Topics in
Probability Theory and Stochastic Processes
Steven R. Dunbar

Nonstandard Examples of Markov Chains

Rating
Mathematicians Only: prolonged scenes of intense rigor.
Section Starter Question

1. What is the “cycle and transposition” notation for elements of the group of permutations on $n$ elements?

2. Explain an algorithm for selecting a random permutation from the group of permutations on $n$ elements.

---

Key Concepts

1. Let $X = S_n$, the symmetric group of $n$ letters, that is, the set of permutations on $n$ letters with the Cayley distance as metric. The metric defines a probability distribution on $S_n$ by

$$
\pi(\sigma) = z^{-1} \theta^{d(\sigma, \sigma_0)}, \sigma, \sigma_0 \in S_n, 0 \leq \theta \leq 1.
$$

Choose a transposition $(i, j)$ uniformly at random and consider $(i, j)\sigma = \sigma^*$. Using the Metropolis algorithm with this transposition defines a Markov chain on the symmetric group. A theorem from Diaconis [3, 1] shows that starting from the identity and using the identity as $\sigma_0$, on the order of $n \log n$ steps are necessary and sufficient to make the distance to the stationary distribution $\pi$ small.

2. The top-to-random shuffle on a deck of $n$ cards is a Markov chain which takes $\Theta(n \log n)$ steps from the identity starting order to reach a near-uniform distribution over the set of permutations.

3. A state of the particles in statistical mechanics is described by a configuration $\omega$ from the configuration space $\Omega$. The physics of a configuration space $\Omega$ is described by an energy function $E : \Omega \rightarrow \mathbb{R}^+$. A basic principle of statistical physics is that Nature seeks low-energy configurations. The total energy of the system is the expected value of the energy function. Defining a probability transition with the Metropolis
algorithm, choose configurations with probability $e^{E(\omega)/kT}$ and weight them uniformly to find the expected value.

4. As a model, the physicist Kirkwood posed the problem of whether a gas of hard disks would show a phase transitions. Consider placement of $n$ discs of radius $\epsilon$ in the unit square. This hard disks problems is the original motivation for the Metropolis algorithm defining a Markov chain in the configuration space.

---

**Vocabulary**

1. Let $X = S_n$, the **symmetric group** of $n$ letters, that is, the set of permutations on $n$ letters. Let $d(\sigma, \sigma_0)$ be a metric on the symmetric group by

$$d(\sigma, \sigma_0) = \text{the minimum number of transpositions required to bring } \sigma \text{ to } \sigma_0.$$

This metric is called **Cayley’s distance**.

2. A **Top-to-Random Shuffle**, takes the top card from a stack of $n$ cards and inserts it in the gap between the $(k-1)$th card and the $k$th card in the deck.

3. If $\mu$ and $\nu$ are probability distributions on $\Omega$, the **total variation distance** of $\mu$ from $\nu$ is

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

4. Say that $\tau_{\text{top}}$ is a **strong stationary time** for $X_t$, $t \geq 0$ if $X_{\tau_{\text{top}}+1} \sim \text{unif}(S_n)$, and $X_{\tau_{\text{top}}+1}$ is independent of $\tau_{\text{top}}$. 
5. In statistical mechanics, a state of the particles is described by a configuration \( \omega \) from the configuration space \( \Omega \). The configuration can be infinite or finite, continuous or discrete.

6. For \( \Omega \) a bounded subset \( \Lambda \) of the integer lattice in the plane, attach a value \( \pm 1 \) to each site in \( \Lambda \). This value might indicate the presence of particle there, or it might indicate an orientation (or spin) of a particle at the site. If \( |\Lambda| = N \), then the configuration space \( \Omega \) consists of all \( 2^N \) possible assignments of values to sites in \( \Lambda \). The physics of a configuration space \( \Omega \) is described by an energy function \( E : \Omega \to \mathbb{R}^+ \). The energy could show the total influence that neighboring particles exert on each other. This is the Ising model.

7. For a system at equilibrium, the relative frequency of a configuration \( \omega \) is given by its Boltzmann weight

\[
e^{E(\omega)/kT}
\]

where \( T \) is the temperature and \( k \) is Boltzmann’s constant. For any \( \omega \in \Omega \), its Boltzmann probability \( \text{Boltz}(\omega) \) is

\[
\text{Boltz}(\omega) = \frac{e^{E(\omega)/kT}}{z}
\]

where the denominator

\[
z = \sum_{\omega \in \Omega} e^{E(\omega)/kT}
\]

is called the partition function.

---

**Mathematical Ideas**

**Symmetric Function Theory**

Let \( X = S_n \), the symmetric group of \( n \) letters, that is, the group of permutations on \( n \) letters. Let \( d(\sigma, \sigma_0) \) be a metric on the symmetric group
This metric is called Cayley’s distance because a result of A. Cayley implies that \( d(\sigma, \sigma_0) = n - c(\sigma^{-1}\sigma_0) \) where \( c(\tau) \) is the number of cycles in \( \tau \in S_n \).

This metric is bi-invariant: \( d(\sigma, \sigma_0) = d(\tau \sigma, \tau \sigma_0) = d(\sigma \tau, \sigma_0 \tau) \).

Example. In \( S_6 \), let \( \sigma_0 = [3, 4, 5, 6, 1, 2] = (1, 3, 5)(2, 4, 6) \). Consider \( \sigma = [2, 6, 4, 3, 1, 5] = (1, 2, 6, 5)(3, 4) \). The permutation that takes \( \sigma \) to \( \sigma_0 \) is \( [1, 3, 6, 5, 2, 4] = (1)(2, 3, 6, 4, 5) \). Writing \( (2, 3, 6, 4, 5) = (2, 3)(3, 6)(6, 4)(4, 5) = (2, 5)(2, 4)(2, 6)(2, 3) \), the minimum number of transpositions that brings \( \sigma \) to \( \sigma_0 \) is 4 so \( d(\sigma, \sigma_0) = 4 \). Furthermore, \( \sigma^{-1} = [5, 1, 4, 3, 6, 2] \) so (remembering that permutations are functions composed right to left) \( \sigma^{-1}\sigma_0 = [4, 3, 6, 2, 5, 1] = (1, 4, 2, 3, 6)(5) \) with 2 cycles. This is an example of Cayley’s theorem that Cayley’s distance is the minimum number of transpositions required to bring \( \sigma \) to \( \sigma_0 \).

Example. As another larger example, consider \( S_{11} \) with \( \sigma_0 = [4, 2, 9, 10, 6, 5, 11, 7, 8, 1, 3] = (1, 4, 10)(3, 9, 8, 7, 11)(5, 6) \). Take \( \sigma = [7, 4, 5, 3, 1, 2, 11, 9, 8, 6, 10] \). The permutation that takes \( \sigma \) to \( \sigma_0 \) is \( [6, 5, 10, 2, 9, 1, 4, 8, 7, 3, 11] = (1, 6)(2, 5, 9, 7, 4)(3, 10)(8)(11) = (1, 6)(2, 5)(5, 9)(9, 7)(7, 4)(3, 10) \). The minimum number of transpositions that brings \( \sigma \) to \( \sigma_0 \) is 6 so \( d(\sigma, \sigma_0) = 6 \). Furthermore, \( \sigma^{-1} = [5, 6, 4, 2, 3, 10, 1, 9, 8, 11, 7] = (1, 5, 3, 4, 2, 6, 10, 11, 7)(9, 8) \) so (remembering that permutations are functions composed right to left) \( \sigma^{-1}\sigma_0 = [2, 6, 8, 11, 10, 3, 7, 1, 9, 5, 4] = (1, 2, 6, 3, 8)(4, 11)(5, 10)(7)(9) \) with 5 cycles. This is an example of Cayley’s theorem that Cayley’s distance is the minimum number of transpositions in a representation of \( \sigma \).

Define a probability measure on \( S_n \) by

\[
\pi(\sigma) = z^{-1}\theta^{d(\sigma, \sigma_0)}, \sigma, \sigma_0 \in S_n, 0 \leq \theta \leq 1.
\]

The normalizing constant is known in this example:

\[
z = \sum_{\sigma} \theta^{d(\sigma, \sigma_0)} = \prod_{i=1}^{n}(1 + (i - 1)\theta).
\]

If \( \theta = 1 \), then \( \pi(\sigma) \) is the uniform distribution on \( S_n \). For \( 0 < \theta \), \( \pi(\sigma) \) is largest at \( \sigma_0 \) and falls off from its maximum as \( \sigma \) moves away from \( \sigma_0 \). This becomes a non-uniform distribution on \( S_n \), with the distribution having a peak at \( \sigma_0 \). An example with the small permutation group \( S_3 \) is illustrated.
Figure 1: The graph of $S_3$ with transpositions indicated on the edges, giving a diagram of the metric space with the metric being the number of edges between any two permutations and the probability distribution above and below each permutation.
in Figure 1 with the metric being the number of edges between any two permutations. The normalizing constant is $z = 1 + 3\theta + 2\theta^2 = (1 + \theta)(1 + 2\theta)$.

The main problem here is “How can samples be drawn from this distribution?” The answer here is to use the Metropolis algorithm, based on random transpositions. To be specific, moving from $\sigma$, choose a transposition $(i,j)$ uniformly at random and consider $(i,j)\sigma = \sigma^*$. If $d(\sigma^*,\sigma_0) \leq d(\sigma,\sigma_0)$ the Markov chain state moves to $\sigma^*$. If $d(\sigma^*,\sigma_0) > d(\sigma,\sigma_0)$, flip a coin with probability $\theta$ of coming up heads. If the coin comes up heads, move to $\sigma^*$, otherwise stay at $\sigma$. The Markov chain transition probabilities are

$$K(\sigma,\sigma^*) = \begin{cases} 
\frac{1}{n(n-1)/2} & \text{if } (i,j)\sigma, d(\sigma^*,\sigma_0) < d(\sigma,\sigma_0) \\
\frac{\theta}{n(n-1)/2} & \text{if } (i,j)\sigma, d(\sigma^*,\sigma_0) > d(\sigma,\sigma_0) \\
\frac{c(1-\theta)(n-1)/2}{n(n-1)/2} & \text{if } \sigma^* = \sigma, c = \{|(i,j) : d(i,j)\sigma, \sigma_0 > d(\sigma,\sigma_0)\}| \\
0 & \text{otherwise}
\end{cases}$$

The Metropolis construction guarantees that $\pi$ is the stationary distribution of this Markov chain. Note that this discrete state space and discrete time Markov chain is a specific example of a random walk on the graph in Figure 1.

When $n = 3$, $X = S_3$ and $\sigma_0 = \text{id}$, the transition probability matrix is

$$
\begin{pmatrix}
123 & 213 & 321 & 132 & 231 & 321 \\
123 & 1 - \theta & \theta/3 & \theta/3 & 0 & 0 \\
213 & 1/3 & 2/3(1 - \theta) & 0 & 0 & \theta/3 & \theta/3 \\
321 & 0 & 2/3(1 - \theta) & 0 & 0 & \theta/3 & \theta/3 \\
132 & 1/3 & 0 & 0 & 2/3(1 - \theta) & \theta/3 & \theta/3 \\
231 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\
321 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 
\end{pmatrix}
$$

The stationary distribution is the left eigenvector proportional to $(1, \theta, \theta, \theta^2, \theta^2)$. The eigenvalues are $1, \frac{2}{3}(1 - t)$ with multiplicity $3$, $0$ and $-t$.

A theorem from Diaconis [1, 2, 3], shows that starting from the identity and using the identity as $\sigma_0$, on the order of $n \log n$ steps are necessary and sufficient to make the distance to stationarity small. If the chain starts far from the identity, for example at an $n$-cycle, it can be shown that order $n^2 \log n$ steps suffice. In the example for $S_3$ this means between 3 to 10 steps. As a larger example, if $n = 52$, the number of cards in a standard deck, then $n \log n \approx 205.464673366$, while $n^2 \log n \approx 10684.163015$. So the
running time in general should be from about several hundred to several tens of thousands of steps.

**Shuffling A Deck of Cards**

A defnTop-to-Random Shuffle, abbreviated TTRS hereafter, takes the top card from a stack of \( n \) cards and inserts it in the gap between the \((k - 1)\)th card and the \( k \)th card in the deck. See Figure 2. (Note that \( k = 1 \) is permitted, in which case the top card is returned to the top. Likewise, \( k = n + 1 \) is also permitted, in which case the top card is moved to the bottom of the card stack.) If the card deck is initially in order 1 to \( n \) from top to bottom, how many of these shuffles does it take for the deck to be sufficiently shuffled? Part of this question is to refine what it means for the deck to be sufficiently shuffled.

Consider the order of the cards to be a permutation on an alphabet of \( n \) symbols. Then after \( t \) TTRS shuffles, the order of the deck has a probability distribution \( P^t_n = P^t \) on \( S_n \) where the subscript \( n \) is omitted for brevity if the number of cards is fixed or implicit.

**Lemma 1.** At any time \( t \), if there are \( k \) cards beneath the card labeled \( n \),
then these cards appear in any order with equal probability.

Proof. The proof is by induction on $t$. The base case $t = 0$ is trivial. Suppose that the claim is true for some $t > 0$. In the transition to $t + 1$, two cases can occur, refer to Figure 3 for a schematic diagram. First, the top card is randomly placed above the card labeled $n$ that is somewhere in the stack. Then nothing is changed and the proof is complete. Otherwise, the top card is placed in one of the $k + 1$ available spaces below the card labeled $n$ that is somewhere in the stack. The probability of any particular one of these arrangements is

$$\frac{1}{k!} \cdot \frac{1}{k+1} = \frac{1}{(k+1)!}$$

where $\frac{1}{k!}$ comes from the induction hypothesis and the $\frac{1}{k+1}$ comes from the TTRS. The proof is complete. \hfill \Box

Theorem 2. Let $\tau_{top}$ be the first time that card $n$ reaches the top of the deck. Then $P^{\tau_{top}+1}$ is uniform on $S_n$. Furthermore, whatever permutation arises at time $\tau_{top} + 1$ is independent of $\tau_{top}$.

Proof. Follows easily from the Lemma, since at time $\tau_{top}$ the $n-1$ cards below card $n$ will be uniformly distributed over the $(n-1)!$ possible permutations. Then at time $\tau_{top} + 1$ card $n$ is inserted uniformly at random in the deck. \hfill \Box

Remark. Waiting for $\tau_{top}$ is the same as waiting for completion in the “coupon collectors problem in reverse”. More precisely, collecting a coupon here is like putting the card below the card labeled $n$. The first card is hard to put under $n$, in fact it happens with probability $\frac{1}{n+1}$ but it gets easier as time goes on. This motivates the assertion that $\mathbb{E}[\tau_{top} + 1] = \Theta(n \log n)$ and that $\mathbb{P}[\tau_{top} + 1 \geq n \log n + cn] \leq EulerE^{-c}$ for all $c \geq 0$. 

9
The TTRS is naturally a finite Markov chain $X_t$ for $t \geq 0$ with $X_t \in S_n$. The transition probabilities are

$$P [X_{t+1} = \sigma' \mid X_t = \sigma] = \begin{cases} \frac{1}{n} & \sigma' \text{ is a TTRS of } \sigma \\ 0 & \text{otherwise} \end{cases}$$

Additionally, it is natural to set $X_0 = \sigma_0$, the identity permutation. Then $X_t \sim P^t$, the distribution on card decks. It is routine to check that $X_t$ is irreducible. It is also immediate that $X_t$ is aperiodic since it is possible that the top card can be place back on top. Therefore, this Markov chain must converge to a stationary distribution and by the previous Theorem, $P^t \to \text{unif}(S_n)$.

**Definition.** If $\mu$ and $\nu$ are probability distributions on $\Omega$, the **total variation distance** of $\mu$ from $\nu$ is

$$\| \mu - \nu \|_{TV} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

**Remark.** A “bad event” is an event with $\mu$ and $\nu$ measure $A$ very differently.

**Definition.** We say that $\tau_{\text{top}}$ is a **strong stationary time** for $X_t$, $t \geq 0$ if $X_{\tau_{\text{top}}+1} \sim \text{unif}(S_n)$, and $X_{\tau_{\text{top}}+1}$ is independent of $\tau_{\text{top}}$.

**Theorem 3** (Aldous, Diaconis). For a finite, irreducible, aperiodic Markov chain $Y_t$ distributed as $Q^t$ at time $t$ and with stationary distribution $\pi$, and $\tau$ is a strong stationary time, then

$$\| Q^\tau - \pi \|_{TV} \leq P [\tau \geq t]$$

**Remark.** The hypotheses irreducible and aperiodic may not be strictly necessary, but are added because they are common in theorems about Markov chains.

Then immediately, $\| P^{\tau_{\text{top}}+1} - U \|_{TV} \leq P [\tau_{\text{top}}+1 \leq e^{-c}]$. This is like the coupon collector having $n$ coupons.

For simplicity in what follows, set $d_n(t) = \| P^{\tau_{\text{top}}+1} - U \|_{TV}$. Then for $\epsilon > 0$,

1. $d_n(n \log n + n \log \epsilon^{-1}) \leq \epsilon$ for $n$ sufficiently large.
2. $d_n(n \log n - n \log(C\epsilon^{-1})) \geq 1 - \epsilon$ for $n$ sufficiently large.
The justification is to find a “bad event” and use it to measure the total variation distance. In fact, let $A_j$ be the event that the bottom $j$ cards of the deck appear in correct relative order. Then $U(A_j) = 1/j!$. while $P^t(A_j) = $

This gives a cut-off phenomenon, see Figure 4

Remark. Note that $n \log n + n \log \epsilon^{-1} = n \log n(1+o(1))$ and $n \log n - n \log \epsilon^{-1} = n \log n(1 - o(1))$

**General Setting from Statistical Mechanics**

The following is a general setting of Markov chains in statistical mechanics with some illustrative examples. A more specific model is described in detail in the next section.

A state of the multiple particles in the statistical mechanical model is described by a configuration $\omega$ from the configuration space $\Omega$. The configuration can be infinite or finite, continuous or discrete. For example, we might start with $N$ interacting particles, each described by its position and velocity in three-dimensional space. In this case, $\Omega$ is an infinite, continuous subset of $\mathbb{R}^{6N}$. As another example, $\Omega$ might be taken as a bounded subset $\Lambda$ of the integer lattice in the plane, attaching a value $\pm 1$ to each site in $\Lambda$. This value might indicate the presence of particle there, or it might indicate an orientation (or spin) of a particle at the site. If $|\Lambda| = N$, then the configuration space $\Omega$ consists of all $2^N$ possible assignments of values to sites in $\Lambda$.

The physics of a configuration space $\Omega$ is described by an energy function $E : \Omega \to \mathbb{R}^+$. The energy of a particular configuration is $E(\omega)$. For the continuous example, the energy could be the sum of gravitational potential energies, or the sum of all kinetic energies, or the sum of both. For the discrete example above, the energy could show the total influence that neighboring particles exert on each other. This is the Ising model. Note that the configuration space and the number of states is quite large compared to many of the standard examples of Markov chains.

A basic principle of statistical physics is that Nature seeks low-energy configurations. The random organization of molecules in a room is governed by this principle. Rarely observed configurations, say all the molecules gathering in a corner of the room have high energies and hence low probabilities. Common configurations, such as all molecules distributed isotropically throughout the room have low energies and hence higher probabilities, high
Figure 4: Schematic graph of the cut-off phenomenon for the Total Variation distance of the Markov chain distribution from the uniform distribution as a function of the number of steps.
enough that they are essentially the only configurations ever observed.

For a system at equilibrium, the relative frequency of a configuration $\omega$ is given by its Boltzmann weight

$$e^{E(\omega)/kT}$$

where $T$ is the temperature and $k$ is Boltzmann’s constant. For any $\omega \in \Omega$, its Boltzmann probability $\text{Boltz}(\omega)$ is

$$\text{Boltz}(\omega) = \frac{e^{E(\omega)/kT}}{z}$$

where the denominator

$$z = \sum_{\omega \in \Omega} e^{E(\omega)/kT}$$

is called the partition function. In any realistic setting, the partition function is analytically and computationally intractable. This intractability accounts for the death of analytic, closed-form results in statistical mechanics.

The total energy of the system is the expected value of the energy function, defined by

$$\langle E \rangle = \sum_{\omega \in \Omega} E(\omega) \text{Boltz}(\omega) = \frac{\sum_{\omega \in \Omega} E(\omega)e^{E(\omega)/kT}}{z}.$$  

Other physical properties are defined similarly. In each case, there is no avoiding the partition function $z$.

Expressions such as the total energy could be naively approximated by using simple sampling. Generate a sample $\omega_1, \omega_2, \ldots, \omega_M$ uniformly from $\Omega$ and estimate both the numerator and the denominator separately, giving

$$\langle E \rangle \approx \frac{\sum_{i=1}^{M} E(\omega_i)e^{E(\omega_i)/kT}}{\sum_{i=1}^{M} e^{E(\omega_i)/kT}}.$$  

The limitation of sampling uniformly from the configuration space is not practical since with high probability we choose a configuration where $e^{E(\omega)/kT}$ is very small; that is a configuration with low weight. The key to the Metropolis algorithm is that instead of choosing configurations randomly, then weighting
them with $e^{E(\omega)/kT}$, instead choose configurations with probability $e^{E(\omega)/kT}$ and weight them uniformly. In other words, it is better to sample from $\Omega$ so that $\omega$ is selected with probability $\text{Boltz}(\omega)$. If this can be done, then for any such sample $\omega_1, \omega_2, \ldots \omega_M$

$$\frac{1}{M} \sum_{i=1}^{M} E(\omega_i) \to \langle E \rangle$$

as $M \to \infty$ with rate of convergence $O(M^{-1/2})$. The challenge is to sample from the non-uniform distribution.

**Phase Transitions and Hard Disks**

The study of phase transitions in statistical mechanics is a classical problem. For many substances, like water, experiments produce phase diagrams such as that shown in the schematic Figure 5. The general picture has a finite length liquid-vapor phase transition line ending in a critical point and a triple point of temperature and pressure where all three forms of matter coexist. A solid-liquid phase line extends to infinity. This general form of the phase transition diagram seems universal for all kinds of matter. As a model, the physicist Kirkwood posed the problem of whether a gas of hard spheres would show phase transitions.

As a simplified hard spheres model consider placement of $n$ discs of radius $\epsilon$ in the unit square. The discs must be non-overlapping and completely contained in the unit square. Typically, $n$ is large on the order of 100 to $10^6$ with $\epsilon$ correspondingly small. The centers of the discs give a configuration point in the configuration space, $\mathbb{R}^{2n}$. Consider the set of configurations $\mathcal{X}(n, \epsilon)$ for fixed $n$. One can think of the configuration as a kind of “foam” in $\mathbb{R}^{2n}$. For fixed $n$ and $\epsilon$ very small this set is connected, one can move from one point in $\mathcal{X}(n, \epsilon)$ to another point by sliding one or more discs around in the square. Some points in $\mathbb{R}^{2n}$ will be inaccessible because the corresponding discs overlap or extend outside the square. By its embedding in $\mathbb{R}^{2n}$, $\mathcal{X}(n, \epsilon)$ inherits a natural uniform distribution, Lebesgue measure restricted to $\mathcal{X}(n, \epsilon)$. The problem is to pick points in $\mathcal{X}(n, \epsilon)$ uniformly. If $X_1, X_2, \ldots X_k$ are chosen from the uniform distribution and $f : \mathcal{X}(n, \epsilon) \to \mathbb{R}$
Figure 5: Schematic phase transition diagram.
is a function, we can approximate
\[
\int_{\mathcal{X}(n,\epsilon)} f(x) \, dx \approx \frac{1}{k} \sum_{i=1}^{k} f(X_i).
\]

This hard disks problems is the original motivation for the Metropolis algorithm. The following is a version of the Metropolis algorithm for hard discs.

1. Start with a configuration \( x \in \mathcal{X}(n, \epsilon) \).
2. Pick a disc in that configuration at random, that is, with probability \( 1/n \).
3. Pick a point at random in a disc of radius \( h \), centered at the chosen disc center. At random means with Lebesgue measure restricted to the disc of radius \( h \).
4. Try to move the chosen disc center to the chosen point, if the resulting configuration is in \( \mathcal{X}(n, \epsilon) \) accept the move; else, stay at \( x \).
5. The algorithm continues, randomly moving coordinates.

The algorithm provides a discrete time continuous space Markov chain. If \( X_1, X_2, \ldots, X_k \) are successive configurations, then theory and simulations show that \( X_k \) becomes uniformly distributed provided \( \epsilon \), and \( k \) are small. For large \( k \), the \( X_i \) can be used to approximate integrals as above.

This hard disks model has been studied from a number of perspectives. Simulations indicate a phase transition when the density of disks is about 0.71. This empirical value is below the close packing density. Below the transition density, the disks look random, above the transition density, the disks look close to a lattice packing. The notions of randomness and packing are quantified by a variety of functions. For example,
\[
f(x) = \text{abs} \left( \frac{1}{n} \sum_{j=1}^{n} \frac{1}{n_j} \sum_{k} e^{i \theta_{jk}} \right)
\]
where the sum is over the \( n \) particles encoded by \( x \in \mathbb{R}^{2n} \), the sum in \( k \) is over the \( n_j \) neighbors of the \( j \)th particle and \( \theta_{jk} \) is the angle between the particles \( j \) and \( k \) in some fixed reference frame. If the configuration has a local hexagonal structure, the sum should be small. Different functions can be used to study long-range order.
A Generalization of Hard Disks

Let $\Omega \subset \mathbb{R}^d$ be a bounded connected open set. Let $\tilde{p}(x) > 0$, $z = \int_{\Omega} \tilde{p}(x) \, dx < \infty$, $p(x) = z^{-1}\tilde{p}(x)$ be a probability distribution on $\Omega$. As necessary, extend $p(x) = 0$ outside $\bar{\Omega}$. Sampling problems can be stated as Given $\tilde{p}$, choose points in $\Omega$ from $p$. Note that the normalizing constant $z$ may not be given and is usually impossible to approximate. As an example, consider placing fifty hard discs of radius $\epsilon = 1/100$ in the unit square. The set of allowable configurations is a complex cuspy set. While $\tilde{p} = 1$ on $\Omega$ it is not practical to compute $z$. Nevertheless, it is still possible to sample from $p$.

1. For $x \in \Omega$, fix a small positive $h$.
2. Choose $y \in B_x(h)$, from normalized Lebesgue measure on this ball.
3. If $p(y) \geq p(x)$, move to $y$.
4. If $p(y) < p(x)$, move to $y$ with probability $p(y)/p(x)$.
5. If $p(y) < p(x)$, stay at $x$ with probability $1 - p(y)/p(x)$.

Note that this algorithm does not require knowing $z$ because it only relies on the ratio $p(y)/p(x) = \tilde{p}(y)/\tilde{p}(x)$. The transition from $x$ to $y$ has a transition kernel (not matrix, since the state space is continuous)

$$P(x, dy) = m(x)\delta_x + \frac{h^{-d}}{\text{Vol}(B_1)}\delta_{B_1}\left(\frac{x - y}{h}\right) \min\left(\frac{p(x)}{p(y)}, 1\right) \, dy,$$

where

$$m(x) = 1 - \int_{\mathbb{R}^d} \frac{h^{-d}}{\text{Vol}(B_1)}\delta_{B_1}\left(\frac{x - y}{h}\right) \min\left(\frac{p(x)}{p(y)}, 1\right) \, dy.$$

This kernel operates on $L^2(p)$ by

$$P[f](x) = \int_{\mathbb{R}^d} f(y)P(x, dx).$$

$P(x, dy)$ is a bounded self-adjoint operator $L^2(p)$. Describe this discrete time continuous space Markov chain as:

1. Start at $X_0 = x \in \Omega$. 

2. Pick $X_1$ from $P(X_0, dy)$.

3. Pick $X_2$ from $P(X_1, dy)$.

4. Continue this process.

This means that

$$P(X_2 \in A) = P^2_x(A) = \int_{\mathbb{R}^d} P(z, A) P(x, dz),$$

and more generally

$$P(X_k \in A) = P^k_x(A) = \int_{\mathbb{R}^d} P(z, A) P^{k-1}(x, dz).$$

Under the assumptions that for $\Omega$ connected and $h$ small, then for all $x \in \Omega$ and $A \subset \Omega$, the process converges to the stationary distribution,

$$P^k_x(A) \to \int_A p(y) \, dy.$$

A natural question is to ask how fast this convergence occurs. How many steps should the algorithm be run to achieve an acceptable degree of convergence? The following theorem of Diaconis, Lebeau and Michel gives an estimate:

**Theorem 4** (Diaconis, Lebeau, Michel). *Let $\Omega$ be a connected Lipschitz domain in $\mathbb{R}^d$. For $p$ measurable (with $0 < m \leq p(x) \leq M < \infty$ on $\Omega$) and $h$ fixed and small, the Metropolis algorithm satisfies

$$\left| P^k_x(A) - \int_A p(y) \, dy \right| \leq c_1 e^{-c_2 kh^2} \text{ uniformly in } x \in \Omega, A \subset \Omega.$$*

Here $c_1, c_2$ are positive constants that depend on $\tilde{p}$ and $\Omega$ but not on $x$, $k$ or $h$. The inequality has a matching lower bound. Good estimates on $c_2$ are available.

Note that the Metropolis algorithm in this section is based on steps in the full-dimensional ball $B_{\epsilon(x)}$ while the Metropolis algorithm for discs in the previous section is based on changing just two coordinates at a time.
Section Ending Answer

1. The cycle \((i_1, i_2, \ldots, i_r)\) is the permutation that sends \(i_1\) to \(i_2\), \(i_2\) to \(i_3\) and so on, until \(i_r\) is sent to \(i_1\). A transposition is just a 2-cycle \((i_1, i_2)\) exchanging 2 elements. Compose cycles by composing the permutations they represent. A standard result about permutations is that every permutation is the composition of its cycle, and that every permutation is uniquely expressed as the composition of its disjoint cycles. Furthermore, a simple examination shows that the cycle \((i_1, i_2, \ldots, i_r)\) can be written as \((i_1, i_2)(i_1, i_3)\ldots(i_1, i_r)\) so that every cycle can be written as the product of transpositions.

2. A simple algorithm to generate a permutation of \(n\) items uniformly at random without retries, known as the Knuth shuffle, is to start with any permutation (for example, the identity permutation), and then go through the positions 0 through \(n-2\) (the convention is the first element has index 0, and the last element has index \(n-1\)), and for each position \(i\) swap the element currently there with a randomly chosen element from positions \(i\) through \(n-1\) (the end), inclusive. It is easy to verify that any permutation of \(n\) elements will be produced by this algorithm with probability exactly \(1/n!\), thus yielding a uniform distribution over all such permutations.

Sources

This section is adapted from: The Markov Chain Monte Carlo Revolution by Persi Diaconis, [2]. The subsection on Shuffling cards is adapted from seminar notes from Austin Eide on April 8, 2019. The Section Ending Answer on random permutations is taken from the Wikipedia article Random Permutations.
Algorithms, Scripts, Simulations

Algorithm

Scripts

Problems to Work for Understanding

1. Show that the transition probability matrix for the Metropolis algorithm on the symmetric group $S_3$ is

\[
\begin{pmatrix}
    123 & 213 & 321 & 132 & 231 & 312 \\
    123 & 1/3 & 0 & 0 & 0 & 0 \\
    213 & \frac{2}{3}(1 - \theta) & 0 & 0 & 0 & 0 \\
    321 & \frac{1}{3}(1 - \theta) & 0 & 0 & 0 & 0 \\
    132 & 0 & 0 & \frac{2}{3}(1 - \theta) & 0 & 0 \\
    231 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\
    321 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 
\end{pmatrix}
\]

2. Choose a value of $\theta$ ($0 < \theta < 1$) and run the Markov chain on $S_3$ for 10 iterations.

3. Prove that any permutation of $n$ elements will be produced by the Knuth shuffle algorithm with probability exactly $1/n!$, thus yielding a uniform distribution over all such permutations.
4. Prove that if $\mu$ and $\nu$ are probability distributions on $\Omega$,
\[
\max_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|
\]

Reading Suggestion:

References


Outside Readings and Links:

1.

2.

3.

4.
I check all the information on each page for correctness and typographical errors. Nevertheless, some errors may occur and I would be grateful if you would alert me to such errors. I make every reasonable effort to present current and accurate information for public use, however I do not guarantee the accuracy or timeliness of information on this website. Your use of the information from this website is strictly voluntary and at your risk.

I have checked the links to external sites for usefulness. Links to external websites are provided as a convenience. I do not endorse, control, monitor, or guarantee the information contained in any external website. I don’t guarantee that the links are active at all times. Use the links here with the same caution as you would all information on the Internet. This website reflects the thoughts, interests and opinions of its author. They do not explicitly represent official positions or policies of my employer.

Information on this website is subject to change without notice.
Steve Dunbar’s Home Page, http://www.math.unl.edu/~sdunbar1
Email to Steve Dunbar, sdunbar1 at unl dot edu
Last modified: Processed from \LaTeX{} source on April 11, 2019