32

Exact Monte Carlo Sampling

▶ 32.1 The problem with Monte Carlo methods

For high-dimensional problems, the most widely used random sampling methods are Markov chain Monte Carlo methods like the Metropolis method, Gibbs sampling, and slice sampling.

The problem with all these methods is this: yes, a given algorithm can be guaranteed to produce samples from the target density $P(\mathbf{x})$ asymptotically, 'once the chain has converged to the equilibrium distribution'. But if one runs the chain for too short a time T, then the samples will come from some other distribution $P^{(T)}(\mathbf{x})$. For how long must the Markov chain be run before it has 'converged'? As was mentioned in Chapter 29, this question is usually very hard to answer. However, the pioneering work of Propp and Wilson (1996) allows one, for certain chains, to answer this very question; furthermore Propp and Wilson show how to obtain 'exact' samples from the target density.

▶ 32.2 Exact sampling concepts

Propp and Wilson's *exact sampling method* (also known as 'perfect simulation' or 'coupling from the past') depends on three ideas.

Coalescence of coupled Markov chains

First, if several Markov chains starting from different initial conditions share a single random-number generator, then their trajectories in state space may *coalesce*; and, having, coalesced, will not separate again. If *all* initial conditions lead to trajectories that coalesce into a single trajectory, then we can be sure that the Markov chain has 'forgotten' its initial condition. Figure 32.1a-i shows twenty-one Markov chains identical to the one described in section 29.4, which samples from $\{0, 1, \ldots, 20\}$ using the Metropolis algorithm (figure 29.12, p.368); each of the chains has a different initial condition but they are all driven by a single random number generator; the chains coalesce after about 80 steps. Figure 32.1a-ii shows the same Markov chains with a different random number seed; in this case, coalescence does not occur until 400 steps have elapsed (not shown). Figure 32.1b shows similar Markov chains, each of which has identical proposal density to those in section 29.4 and figure 32.1a; but in figure 32.1b, the proposed move at each step, 'left' or 'right', is obtained in the same way by all the chains at any timestep, independent of the current state. This coupling of the chains changes the statistics of coalescence. Because two neighbouring paths merge only when a rejection occurs, and rejections occur only at the walls (for this particular Markov chain), coalescence will occur only when the chains are all in the leftmost state or all in the rightmost state.

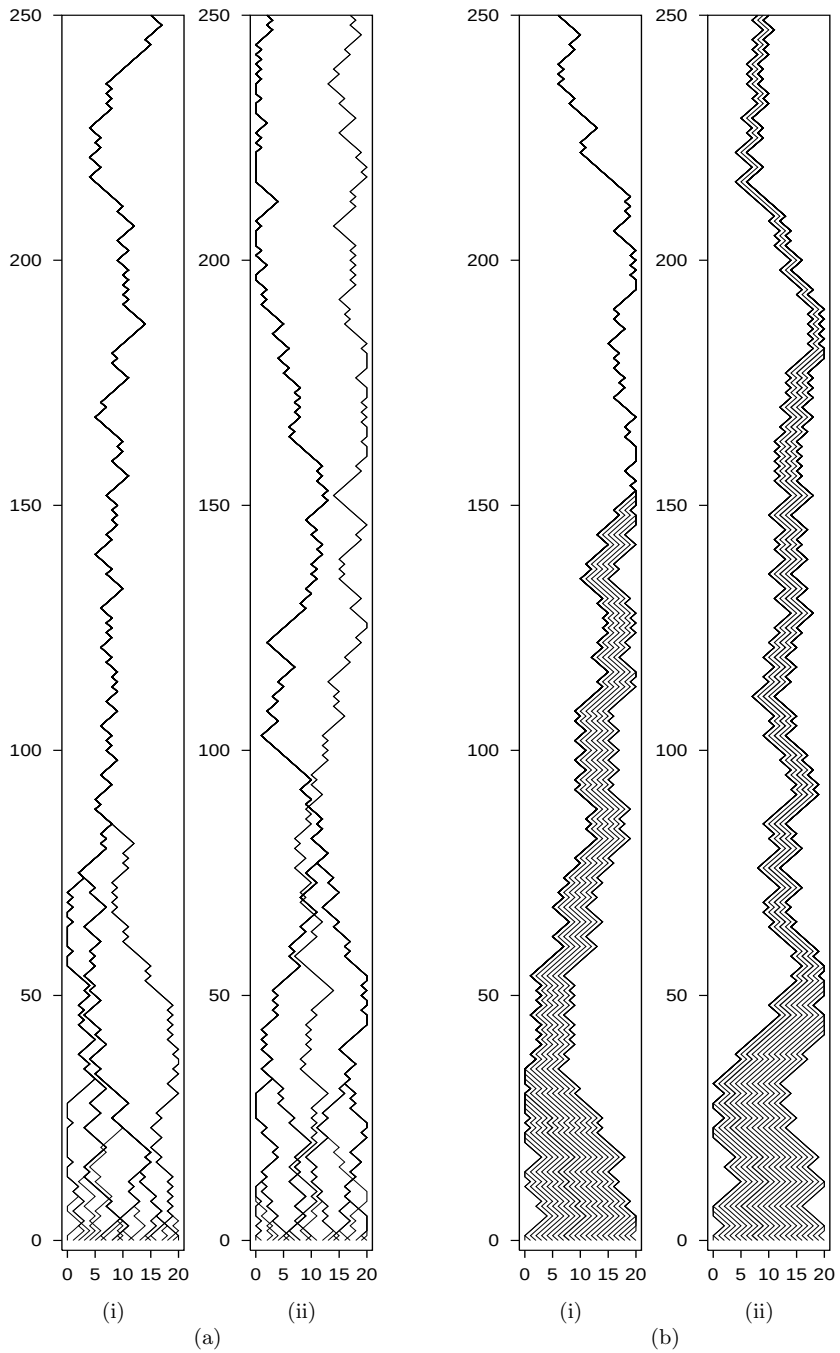

Figure 32.1. Coalescence, the first idea behind the exact sampling method. Time runs from bottom to top. In the leftmost panel, coalescence occurred within 100 steps. Different coalescence properties are obtained depending on the way each state uses the random numbers it is supplied with. (a) Two runs of a Metropolis simulator in which the random bits that determine the proposed step depend on the current state; a different random number seed was used in each case. (b) In this simulator the random proposal ('left' or 'right') is the same for all states. In each panel, one of the paths, the one starting at location $x = 8$, has been highlighted.

Coupling from the past

How can we use the coalescence property to find an exact sample from the equilibrium distribution of the chain? The state of the system at the moment when complete coalescence occurs is not a valid sample from the equilibrium distribution; for example in figure 32.1b, final coalescence always occurs when the state is against one of the two walls, because trajectories merge only at the walls. So sampling forward in time until coalescence occurs is not a valid method.

The second key idea of exact sampling is that we can obtain exact samples by sampling *from a time T_0 in the past, up to the present*. If coalescence has occurred, the present sample is an unbiased sample from the equilibrium distribution; if not, we restart the simulation from a time T_0 further into the past, *reusing the same random numbers*. The simulation is repeated at a sequence of ever more distant times T_0, with a doubling of T_0 from one run to the next being a convenient choice. When coalescence occurs at a time before 'the present', we can record $x(0)$ as an *exact sample* from the equilibrium distribution of the Markov chain.

Figure 32.2 shows two exact samples produced in this way. In the leftmost panel of figure 32.2a, we start twenty-one chains in all possible initial conditions at $T_0 = -50$ and run them forward in time. Coalescence does not occur. We restart the simulation from all possible initial conditions at $T_0 = -100$, and reset the random number generator in such a way that the random numbers generated at each time t (in particular, from $t = -50$ to $t = 0$) will be identical to what they were in the first run. Notice that the trajectories produced from $t = -50$ to $t = 0$ by these runs that started from $T_0 = -100$ are identical to a *subset* of the trajectories in the first simulation with $T_0 = -50$. Coalescence still does not occur, so we double T_0 again to $T_0 = -200$. This time, all the trajectories coalesce and we obtain an exact sample, shown by the arrow. If we pick an earlier time such as $T_0 = -500$, all the trajectories must still end in the same point at $t = 0$, since every trajectory must pass through *some* state at $t = -200$, and *all* those states lead to the same final point. So if we ran the Markov chain for an infinite time in the past, from any initial condition, it would end in the same state. Figure 32.2b shows an exact sample produced in the same way with the Markov chains of figure 32.1b.

This method, called *coupling from the past*, is important because it allows us to obtain exact samples from the equilibrium distribution; but, as described here, it is of little practical use, since we are obliged to simulate chains starting in *all* initial states. In the examples shown, there are only twenty-one states, but in any realistic sampling problem there will be an utterly enormous number of states – think of the 2^{1000} states of a system of 1000 binary spins, for example. The whole point of introducing Monte Carlo methods was to try to avoid having to visit all the states of such a system!

Monotonicity

Having established that we can obtain valid samples by simulating forward from times in the past, starting in *all* possible states at those times, the third trick of Propp and Wilson, which makes the exact sampling method useful in practice, is the idea that, for some Markov chains, it may be possible to detect coalescence of all trajectories *without simulating all those trajectories*. This property holds, for example, in the chain of figure 32.1b, which has the property that *two trajectories never cross*. So if we simply track the two trajectories starting from the leftmost and rightmost states, we will know that

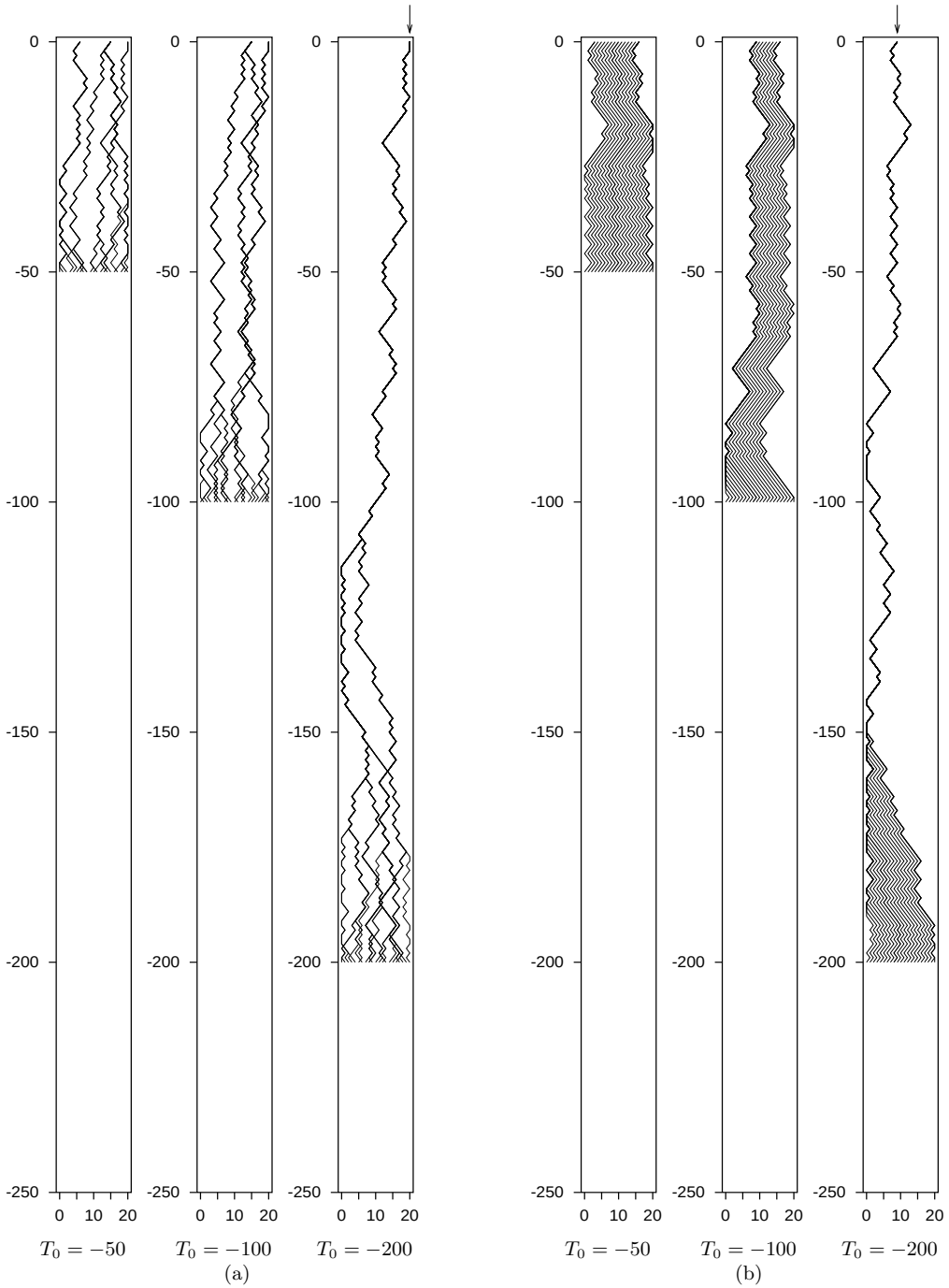

Figure 32.2. 'Coupling from the past', the second idea behind the exact sampling method.

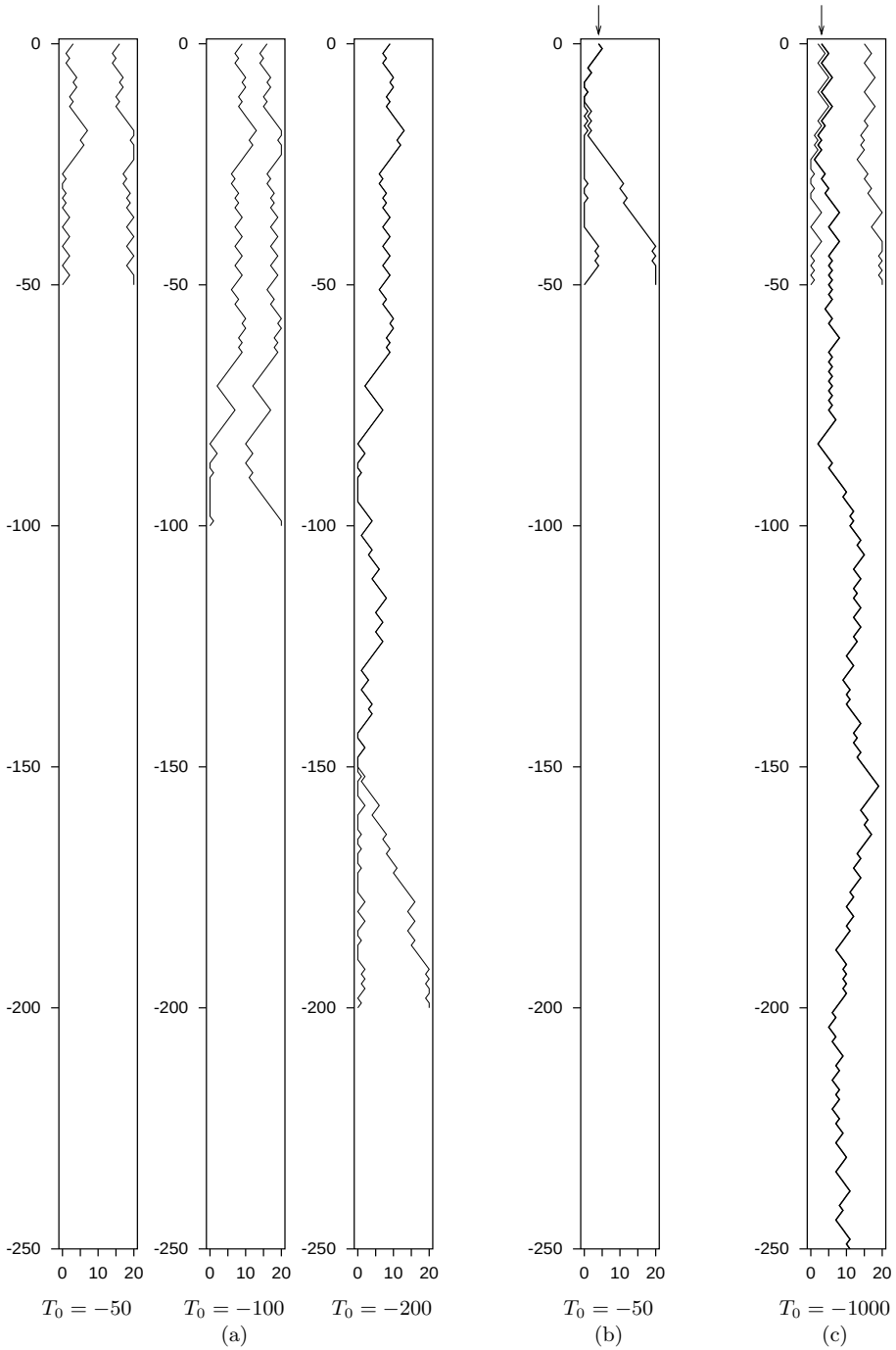

Figure 32.3. (a) Ordering of states, the third idea behind the exact sampling method. The trajectories shown here are the left-most and right-most trajectories of figure 32.2b. In order to establish what the state at time zero is, we only need to run simulations from $T_0 = -50$, $T_0 = -100$, and $T_0 = -200$, after which point coalescence occurs.

(b,c) Two more exact samples from the target density, generated by this method, and different random number seeds. The initial times required were $T_0 = -50$ and $T_0 = -1000$, respectively.

coalescence of *all* trajectories has occurred when *those two* trajectories coalesce. Figure 32.3a illustrates this idea by showing only the left-most and right-most trajectories of figure 32.2b. Figure 32.3(b,c) shows two more exact samples from the same equilibrium distribution generated by running the 'coupling from the past' method starting from the two end-states alone. In (b), two runs coalesced starting from $T_0 = -50$; in (c), it was necessary to try times up to $T_0 = -1000$ to achieve coalescence.

▶ 32.3 Exact sampling from interesting distributions

In the toy problem we studied, the states could be put in a one-dimensional order such that no two trajectories crossed. The states of many interesting state spaces can also be put into a *partial order* and coupled Markov chains can be found that respect this partial order. [An example of a partial order on the four possible states of two spins is this: $(+,+) > (+,-) > (-,-)$; and $(+,+) > (-,+) > (-,-)$; and the states $(+,-)$ and $(-,+)$ are not ordered.] For such systems, we can show that coalescence has occurred merely by verifying that coalescence has occurred for all the histories whose initial states were 'maximal' and 'minimal' states of the state space.

As an example, consider the Gibbs sampling method applied to a ferromagnetic Ising spin system, with the partial ordering of states being defined thus: state $\mathbf{x}$ is 'greater than or equal to' state $\mathbf{y}$ if $x_i \geq y_i$ for all spins i. The maximal and minimal states are the the all-up and all-down states. The Markov chains are coupled together as shown in algorithm 32.4. Propp and Wilson (1996) show that exact samples can be generated for this system, although the time to find exact samples is large if the Ising model is below its critical temperature, since the Gibbs sampling method itself is slowly-mixing under these conditions. Propp and Wilson have improved on this method for the Ising model by using a Markov chain called the single-bond heat bath algorithm to sample from a related model called the random cluster model; they show that exact samples from the random cluster model can be obtained rapidly and can be converted into exact samples from the Ising model. Their ground-breaking paper includes an exact sample from a 16-million-spin Ising model at its critical temperature. A sample for a smaller Ising model is shown in figure 32.5.

Compute $a_i := \sum_j J_{ij} x_j$
Draw u from Uniform$(0, 1)$
If $u < 1/(1 + e^{-2a_i})$
  $x_i := +1$
Else
  $x_i := -1$

Algorithm 32.4. Gibbs sampling coupling method. The Markov chains are coupled together by having all chains update the same spin i at each time step and having all chains share a common sequence of random numbers u.

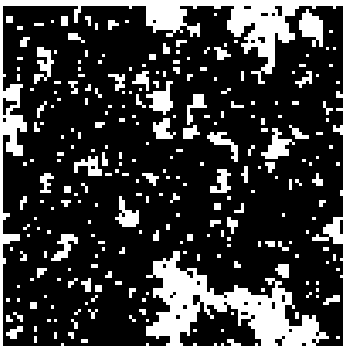

Figure 32.5. An exact sample from the Ising model at its critical temperature, produced by D.B. Wilson. Such samples can be produced within seconds on an ordinary computer by exact sampling.

A generalization of the exact sampling method for 'non-attractive' distributions

The method of Propp and Wilson for the Ising model, sketched above, can be applied only to probability distributions that are, as they call them, 'attractive'. Rather than define this term, let's say what it means, for practical purposes: the method can be applied to spin systems in which all the couplings are positive (e.g., the ferromagnet), and to a few special spin systems with negative couplings (e.g., as we already observed in Chapter 31, the rectangular ferromagnet and antiferromagnet are equivalent); but it cannot be applied to general spin systems in which some couplings are negative, because in such systems the trajectories followed by the all-up and all-down states are not guaranteed to be upper and lower bounds for the set of all trajectories. Fortunately, however, we do not need to be so strict. It is possible to re-express the Propp and Wilson algorithm in a way that generalizes to the case of spin systems with negative couplings. The idea of the *summary state* version of exact sampling is still that we keep track of bounds on the set of

all trajectories, and detect when these bounds are equal, so as to find exact samples. But the bounds will not themselves be actual trajectories, and they will not necessarily be *tight* bounds.

Instead of simulating two trajectories, each of which moves in a state space $\{-1,+1\}^N$, we simulate one *trajectory envelope* in an augmented state space $\{-1,+1,?\}^N$, where the symbol `?` denotes 'either -1 or $+1$'. We call the state of this augmented system the 'summary state'. An example summary state of a six-spin system is `++-?+?`. This summary state is shorthand for the set of states

`++-+++`, `++-++-`, `++--++`, `++--+-` .

The update rule at each step of the Markov chain takes a single spin, enumerates all possible states of the neighbouring spins that are compatible with the current summary state, and, for each of these local scenarios, computes the new value (`+` or `-`) of the spin using Gibbs sampling (coupled to a random number u as in algorithm 32.4). If all these new values agree, then the new value of the updated spin in the summary state is set to the unanimous value (`+` or `-`). Otherwise, the new value of the spin in the summary state is '`?`'. The initial condition, at time T_0, is given by setting all the spins in the summary state to '`?`', which corresponds to considering all possible start configurations.

In the case of a spin system with positive couplings, this summary state simulation will be identical to the simulation of the uppermost state and lowermost states, in the style of Propp and Wilson, with coalescence occuring when all the '`?`' symbols have disappeared. The summary state method can be applied to general spin systems with any couplings. The only shortcoming of this method is that the envelope may describe an unnecessarily large set of states, so there is no guarantee that the summary state algorithm will converge; the time for coalescence to be *detected* may be considerably larger than the actual time taken for the underlying Markov chain to coalesce.

The summary state scheme has been applied to exact sampling in belief networks by Harvey and Neal (2000), and to the triangular antiferromagnetic Ising model by Childs *et al.* (2001). Summary state methods were first introduced by Huber (1998); they also go by the names sandwiching methods and bounding chains.

Further reading

For further reading, impressive pictures of exact samples from other distributions, and generalizations of the exact sampling method, browse the perfectly-random sampling website.[1]

For beautiful exact-sampling demonstrations running live in your web-browser, see Jim Propp's website.[2]

Other uses for coupling

The idea of coupling together Markov chains by having them share a random number generator has other applications beyond exact sampling. Pinto and Neal (2001) have shown that the accuracy of estimates obtained from a Markov chain Monte Carlo simulation (the second problem discussed in section 29.1, p.357), using the estimator

$$\hat{\Phi}_P \equiv \frac{1}{T} \sum_t \phi(\mathbf{x}^{(t)}), \tag{32.1}$$

[1] `http://www.dbwilson.com/exact/`

[2] `http://www.math.wisc.edu/~propp/tiling/www/applets/`

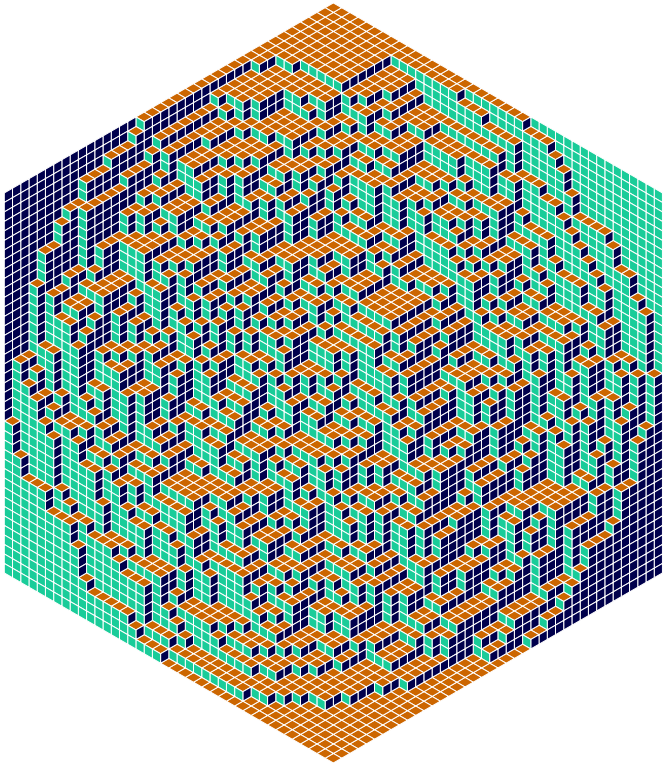

Figure 32.6. A perfectly random tiling of a hexagon by lozenges, provided by J.G. Propp and D.B. Wilson.

can be improved by coupling the chain of interest, which converges to P, to a second chain, which generates samples from a second, simpler distribution, Q. The coupling must be set up in such a way that the states of the two chains are strongly correlated. The idea is that we first estimate the expectations of a function of interest, ϕ, under P and under Q in the normal way (32.1) and compare the estimate under Q, $\hat{\Phi}_Q$, with the true value of the expectation under Q, Φ_Q which we assume can be evaluated exactly. If $\hat{\Phi}_Q$ is an overestimate then it is likely that $\hat{\Phi}_P$ will be an overestimate too. The difference $(\hat{\Phi}_Q - \Phi_Q)$ can thus be used to correct $\hat{\Phi}_P$.

32.4 Exercises

▷ Exercise 32.1.[2, p.421] Is there any relationship between the probability distribution of the time taken for all trajectories to coalesce, and the equilibration time of a Markov chain? Prove that there is a relationship, or find a single chain that can be realized in two different ways that have different coalescence times.

▷ Exercise 32.2.[2] Imagine that Fred ignores the requirement that the random bits used at some time t, in every run from increasingly distant times T_0, must be identical, and makes a coupled-Markov-chain simulator that uses fresh random numbers every time T_0 is changed. Describe what happens if Fred applies his method to the Markov chain that is intended to sample from the uniform distribution over the states 0, 1, and 2, using the Metropolis method, driven by a random bit source as in figure 32.1b.

Exercise 32.3.[5] Investigate the application of perfect sampling to linear regression in Holmes and Mallick (1998) or Holmes and Denison (2002) and try to generalize it.

Exercise 32.4.[3] The concept of coalescence has many applications. Some surnames are more frequent than others, and some die out altogether. Make

a model of this process; how long will it take until everyone has the same surname?

Similarly, variability in any particular portion of the human genome (which forms the basis of forensic DNA fingerprinting) is inherited like a surname. A DNA fingerprint is like a string of surnames. Should the fact that these surnames are subject to coalescences, so that some surnames are by chance more prevalent than others, affect the way in which DNA fingerprint evidence is used in court?

▷ Exercise 32.5.[2] How can you use a coin to create a random ranking of 3 people? Construct a solution that uses exact sampling. For example, you could apply exact sampling to a Markov chain in which the coin is repeatedly used alternately to decide whether to switch first and second, then whether to switch second and third.

Exercise 32.6.[5] Finding the partition function Z of a probability distribution is a difficult problem. Many Markov chain Monte Carlo methods produce valid samples from a distribution without ever finding out what Z is.

Is there any probability distribution and Markov chain such that either the time taken to produce a perfect sample or the number of random bits used to create a perfect sample are related to the value of Z? Are there some situations in which the time to coalescence conveys information about Z?

▶ 32.5 Solutions

Solution to exercise 32.1 (p.420). It is perhaps surprising that there is no direct relationship between the equilibration time and the time to coalescence. We can prove this using the example of the uniform distribution over the integers $\mathcal{A} = \{0, 1, 2, \ldots, 20\}$. A Markov chain that converges to this distribution in exactly one iteration is the chain for which the probability of state s_{t+1} given s_t is the uniform distribution, for all s_t. Such a chain can be coupled to a random number generator in two ways: (a) we could draw a random integer $u \in \mathcal{A}$, and set s_{t+1} equal to u regardless of s_t; or (b) we could draw a random integer $u \in \mathcal{A}$, and set s_{t+1} equal to $(s_t + u) \bmod 21$. Method (b) would produce a cohort of trajectories locked together, similar to the trajectories in figure 32.1, except that no coalescence ever occurs. Thus, while the equilibration times of methods (a) and (b) are both one, the coalescence times are respectively one and infinity.

It seems plausible on the other hand that coalescence time provides some sort of upper bound on equilibration time.