# Chapter 4
# Markov chain Monte Carlo methods

In this chapter we discuss the need for random-sampling methods, and justify their use rigorously. Given a random variable $X(\pi)$, such as any of those presented in chapter 3, the expectation of $X$ is (equation (2.1.6))

$$\mathbb{E}[X] = \sum_{\pi \in \mathcal{S}_N} P_{\text{Gibbs}}(\pi) X(\pi).$$

This is a real number, with no uncertainty. The problem is that the number of permutations, $N!$, grows intractably in $N$: even for $L = 10$ (and we consider $L$ up to 80), $N = 1000$ and $N!$ is a number with over 5,000 digits. The true expectation is effectively incomputable. Expectations are instead *estimated* by summing over some number $M$ (in the current work, $10^5$ or $10^6$) of typical permutations. The sample mean

$$\langle X \rangle_M = \frac{1}{M} \sum_{k=1}^{M} X(\pi_k) \tag{4.0.1}$$

depends on the random sequence $\pi_1, \ldots, \pi_M$. It is now a random variable with its own variance. The two main sources of error in MCMC simulations are *initialization bias* and *sampling variability.* The former involves *thermalization* (section 9.6) and *multimodality of distributions* (sections 5.4 and 7.8); the latter involves *autocorrelation* (section 9.15 and appendix B).

To create such a sequence of system states (for us, permutations), the method used throughout the computational physics community [Berg, LB] is Markov chain Monte Carlo. Namely, given a permutation $\pi_k$, one selects a successor permutation $\pi_{k+1}$ in some random way. This is the Monte Carlo part. Moreover, the transition probabilities from $\pi_k$ to each candidate $\pi_{k+1}$ depend only on $\pi_k$, and not on any previous choices. This is the Markov chain part.

In the next sections we show (1) we can construct Markov chains which sample from the Gibbs distribution $P_{\text{Gibbs}}(\pi)$ (equation (2.1.4)); (2) other distributions (induced by the selection of initial state) converge to the Gibbs distribution; (3) the sample mean $\langle X \rangle_M$ converges almost surely to the true expectation $\mathbb{E}[X]$; and (4) the variance of $\langle X \rangle_M$ can be estimated, allowing us to place error bars on our estimates of $\mathbb{E}[X]$. Most of the material in this chapter is familiar: see [Berg, LB, CB, FG, GS] to name only a few. Results are restated here for self-containment of presentation.

## 4.1 Markov chains and Markov matrices

Before continuing to discuss random permutations and the Gibbs measure, we spend some time discussing more general random sequences, including Markov chains as a special case. This will turn out to be worthwhile: one of the strengths of this dissertation, in the author's estimation, is the careful disambiguation of some misleading notation and terminology (principally, overuse of the single letter $P$) which are encountered from time to time in the literature.

Let $\Omega$ be a finite set, and put $\#\Omega = K$. (For example, $\Omega = \mathcal{S}_N$ with $K = N!$.) The set of all sequences of elements of $\Omega$, indexed by the non-negative integers, is $\Omega^{\mathbb{N}}$. Just as we can have an arbitrary probability measure $P$ on $\Omega$, we can have an arbitrary probability measure $\mathbf{P}$ on $\Omega^{\mathbb{N}}$. Marginal distributions on the $k$th slot are written $P_k$. If $S_k$ is an $\Omega$-valued random variable, e.g. a random selection from $\Omega$ at the $k$th slot, then $S_0, S_1, S_2, \ldots$ is a *random sequence*, or *discrete-time random process*. Since $\mathbf{P}$ is arbitrary, the $P_i$ are not necessarily the same distributions, and samples $S_i$ and $S_j$ at the $i$th and $j$th slots are not necessarily independent.

Repeatedly using the conditional-probability formula $P(E \mid F) = P(E, F)/P(F)$ for events $E$ and $F$, we can always split up the probability of a finite sequence of samples into a sequencing of initial and conditional probabilities:

$$\begin{aligned}\mathbf{P}(S_1 = \omega_1, S_2 = \omega_2, \ldots, S_n = \omega_n) =& \mathbf{P}(S_1 = \omega_1) \\ & \cdot \mathbf{P}(S_2 = \omega_2 \mid S_1 = \omega_1) \\ & \cdot \mathbf{P}(S_3 = \omega_3 \mid S_1 = \omega_1, S_2 = \omega_2) \\ & \cdot \mathbf{P}(S_n = \omega_n \mid S_1 = \omega_1, \cdots, S_{n-1} = \omega_{n-1}).\end{aligned} \tag{4.1.1}$$

A *Markov process* (or *Markov chain* if the state space $\Omega$ is finite) is a discrete-time random process such that for all $k > 0$,

$$\mathbf{P}(S_k = \omega_k \mid S_1 = \omega_1, S_2 = \omega_2, \ldots, S_{k-1} = \omega_{k-1}) = \mathbf{P}(S_k = \omega_k \mid S_{k-1} = \omega_{k-1}).$$

That is, if the probability of moving from one state to another depends only on the previous sample, and on nothing farther into the past, then the process is Markov. Now we have

$$\begin{aligned}\mathbf{P}(S_1 = \omega_1, \ldots, S_n = \omega_n) = & \mathbf{P}(S_1 = \omega_1) \\ & \cdot \mathbf{P}(S_2 = \omega_2 \mid S_1 = \omega_1) \cdots \cdot \mathbf{P}(S_n = \omega_n \mid S_{n-1} = \omega_{n-1}).\end{aligned} \tag{4.1.2}$$

We have the initial distribution for the first state, then transition probabilities for subsequent states. Precisely, one says a Markov chain is a discrete-time random process with this Markov property. With slight abuse of notation, though, we also refer to the probability distribution $\mathbf{P}$ as a Markov chain if it has this property, since given $\mathbf{P}$ we can always construct a discrete-time random process $S_0, S_1, S_2, \ldots$.

Additionally, if for all $\omega, \omega' \in \Omega$ the conditional probabilities $\mathbf{P}(S_{k+1} = \omega' \mid S_k = \omega)$ are the same for all $k$, then we say the Markov chain is *homogeneous*.

Observe that $\mathbf{P}(S_{k+1} = \omega_j \mid S_k = \omega_i)$ is a $K \times K$ matrix of numbers between zero and one, with the property that rows sum to one (since each $\omega_i$ must transition to *some* $\omega_j$). Such a matrix is called a *stochastic matrix* or *Markov matrix*. We might as well name that matrix $A_k$, with the entry in the $i$th row and $j$th column given by

$$(A_k)_{ij} = \mathbf{P}(S_{k+1} = \omega_j \mid S_k = \omega_i).$$

If the chain is homogeneous, we omit the subscript and write $A$. The key to making linear algebra out of this setup is the following *law of total probability*:

$$\begin{aligned} \mathbf{P}(S_{k+1} = \omega_j) &= \sum_{\omega_i} \mathbf{P}(S_k = \omega_i, S_{k+1} = \omega_j) \\ &= \sum_{\omega_i} \mathbf{P}(S_k = \omega_i)\mathbf{P}(S_{k+1} = \omega_j \mid S_k = \omega_i) \\ &= \sum_{\omega_i} \mathbf{P}(S_k = \omega_i)(A_k)_{ij}. \end{aligned} \tag{4.1.3}$$

The probability mass functions $P_k$ are row vectors. The PMF $P_{k+1}$ of $S_{k+1}$ is the PMF $P_k$ of $S_k$ times the Markov matrix $A_k$. In vector/matrix notation,

$$P_{k+1} = P_k A_k.$$

Throughout this section, we supposed we had been given a probability distribution $\mathbf{P}$ on $\Omega^{\mathbb{N}}$ which satisfied the Markov property; we obtained Markov transition matrices. If, on the other hand, we start with an initial distribution $P_0$ and stochastic matrices $A_0, A_1, \ldots$, then we can re-use equation (4.1.3) to obtain $P_1 = P_0 A_0$ and, inductively, $P_{k+1} = P_k A_k$. We can then use equation (4.1.2) in reverse to obtain a probability distribution $\mathbf{P}$ on $\Omega^{\mathbb{N}}$. Note that now the process is Markov by construction.

In summary, a Markov chain is specified by a sequence-space distribution $\mathbf{P}$ with the Markov property, or an initial distribution $P_0$ and a sequence of transition matrices. Furthermore, we can go back and forth between these two points of view:

$$\mathbf{P} \longleftrightarrow (P_0, A_0, A_1, A_2, \ldots).$$

If the chain is homogeneous, we write

$$\mathbf{P} \longleftrightarrow (P_0, A).$$

## 4.2 Invariant and limiting distributions

The only probability distribution for random permutations considered thus far is the Gibbs distribution $P_{\text{Gibbs}}$ of equation (2.1.4) on page 23. However, others exist. If one is asked for a permutation $\pi$ and one always replies with the identity, then the distribution of that answer is the singleton measure supported on the identity. This is not the Gibbs measure (unless $\alpha = 0$ and $\beta = 0$). A third distribution is the uniform distribution, where each permutation has probability $1/N!$. (This is the same as the Gibbs measure only for $\alpha = 0$ and $T = 0$.)

Following the construction of the final paragraph of section 4.1, suppose we select an initial permutation $\pi_0$ from some probability distribution: for computational work done in this dissertation, this will be the singleton measure supported at the identity permutation, although a uniform-random $\pi_0$ is another possibility. Given Markov transition matrices $A_k$ to be constructed in sections 5.2 and 7.6, we obtain a random sequence of permutations $\pi_0$, $\pi_1$, $\pi_2$, .... We write $\mathbf{P}^{(\pi_0)}$ for the probability distribution on this sequence space, with $P_k^{(\pi_0)}$ being the marginal at the $k$th slot.

Below, we will construct Markov transition matrices $A_k$ which preserve the Gibbs measure $P_{\text{Gibbs}}$, i.e. $P_{\text{Gibbs}} = P_{\text{Gibbs}} A_k$. The following terminology applies.

**Definition 4.2.1.** A probability distribution $P$ is *invariant* for a Markov transition matrix $A$ if $P = PA$, that is, if for all $j = 1, \ldots, K$,

$$P(S_2 = \omega_j) = \sum_{i=1}^{K} A_{ij} P(S_1 = \omega_i).$$

In vector/matrix notation, this means

$$P = PA.$$

In other words, $P$ is invariant for $A$ if $A$-transitions preserve the distribution $P$. If $S_1$ is distributed according to $P$, then $S_2$ will also be distributed according to $P$, and so on. Such a sequence of states $S_1, S_2, \ldots$, while not in general independent, is identically distributed.

**Remark 4.2.2.** A homogeneous chain is not the same as a stationary sequence. In the former case, the transition matrix is the same at each time step; in the latter case, the probability distributions are the same at each time step.

**Example 4.2.3.** $\triangleright$ An illustrative example uses die-tipping. Recall that an ordinary six-sided die has pips on opposite faces summing to seven. There are six states, which we assume to be uniformly distributed if the die is rolled. Given that the die has $n$ pips facing up, we may *tip* the die by picking one of the four sides at uniform random and putting that side up. E.g. if 1 is up, then after the tip, 2, 3, 4, or 5 will appear each with probability 1/4; 1 or 6 will appear with probability 0.

The transition matrix, with rows indexing the current state and columns indicating the successor state, is

$$A = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{pmatrix}.$$

Suppose the die is initially set on table with 1 up, i.e.

$$P_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$P_1 = P_0\,A = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{pmatrix}$$

If, instead, the die is initially rolled, then $P_0$ is uniform:

$$P_0 = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

One computes $P_1$ to be uniform as well:

$$P_1 = P_0\,A = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

The 1-up distribution is not invariant for the die-tipping transition rule, but the die-roll distribution is. $\triangleleft$

Given a Markov matrix $A$, one may wish to find an invariant distribution $P$. Going the other way, given a distribution $P$, one may wish to construct a Markov matrix $A$ such that $P$ is invariant with respect to $A$. The latter is our main goal here: the distribution of interest is the Gibbs distribution $P_{\text{Gibbs}}$. The following theorem is key. First, we define the terminology necessary to state it.

**Definition 4.2.4.** A Markov chain $\mathbf{P}$ on a state space $\Omega$ is *irreducible* if for all $\omega, \omega' \in \Omega$ there exists an $n > 0$ such that $\mathbf{P}(S_n = \omega' \mid S_0 = \omega) > 0$.

**Definition 4.2.5.** The *period* of $\omega \in \Omega$ is

$$p(\omega) = \gcd\{n : \mathbf{P}(S_n = \omega \mid S_0 = \omega) > 0\}$$

We say that $\omega$ has period $p$ if it reappears with probability 1 after every $p$ steps. A state $\omega$ is *aperiodic* if $p(\omega) = 1$. A chain is *aperiodic* if $p(\omega) = 1$ for every $\omega$.

**Remark.** An irreducible, aperiodic chain on a finite state space is sometimes called *ergodic*.

**Definition 4.2.6.** A Markov matrix $A$ on a state space $\Omega$ (with $\#\Omega = K$) and a distribution $P$ on $\Omega$ are *reversible*, or satisfy *detailed balance*, if for all $1 \leq i, j \leq K$,

$$A_{ij}\, P(\omega_i) = A_{ji}\, P(\omega_j).$$

**Theorem 4.2.7** (Invariant-distribution theorem)**.** *Fix a finite set $\Omega$. Let $A$ be a Markov transition matrix on $\Omega$, as above, and let $P$ be a probability distribution on $\Omega$. If the homogeneous Markov chain $(P, A)$ is irreducible, aperiodic, and satisfies detailed balance, then $P$ is invariant for $A$. That is, the Markov chain $(P, A)$ is stationary. Furthermore, if another homogeneous Markov chain $(P_0, A)$ is irreducible and aperiodic, then for all $\omega \in \Omega$, $P_n(\omega) \to P(\omega)$ as $n \to \infty$.*

**Remark.** Some authors call this the *ergodic theorem*; yet, others call our theorem 4.2.9 the ergodic theorem. It also may be thought of as a special case of the Perron-Frobenius theorem, applied to the Markov transition matrix $A$.

**Proof.** See [Lawler], sections 1.2 and 1.3. □

We say $P$ ($P_{\text{Gibbs}}$ in the context of random permutations) is the *stationary distribution* or *invariant distribution* for $A$. We say that it is also a *limiting distribution* for any initial distribution $P_0$ satisfying the above hypotheses.

**Example 4.2.8.** $\triangleright$ Continuing example 4.2.3: If we initially roll the die, we start with the uniform distribution which is stationary for the die-tipping rule:

$$\begin{aligned} P_0 &= \begin{pmatrix} 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 \end{pmatrix} \\ P_1 = P_0\, A &= \begin{pmatrix} 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 \end{pmatrix} \\ &\vdots \end{aligned}$$

If we start with the one-face up, we begin with the singleton initial distribution which is not stationary for the die-tipping rule. Yet, subsequent tips have a distribution which tends toward the limiting, uniform distribution:

$$\begin{aligned}
P_0 &= \begin{pmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix} \\
P_1 = P_0\,A &= \begin{pmatrix} 0.0000 & 0.2500 & 0.2500 & 0.2500 & 0.2500 & 0.0000 \end{pmatrix} \\
P_2 = P_1\,A &= \begin{pmatrix} 0.2500 & 0.1250 & 0.1250 & 0.1250 & 0.1250 & 0.2500 \end{pmatrix} \\
P_3 = P_2\,A &= \begin{pmatrix} 0.1250 & 0.1875 & 0.1875 & 0.1875 & 0.1875 & 0.1250 \end{pmatrix} \\
P_4 = P_3\,A &= \begin{pmatrix} 0.1875 & 0.1562 & 0.1562 & 0.1562 & 0.1562 & 0.1875 \end{pmatrix} \\
P_5 = P_4\,A &= \begin{pmatrix} 0.1562 & 0.1719 & 0.1719 & 0.1719 & 0.1719 & 0.1562 \end{pmatrix} \\
&\vdots \\
P_{14} = P_{13}\,A &= \begin{pmatrix} 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 \end{pmatrix} \\
P_{15} = P_{14}\,A &= \begin{pmatrix} 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 \end{pmatrix} \\
&\vdots
\end{aligned}$$

$\triangleleft$

Importantly, for simulations using the model of random spatial permutations, we need not know *a priori* what a typical permutation looks like. We may start always with the identity permutation, i.e. $P_0$ is the singleton distribution supported on the identity permutation. We may then run the Markov chain, producing a sequence of permutations. As the number of iterations goes to infinity, the distribution of permutations approaches $P_{\text{Gibbs}}$: for all $\pi \in \mathcal{S}_N$, $P_k^{(\pi_0)}(\pi) \to P_{\text{Gibbs}}(\pi)$ as $k \to \infty$.

The specific number $k$ of iterations needed for convergence of $P_k^{(\pi_0)}$ to $P_{\text{Gibbs}}$ is another matter entirely. The theory exposited by [Lawler], as noted above, guarantees that a Markov matrix $A$ with $P_{\text{Gibbs}}$ as its invariant distribution has no other invariant distribution: $A$ has a single eigenvalue 1 with eigenvector $P_{\text{Gibbs}}$. The rate of convergence of $P_k^{(\pi_0)}$ to $P_{\text{Gibbs}}$ depends on the second-largest eigenvalue for $A$, which one in general does not know how to compute. In practice, this *mixing time* (or *burn-in time* or *thermalization time*) is estimated using techniques such as those in section 9.6.

The theory above also does not tell us how to construct a Markov matrix $A$ having a desired distribution $P_{\text{Gibbs}}$ as its invariant distribution — it simply tells us what we can do once we have constructed such a matrix. A specific construction is due to Nicholas Metropolis [Berg, LB, CB]. The essence is that if the invariant probability distribution $P_{\text{Gibbs}}$ is defined as a Gibbs measure via an energy function $H$ on $\Omega$, then proposed successor states $\pi_{k+1}$ of $\pi_k$ are accepted with probability $\min\{1, e^{-\Delta H}\}$ where $\Delta H$ is the energy difference for the state change. In this dissertation, I directly prove that such methods result in detailed balance. Thus, the reader is referred to propositions 5.3.6 and 7.7.5 for details.

We next show that the sample mean $\langle X \rangle_M$ (equation (4.0.1) on page 40) converges to the true mean $\mathbb{E}[X]$ (equation (2.1.6) on page 23).

**Theorem 4.2.9** (Ergodic sampling theorem)**.** *Let $X$ be a random variable on the finite probability space $(\Omega, 2^\Omega, P)$ where $2^\Omega$ is the power set of $\Omega$. If the stationary, homogeneous Markov chain $(P, A)$ satisfies the hypotheses of theorem 4.2.7, then*

$$\frac{1}{M}\sum_{k=1}^{M} X(S_k) \to \mathbb{E}[X] \quad \textit{as} \quad M \to \infty.$$

**Remark.** Some authors call this this *ergodic theorem*; yet, others call our theorem 4.2.7 the ergodic theorem.

**Proof.** This follows from theorem 4.2.7 using the central-limit theorem for identically distributed but non-independent $S_i$'s. □

## 4.3 Sample means and their variances

There is one caveat to replacing the true mean $\mathbb{E}[X]$ with the MCMC sample mean $\langle X\rangle_M$: the naive computation of the standard deviation of the sample mean, which is correct for independent identically distributed states, is a significantly incorrect underestimate for the standard deviation of the sample mean in the case of identically distributed but correlated states. This issue is so important that appendix B is devoted to it. The key result of that appendix is that the standard deviation (error bar) of the sample mean $\langle X\rangle_M$ is off by a factor of the square root of the *integrated autocorrelation time*, $\sqrt{\hat{\tau}_{\text{int}}(X)}$, which is computed as described in section B.10.

## 4.4 Simple example: 1D Ising

For an example which is more complex than die-tipping but less complex than random spatial permutations, consider the 1D $N$-point Ising model. Namely, the configuration space is $\Omega = \{\pm 1\}^N$, i.e. $N$ particles which may be in either an up (filled) or a down (hollow) state:

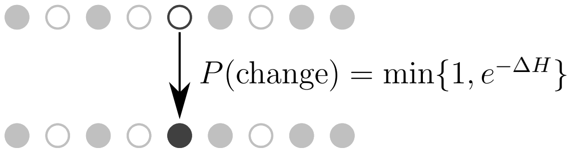
● ○ ● ○ ○ ● ○ ● ●

A state is described by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$. The configuration space $\Omega$ has $2^N$ possible configurations. The system is endowed with an energy function. For the 1D Ising model, one has

$$H(\boldsymbol{\omega}) = \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}\omega_i\omega_j + \sum_{i=1}^{n} h_i\omega_i.$$

where the $C_{ij}$'s are interaction terms (non-interacting, nearest neighbor, mean-field, etc.) and the $h_i$'s are magnetization terms. Given a temperature-related parameter $\beta$, one sets the (Gibbs) probability of each configuration to be proportional to $e^{-\beta H}$.

One picks an initial configuration. There are three obvious choices: (1) Start with all spins down, i.e. $\boldsymbol{\omega} = (-1, \dots, -1)$. (2) Start with all spins up, i.e. $\boldsymbol{\omega} = (+1, \dots, +1)$. (3) Start with $\boldsymbol{\omega}$ selected from a uniform probability distribution on $\Omega$. Then, one selects a site $i$ and decides whether to flip $\omega_i$ to $-\omega_i$.



This decision is made using the Metropolis prescription, namely:

- One computes the change in energy $\Delta H = H(\boldsymbol{\omega}') - H(\boldsymbol{\omega})$ which would be obtained if $\boldsymbol{\omega}$ were sent to $\boldsymbol{\omega}'$ by flipping $\omega_i$.

- One may compute $\Delta H$ by separately computing $H(\boldsymbol{\omega}')$ and $H(\boldsymbol{\omega})$ and subtracting the two. However, since the only change is at the site $i$, one may do some ad-hoc algebra to derive an expression for $\Delta H$ which is less computationally expensive.

- One accepts the change with probability $\min\{1, e^{-\Delta H}\}$.

This is called a *Metropolis step*.

Looping through all $n$ sites from $i = 1$ to $i = n$, performing a Metropolis step at each site $i$, is called a *Metropolis sweep*. If one realizes a random variable $X(\boldsymbol{\omega})$ at each of $M$ sweeps, averaging $X$ over the $M$ sweeps, one obtains an approximation $\langle X \rangle_M$ for the expectation $\mathbb{E}[X]$.

As discussed at the end of section 4.2, one should first run some number $B$ of Metropolis sweeps of the system until it is thermalized, i.e. until the Gibbs distribution has been approached. One should discard the $B$ realizations of the random variable $X$ obtained during thermalization, before running the $M$ sweeps in which data are accumulated. The $B$ sweeps are called the thermalization phase; the $M$ sweeps are called the accumulation phase.

## 4.5 Recipe for MCMC algorithms

The naive outline of an MCMC run is simple:

- Use a Markov chain, discussed at the end of this section, to generate a sequence $\pi_1, \dots, \pi_M$ of permutations.

- For each permutation $\pi_k$, for each random variable $X$ of interest, remember the value $X_k = X(\pi_k)$.

- Compute the sample mean $\overline{X} = \frac{1}{M}\sum_{k=1}^{M} X_k$. Also compute the sample standard deviation, and any other desired statistics.
- Display the statistics.

Since the initial permutation is the identity, the initial distribution is the singleton supported at the identity, which is not the Gibbs distribution $P_{\text{Gibbs}}$. Furthermore, as a very low-energy state, the identity is highly non-typical with respect to the Gibbs distribution (equation (2.1.4)) for the model of spatial permutations. As discussed at the end of section 4.2, one runs the chain until it is thermalized, i.e. until the Gibbs distribution has been approached. (The number of steps $\tau$ required for this is random, but it turns out to fall within a narrow range.) Renumbering $\pi_\tau$ to $\pi_0$, one then accumulates statistics over the $M$ permutations $\pi_1, \ldots, \pi_M$. See section 9.6 for the thermalization-detection algorithm used in this dissertation.

Thus the computational recipe is as follows:

- Start with the initial permutation being the identity permutation.
- Run the Markov chain, generating a sequence of permutations until thermalization has been detected. At that point, rename the current permutation $\pi_0$.
- Continue generating a sequence $\pi_1, \ldots, \pi_M$ of permutations. The mechanics of transitioning from $\pi_k$ to $\pi_{k+1}$ comprises a series of *steps*, which collectively form the $k$th *sweep*. Various types of sweep — swap-only, swap-and-reverse, swaps with band updates, and worm — are presented in chapters 5, 6, and 7.
- For each permutation $\pi_k$, for each random variable $X$ of interest, remember the value $X_k = X(\pi_k)$.
- After all $M$ permutations have been generated, compute the sample mean $\overline{X} = \frac{1}{M}\sum_{k=1}^{M} X_k$. Also compute the sample standard deviation and its error bar (using estimated integrated autocorrelation time), and any other desired statistics, such as histograms.
- Display the statistics.

Given the framework established by previous sections of this chapter, the recipe to prove correctness of this algorithm reduces to the following: define a Markov chain, then prove irreducibility, aperiodicity, and detailed balance. We devote the next chapters to present three Markov chains: the swap-only algorithm, the swap-and-reverse algorithms, and the worm algorithm.