thought of as describing a set of coupled 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

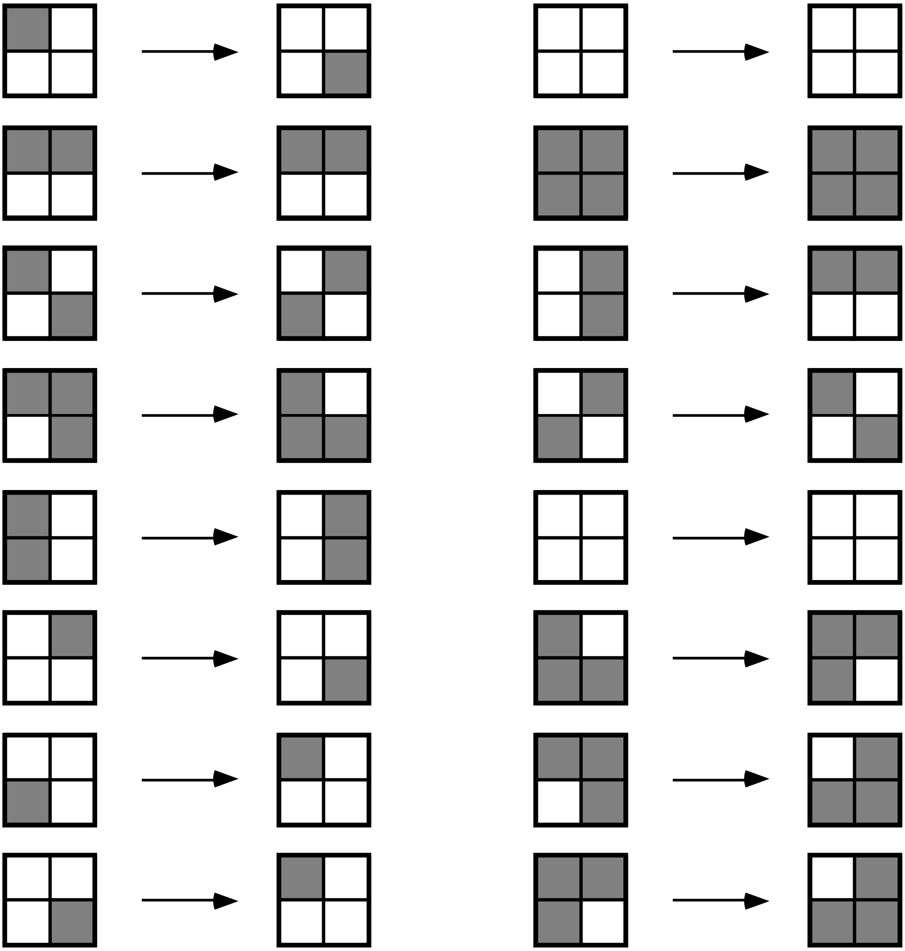


Figure 1.5.12 Illustration of a particular 2-d Margolus rule that preserves the number of ON cells which may be thought of as particles in a gas. The requirement for conservation of number of particles is that every initial configuration is matched with a final configuration having the same number of 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): $[A,0] \leftrightarrow [0,A]$, $[B,0] \leftrightarrow [0,B]$, $[C,A] \leftrightarrow [A,C]$,and $[C,B] \leftrightarrow [B,C]$.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: $[C,0] \leftrightarrow [A,B]$,$[0,C] \leftrightarrow [B,A]$,$[C,0] \leftrightarrow [B,A]$ and $[0,C] \leftrightarrow [A,B]$. 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-

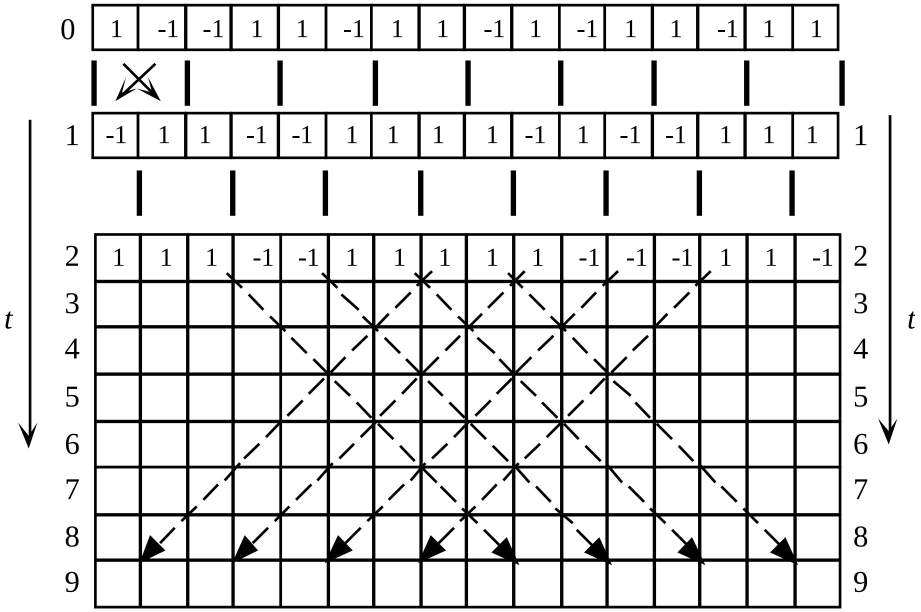


Figure 1.5.13 A simple 1-d CA using a Margolus rule, which switches the values of the two adjacent cells in the partition, can be used to model the wave equation. The partitions alternate between the two possible ways of partitioning the cells every time step. It can be seen that the initial state is 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*

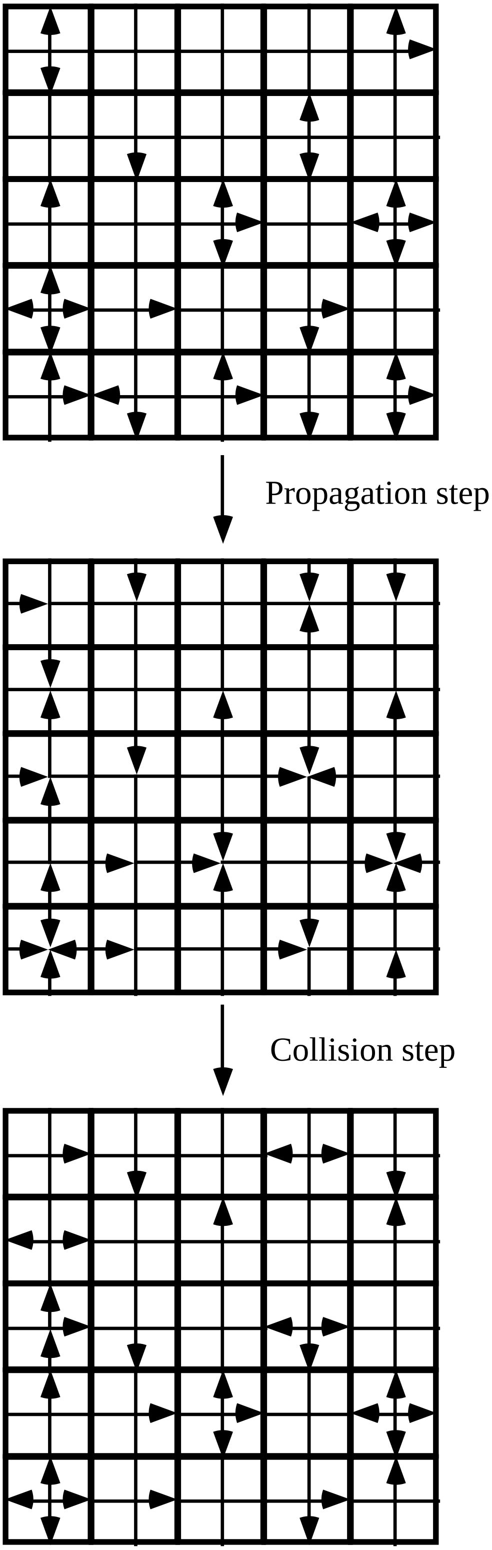
One of the natural physical systems to model using CA is the problem of layer-by-layer material growth such as is achieved in molecular beam epitaxy. There are many areas of study of the growth of materials. For example,in cases where the material is formed of only a single type of atom,it is the surface 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. ∎



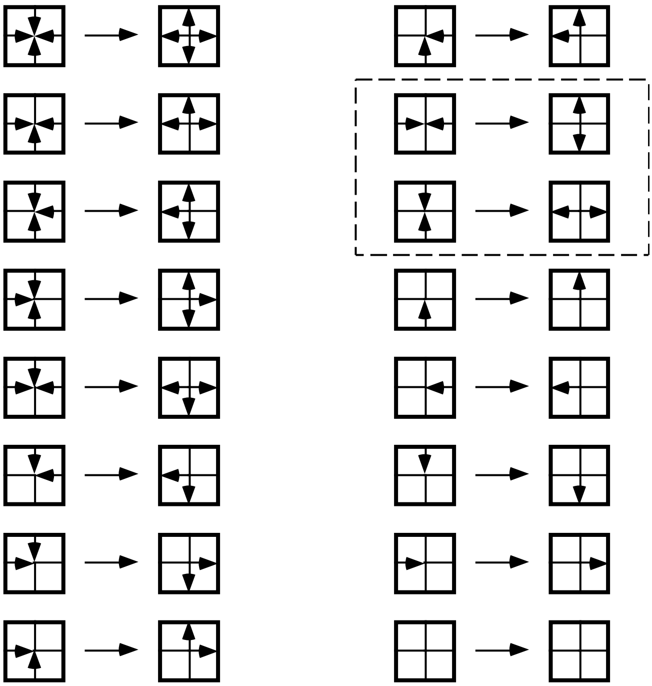


Figure 1.5.15 The collision rule for the HPP lattice gas. With the exception of the case of two particles coming in from N and S and leaving from E and W, or vice versa (dashed box), there are no changes in the particle velocities as a result of collisions in this rule. Momentum conservation does not allow any other changes. ∎

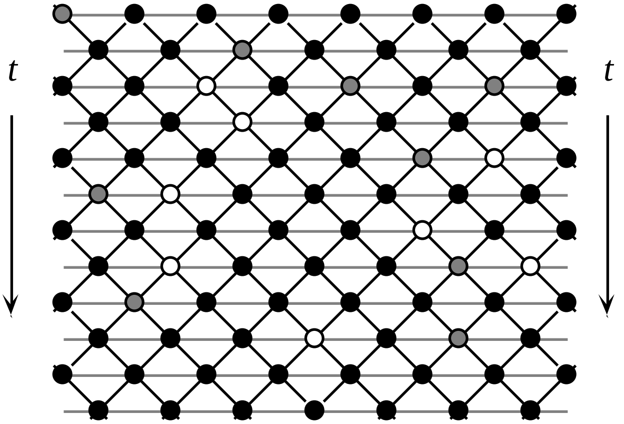


Figure 1.5.16 Illustration of the time history of a CA and its use to model the structure of a material (alloy) formed by a layer by layer growth. Each horizontal dashed line represents a layer of the material. The alloy has three types of atoms. The configuration of atoms in each layer depends only 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, $s_i(t) = \pm1,0$. 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 $(\mathrm{GaAs})_{1-x}\mathrm{Si}_x$ 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:

$$E[\{s_i\}] = \sum_i e_i(s_i) \qquad (1.6.1)$$

Where $e_i(s_i)$ is the energy of the ith spin that does not depend on the values of any of the other spins. Since s_i are binary we can write this as:

$$E[\{s_i\}] = \frac{1}{2}\sum_i (e_i(1) - e_i(-1))s_i + (e_i(1) + e_i(-1)) = E_0 - \sum_i h_i s_i \rightarrow -\sum_i h_i s_i \qquad (1.6.2)$$

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:

$$P[\{s_i\}] = \frac{e^{-\beta E[\{s_i\}]}}{Z} = \frac{e^{\beta \sum_i h_i s_i}}{Z} = \frac{\prod_i e^{\beta h_i s_i}}{Z} \qquad (1.6.3)$$

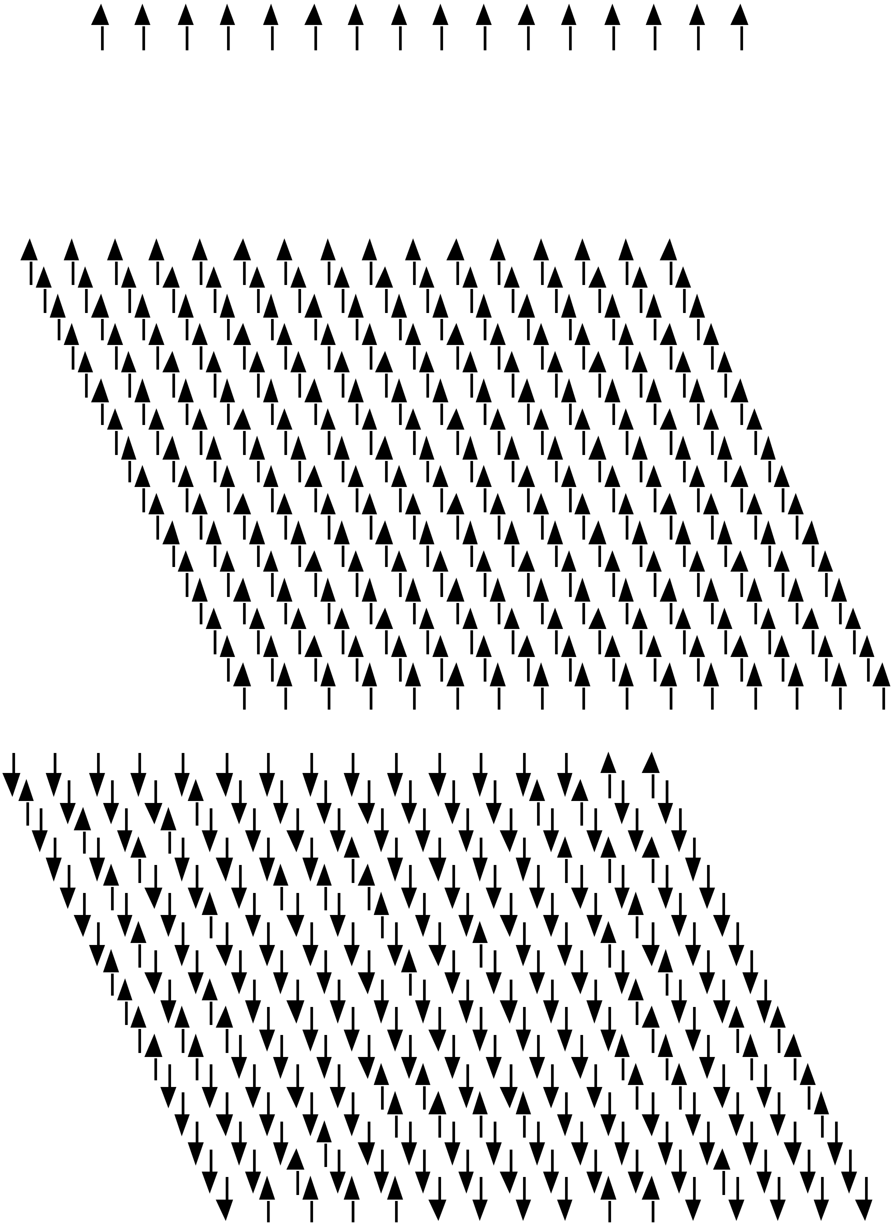


Figure 1.6.1 One way to visualize the Ising model is as a spatial array of binary variables called spins, represented as UP or DOWN arrows. A one-dimensional (1-d) example with all spins UP is shown on top. The middle and lower figures show two-dimensional (2-d) arrays which have all spins UP (middle) or have some spins UP and some spins DOWN (bottom). ∎

where $\beta = 1/kT$. The partition function Z is given by:

$$Z = \sum_{\{s_i\}} e^{-\beta E[\{s_i\}]} = \sum_{\{s_i\}} \prod_i e^{\beta h_i s_i} = \prod_i \sum_{s_i} e^{\beta h_i s_i} = \prod_i \left(e^{\beta h_i} + e^{-\beta h_i} \right) \qquad (1.6.4)$$

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:

$$P[\{s_i\}] = \prod_i P(s_i) = \prod_i \left(\frac{e^{\beta h_i s_i}}{e^{\beta h_i} + e^{-\beta h_i}} \right) \qquad (1.6.5)$$

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):

$$F = -kT\ln(Z) = -kT \sum_i \ln\left(e^{\beta h_i} + e^{-\beta h_i} \right) = -kT \sum_i \ln\left(2\cosh\left(\beta h_i\right)\right) \qquad (1.6.6)$$

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

$$U = -\frac{\partial \ln(Z)}{\partial \beta} = -\sum_i \frac{h_i(e^{\beta h_i} - e^{-\beta h_i})}{(e^{\beta h_i} + e^{-\beta h_i})} = -\sum_i h_i \tanh(\beta h_i) \qquad (1.6.7)$$

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:

$$\begin{aligned} U &= \left\langle E[\{s_i\}] \right\rangle = \left\langle -\sum_i h_i s_i \right\rangle = -\sum_i h_i \left\langle s_i \right\rangle = -\sum_i h_i \sum_{s_i} s_i P(s_i) \\ &= -\sum_i h_i \frac{(e^{\beta h_i} - e^{-\beta h_i})}{(e^{\beta h_i} + e^{-\beta h_i})} = -\sum_i h_i \tanh(\beta h_i) \end{aligned} \qquad (1.6.8)$$

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:

$$m_i = \left\langle s_i \right\rangle = \sum_{s_i = \pm 1} s_i P_{s_i}(s_i) = P_{s_i}(1) - P_{s_i}(-1) \qquad (1.6.9)$$

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)

$$S = k\beta U + k \ln Z = -k\sum_i \beta h_i \tanh(\beta h_i) + k\sum_i \ln\big(2\cosh(\beta h_i)\big) \tag{1.6.12}$$

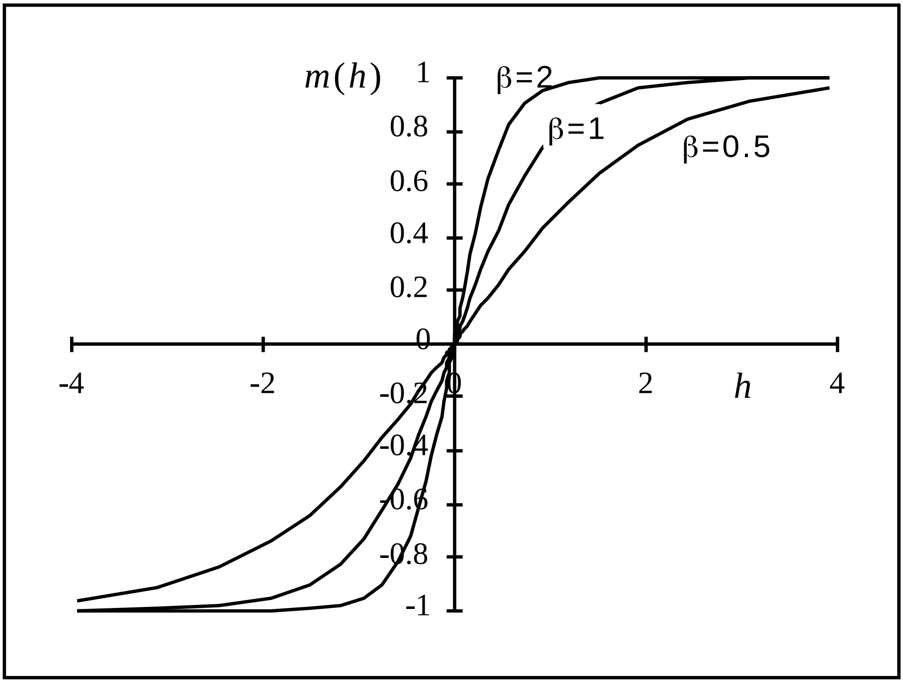


Figure 1.6.2 Plot of the magnetization at a particular site as a function of the magnetic field for independent spins in a magnetic field. The magnetization is the average of the spin value, so the magnetization shows the degree to which the spin is aligned to the magnetic field. The different curves are for several temperatures $\beta = 0.5,1,2$ ($\beta = 1/kT$). 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:

$$S = +k\left[N\ln(2) - \frac{1}{2}\sum_i \left((1+m_i)\ln\left(1+m_i\right) + (1-m_i)\ln\left(1-m_i\right)\right)\right] \tag{1.6.16}$$

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:

$$S = k\ln(\Omega\,(E,N)) = k[N(\ln N - 1) - N(1)(\ln N(1) - 1) - N(-1)(\ln N(-1) - 1]$$
$$= k[N\ln N - N(1)\ln N(1) - N(-1)\ln N(-1)] \qquad (1.6.20)$$

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:

$$\begin{aligned} P_{s_i}(1) &= (1 + m)\,/\,2 \\ P_{s_i}(-1) &= (1 - m)\,/\,2 \end{aligned} \qquad (1.6.21)$$

Since there are many spins in the system, we can obtain the number of UP spins using

$$\begin{aligned} N(1) &= NP_{s_i}(1) = N(1 + m)\,/\,2 \\ N(-1) &= NP_{s_i}(1) = N(1 - m)\,/\,2 \end{aligned} \qquad (1.6.22)$$

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:

$$\begin{aligned} e(s,s') = e(1,1)\frac{(1+s)(1+s')}{4} + e(1,-1)\frac{(1+s)(1-s')}{4} \\ + e(1,-1)\frac{(1-s)(1+s')}{4} + e(-1,-1)\frac{(1-s)(1-s')}{4} \end{aligned} \tag{1.6.28}$$

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:

$$U(m) = < -J\sum_{<ij>} s_i s_j > = -\frac{1}{2}NJzm^2 \tag{1.6.35}$$

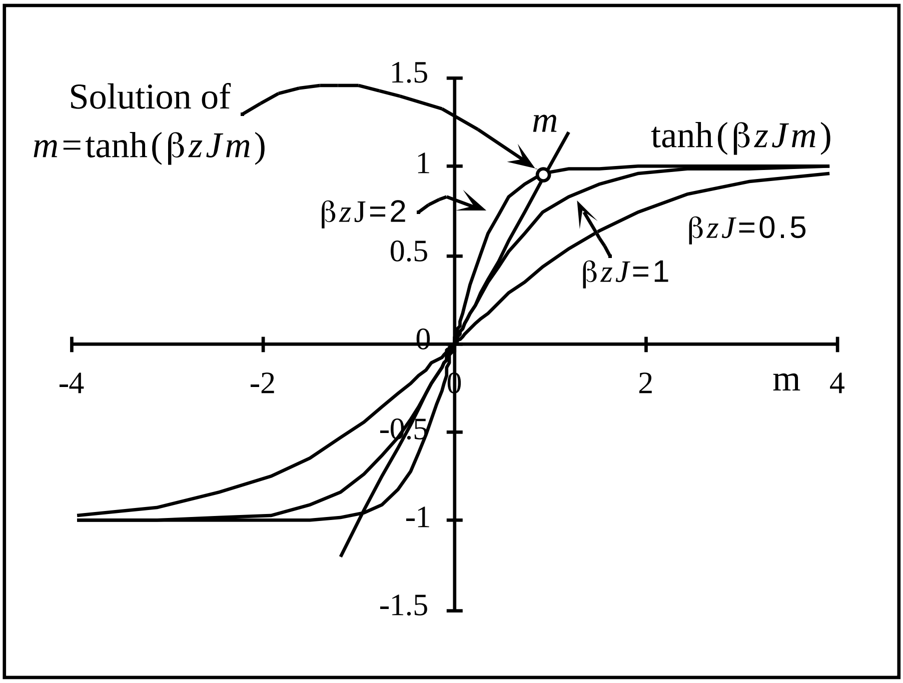


Figure 1.6.3 Graphical solution of Eq. (1.6.32) $m = \tanh(\beta zJm)$ by plotting both the left- and right-hand sides of the equation as a function of m and looking for the intersections. $m = 0$ is always a solution. To consider other possible solutions we note that both functions are antisymmetric 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-

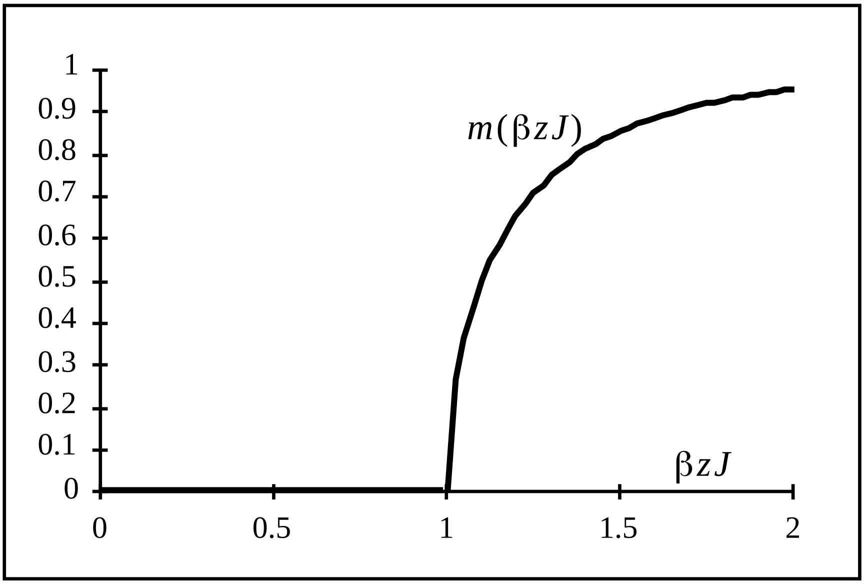


Figure 1.6.4 The mean field approximation solution of the Ising model gives the magnetization (average value of the spin) as a solution of Eq. (1.6.32). The solution is shown as a function of βzJ. As discussed in Fig. 1.6.3 and the text for $\beta zJ > 1$ there are three solutions. Only the positive 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 $h_i \neq 0$.

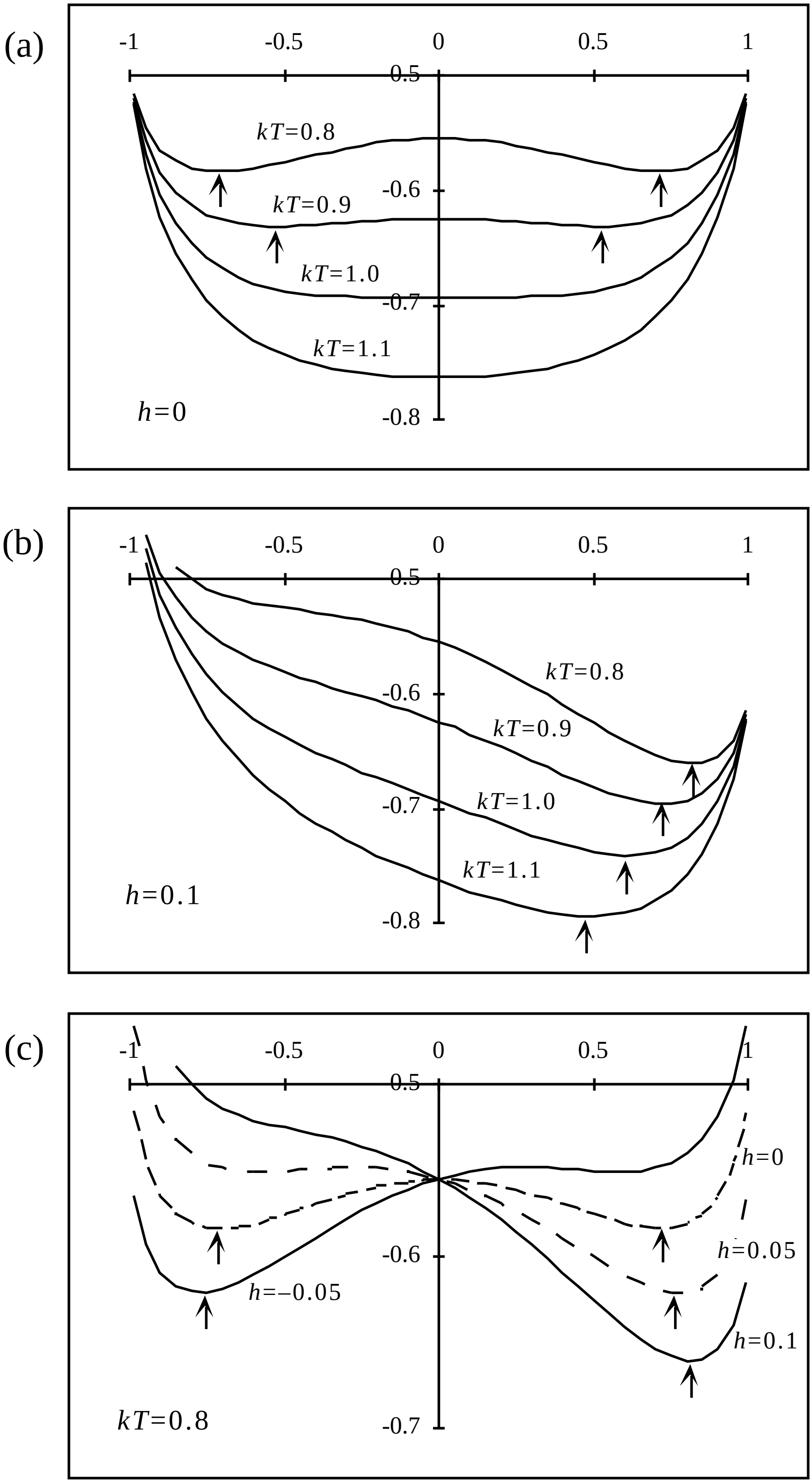
Question 1.6.3 Show that the minima of the free energy are the solutions of Eq.(1.6.32). This shows that our derivation is internally consistent. Specifically, that our two ways of defining the mean field approximation, first using Eq. (1.6.29) and then using Eq. (1.6.34), are compatible.

Solution 1.6.3 Taking the derivative of Eq. (1.6.35) with respect to m and setting it to zero gives:

$$0 = -Jzm - kT\left[-\frac{1}{2}\left(\ln\left(1+m\right) - \ln\left(1-m\right)\right)\right] \qquad (1.6.37)$$

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. ∎

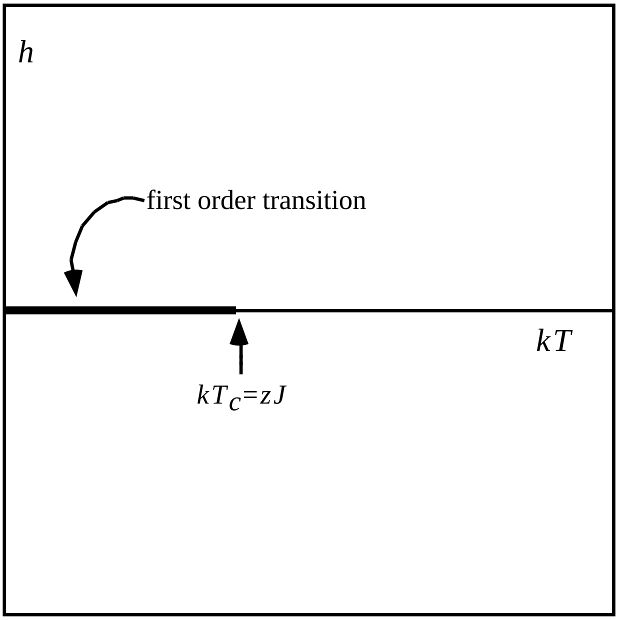
value of the magnetization will so dominate other magnetizations that any other possibility is negligible. This is reasonable except for the case when the magnetic field is close to zero, below T_c, and we have two equally likely magnetizations. In this case,the usual justification does not hold, though 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. ▮



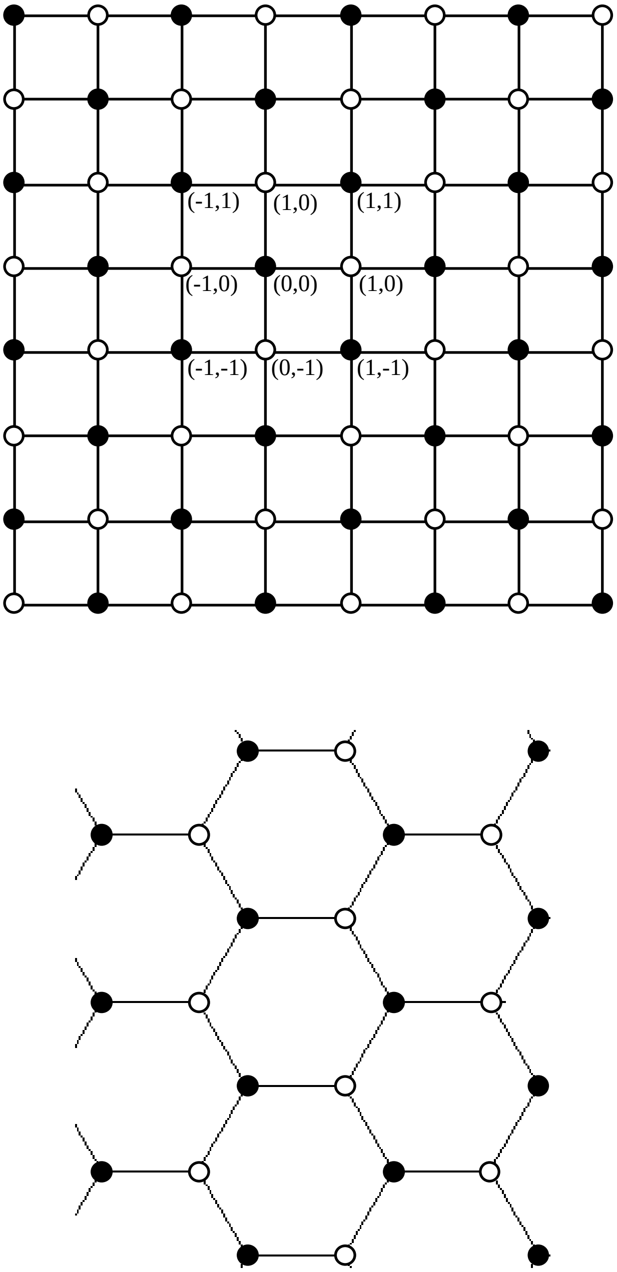


Figure 1.6.7 In order to obtain mean field equations for the antiferromagnetic case $J < 0$ we consider a square lattice (top) and label every site according to the sum of its rectilinear indices as odd (open circles) or even (filled circles). A few sites are shown with indices. Each site is understood to be the location of a spin. We then invert the spins (redefine them by $s \rightarrow -s$) 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.

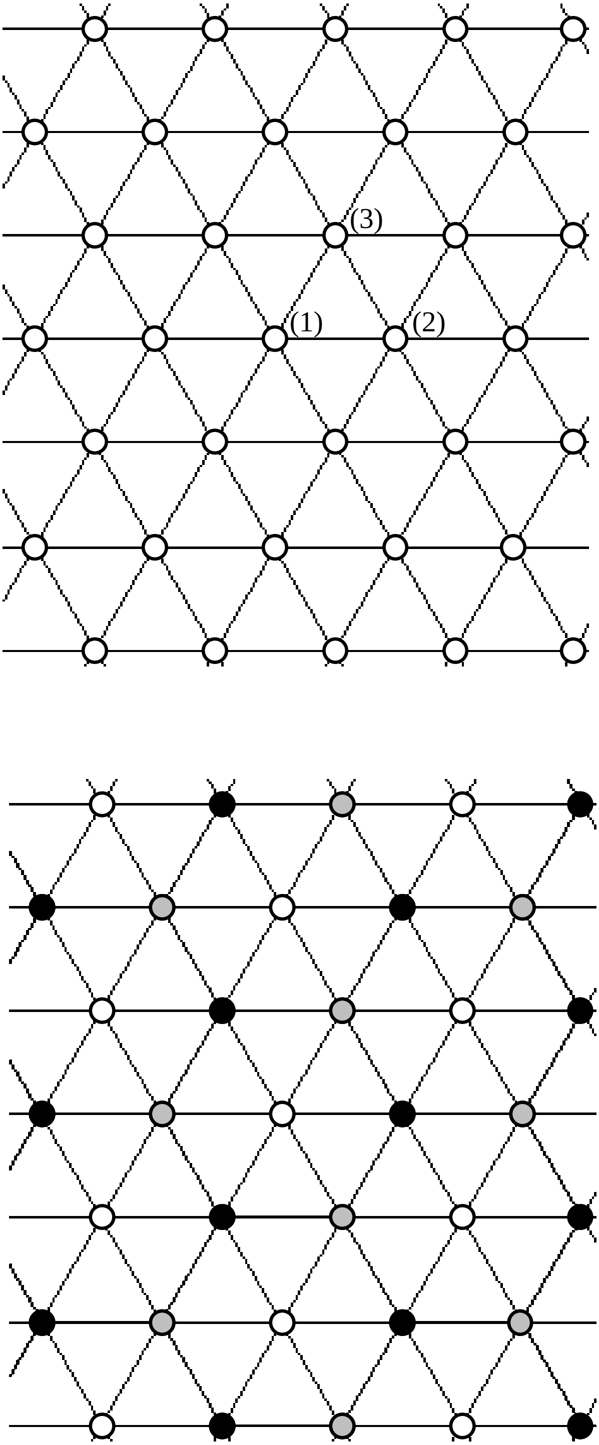


Figure 1.6.8 A triangular lattice (top) is not a bipartite lattice. In this case we cannot solve the antiferromagnet J < 0 by the same method as used for the square lattice (see Fig. 1.6.7). If we try to assign magnetizations to different sites we find that assigning a magnetization to site (1) would 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:

$$\begin{aligned} F(m) &= NJm^2 - \frac{1}{3}NkT\ln(2) \\ &\quad -\frac{2}{3}NkT\left[\ln(2) - \frac{1}{2}\left((1+m)\ln\left(1+m\right) + (1-m)\ln\left(1-m\right)\right)\right] \end{aligned} \tag{1.6.44}$$

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 the 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:

$$\begin{gathered} < s_i > \\ < s_i s_j > - < s_i > < s_j > \\ < s_i s_j s_k > - < s_i s_j > < s_k > - < s_i s_k > < s_j > - < s_j s_k > < s_i > +2 < s_i > < s_j > < s_k > \end{gathered} \tag{1.6.48}$$

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. $P_{s_is_j}(1,-1) = 1$, 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:

$$< s_i s_j > - < s_i > < s_j > \ = < (s_i - < s_i >) \ (s_j - < s_j >) > \tag{1.6.49}$$

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.

$$E[\{s_i\}] = -\sum_i h_i s_i - \frac{1}{2} J \sum_{r_{ij} < r_0} s_i s_j \qquad (1.6.51)$$

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:

$$E[\{s_i\}] = -\sum_i h_i s_i - \frac{1}{2} \sum_{i,j} J(r_{ij}) s_i s_j \qquad (1.6.52)$$

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 $J \to J/N$. The energy is given by:

$$E[\{s_i\}] = -h\sum_i s_i - \frac{1}{2N} J \sum_{i,j} s_i s_j \qquad (1.6.53)$$

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

$$m = \frac{1}{N}\sum_i s_i \qquad (1.6.54)$$

in terms of which the energy becomes:

$$E(\{s_i\}) = hNm - \frac{1}{2} JNm^2 \qquad (1.6.55)$$

This is the same as the mean field Eq. (1.6.39) with the substitution $Jz \to J$. 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:

$$E_{+i}(\{s_j\}_{j \neq i}) = E(s_i = +1,\{s_j\}_{j \neq i}) - E(s_i = -1,\{s_j\}_{j \neq i}) \qquad (1.6.61)$$

The probability of the spin being UP or DOWN is given by Eq. (1.4.14) as:

$$P_{s_i}(1) = \frac{1}{1+e^{E_{+i}/kT}} = f(E_{+i}) \qquad (1.6.62)$$

$$P_{s_i}(-1) = 1 - f(E_{+i}) = f(-E_{+i}) \qquad (1.6.63)$$

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 ηNP(1;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

$$\eta = (R(1|-1) + R(-1|1))\Delta t = \Delta t/\tau \qquad (1.6.65)$$

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:

$$\Delta t = \frac{1}{N(R(1|-1)+R(-1|1))} = \frac{\tau}{N} \qquad (1.6.66)$$

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:

$$\sum_k \frac{N_k \Delta t}{\tau_k} \rightarrow N \frac{\Delta t}{\tau} \qquad (1.6.67)$$

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

$$1/\tau = (R(1|-1) + R(-1|1)) = \nu(e^{-(E_B - E_1)/kT} + e^{-(E_B - E_{-1})/kT}) \qquad (1.6.68)$$

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

$$< s_i(t')s_i(t) > - < s_i >^2 \quad e^{-|t-t'|/\tau_s} \qquad (1.6.69)$$

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:

$$\begin{aligned} < s_i(t')s_i(t) > &= P_{s_i(t'),s_i(t)}(1,1) + P_{s_i(t'),s_i(t)}(-1,-1) \\ &\quad - P_{s_i(t'),s_i(t)}(1,-1) - P_{s_i(t'),s_i(t)}(-1,1) \end{aligned} \qquad (1.6.70)$$

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'.

$$P_{s_i(t'),s_i(t)}(s_2,s_1) = P_{s_i(t'),s_i(t)}(s_2 | s_1) P_{s_i(t)}(s_1) \qquad (1.6.71)$$

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 $P_{s_i(t'),s_i(t)}(1$ç$s_1) + P_{s_i(t'),s_i(t)}(-1$ç$s_1) = 1$, so we can simplify Eq. (1.6.70) to:

$$< s_i(t')s_i(t) > = (2P_{s_i(t'),s_i(t)}(1|1) - 1)P(1) + (2P_{s_i(t'),s_i(t)}(-1|-1) - 1)P(-1) \qquad (1.6.72)$$

The evolution of the probabilities are described by Eq.(1.4.45), repeated here:

$$P(1;t) = (P(1;0) - P(1;\infty))e^{-t/\tau} + P(1;\infty) \qquad (1.6.73)$$

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:

$$\begin{aligned} P_{s(t'),s(t)}(1|1) &= (1 - P(1))e^{-(t'-t)/\tau} + P(1) \\ P_{s(t'),s(t)}(-1|-1) &= (1 - P(-1))e^{-(t'-t)/\tau} + P(-1) \end{aligned} \qquad (1.6.74)$$

Inserting these into Eq. (1.6.72) gives:

$$\begin{aligned} < s_i(t')s_i(t) > &= (2\left[(1-P(1))e^{-(t'-t)/\tau} + P(1)\right] - 1)P(1) \\ &\quad + (2\left[(1-P(-1))e^{-(t'-t)/\tau} + P(-1)\right] - 1)P(-1) \\ &= 4P(1)P(-1)e^{-(t'-t)/\tau} + (P(1) - P(-1))^2 \end{aligned} \qquad (1.6.75)$$

The constant term on the right is the same as the square of the average of the spin:

$$<s_i(t)>^2 = (P(1) - P(-1))^2 \qquad (1.6.76)$$

Inserting into Eq.(1.6.69) leads to the desired result (we have assumed that $t' > t$):

$$<s_i(t')s_i(t)> - <s_i(t)>^2 = 4P(1)P(-1)e^{-(t'-t)/\tau} \quad e^{-(t'-t)/\tau} \qquad (1.6.77)$$ ▮

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*

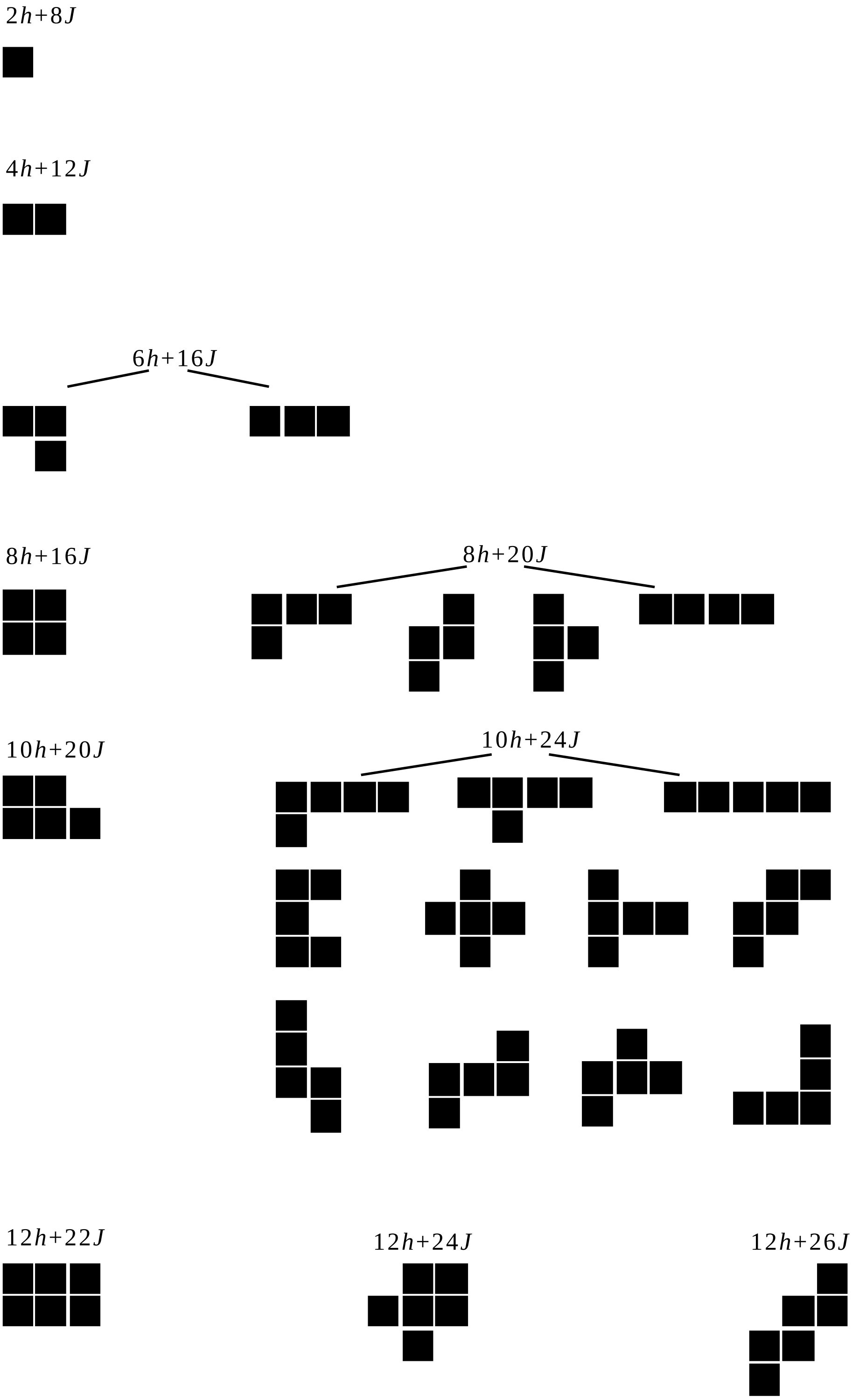
In this section we discuss the first-order transition kinetics in the Ising model. Similar arguments apply to other first-order transitions like the freezing or boiling of water. If we start with an Ising model in equilibrium at a temperature $T < T_c$ and a small positive magnetic field $h << zJ$, the 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 $fJN_c^{(d-1)/d}$, 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

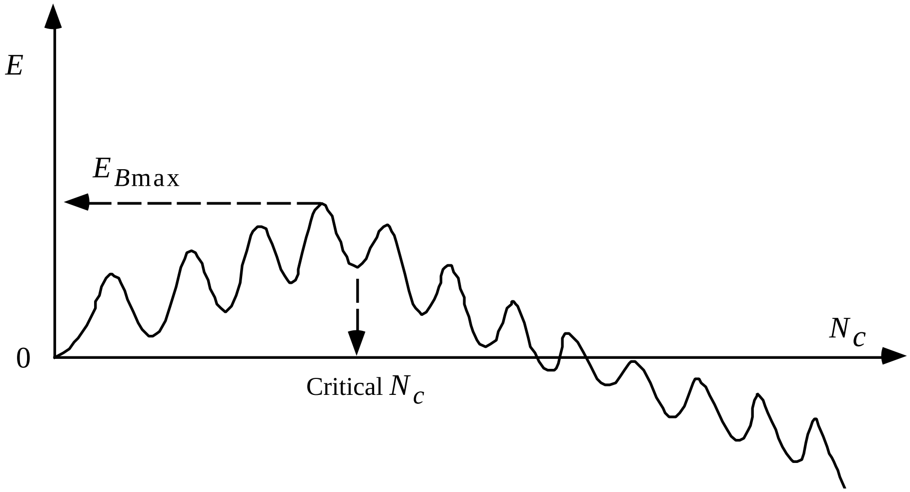


Figure 1.6.10 Schematic illustration of the energies that control the kinetics of a first-order phase transition. The horizontal axis is the size of a cluster of DOWN spins N_c that are the equilibrium phase. The cluster is in a background of UP spins that are the metastable phase. The vertical axis 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

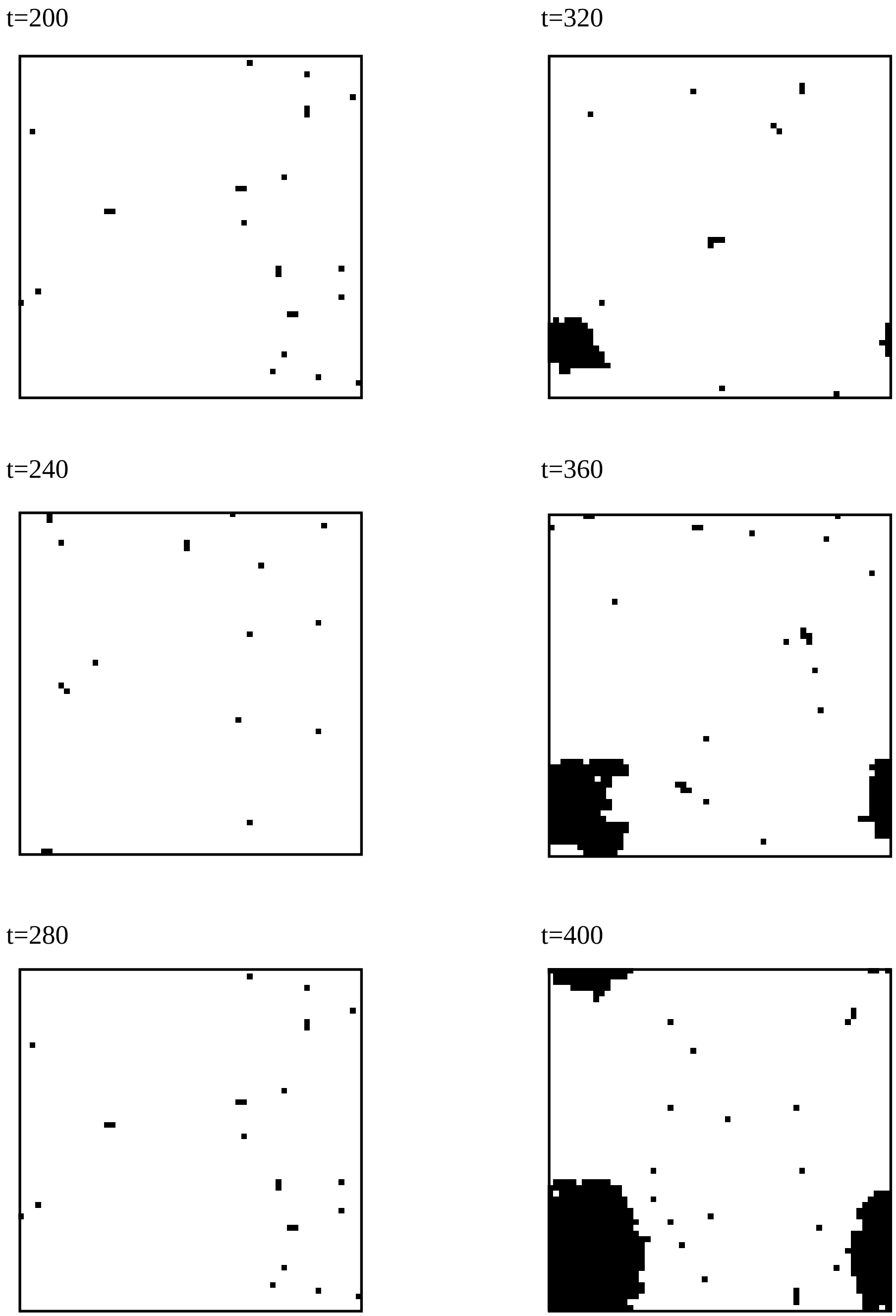
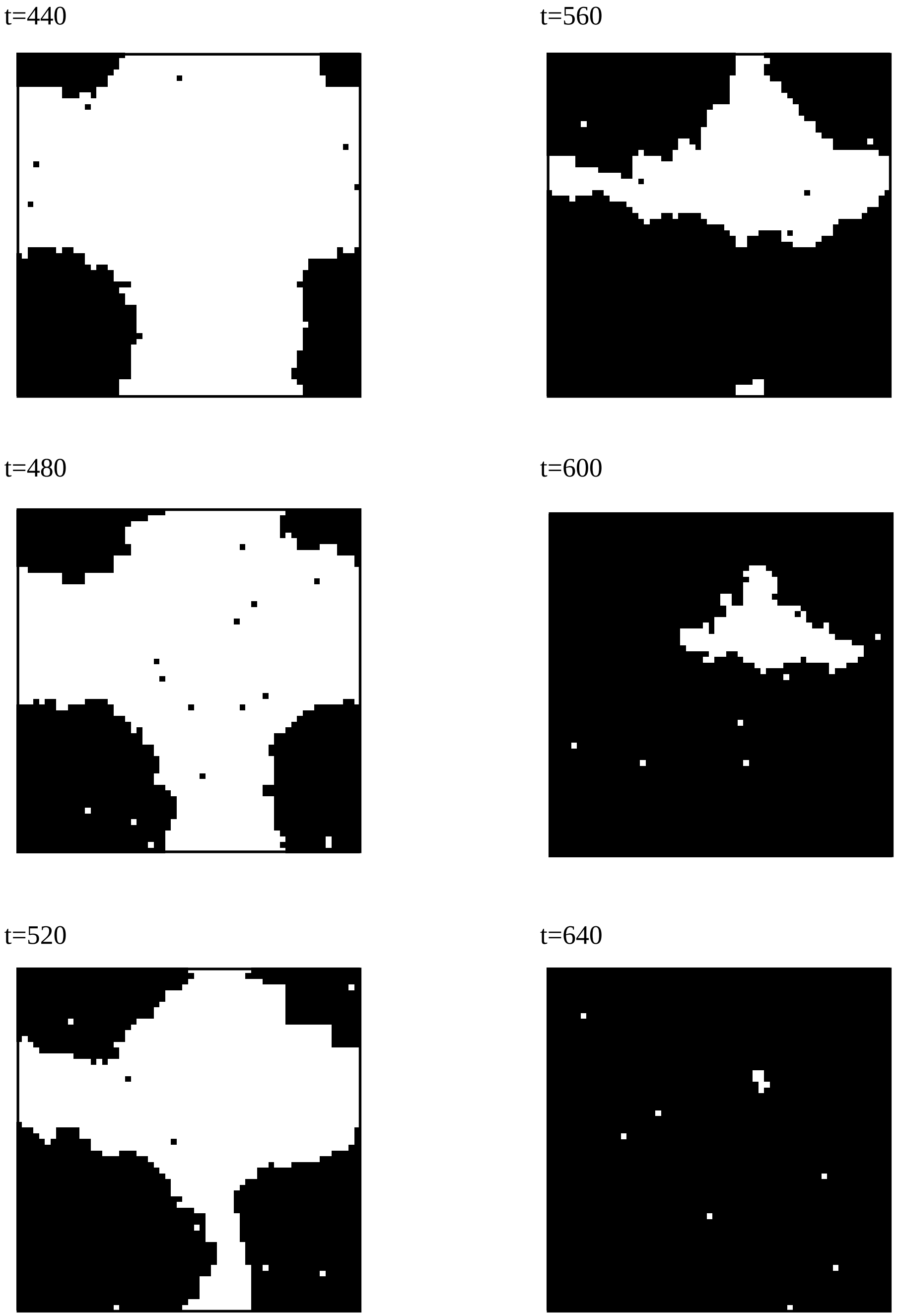


Figure 1.6.11 Frames from a simulation illustrating nucleation and growth in an Ising model in 2-d. The temperature is $T = zJ/3$ and the magnetic field is $h = -0.25$. Glauber dynamics was used. Each time step consists of N updates where the space size is $N = 60 \times 60$. Frames shown are in intervals 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