Markov Chains Continued

Apurva Nakade

2025-03-02

Table of contents

We'll continue our discussion on Markov Chains and study a few more theoretical results.

Sampling from a Markov Chain

The algorithm to generate sample paths of length n of a Markov Chain is simple. Suppose P is the transition matrix of the Markov Chain with mixing time t. The algorithm is as follows:

1. Start at some initial state X_0.
2. For $i = 0, 1, \dots, N-1$
    - Generate $X_{i+1} \sim P(X_i, \cdot)$.
3. Discard the first T samples and return $X_{T+1}, X_{T+2}, \dots, X_N$.

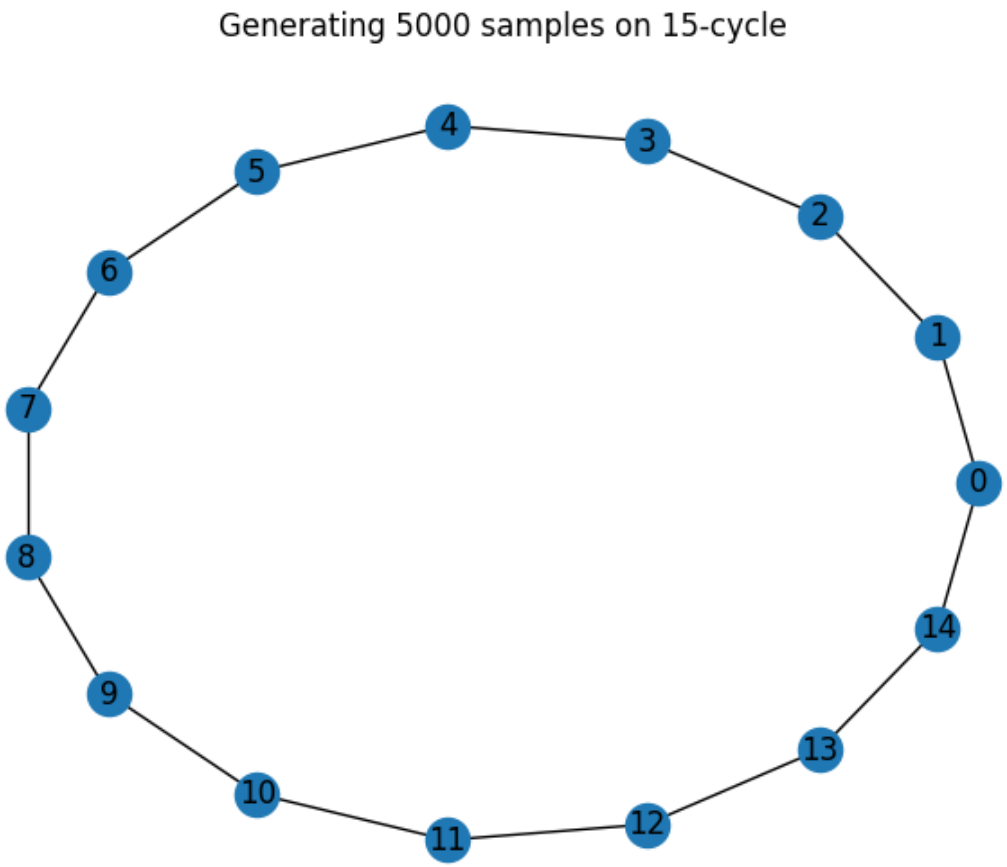
We can interpret this algorithm as generating $N-T$ samples from the stationary distribution of the Markov Chain. The number of samples discarded is called the **burn-in period**.

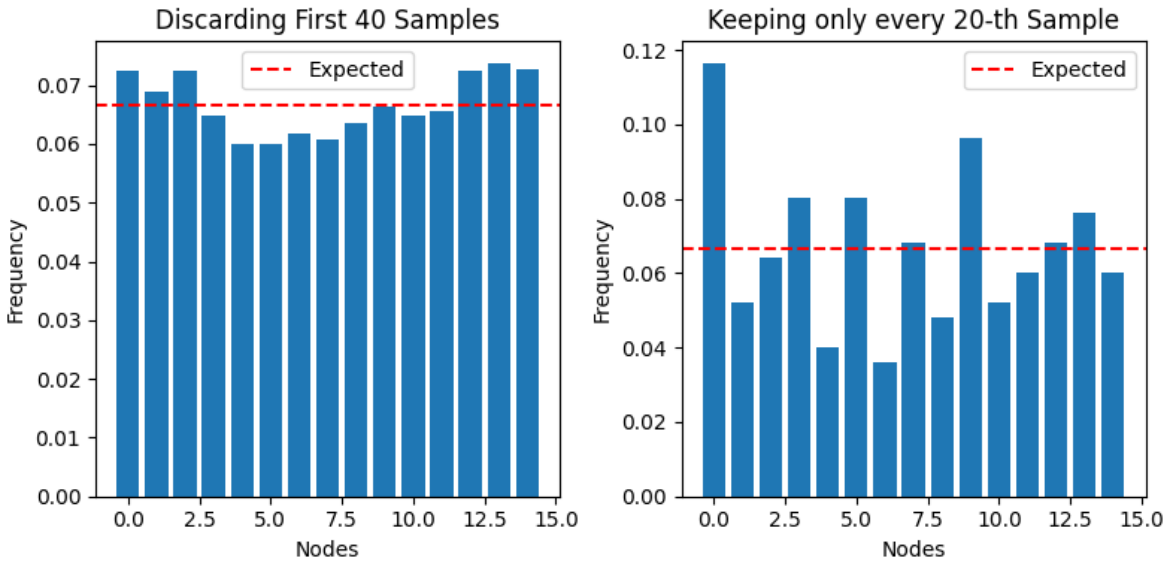
One big issue with this algorithm is that the samples are highly correlated. If independence is important, selecting every 40th sample may be beneficial. However, this leads to a lot of wasted samples and it might not get rid of all the correlation. In practice, it is better to generate a large number of samples than to "thin" the samples.

Example 0.1. In the example below we generate a uniform distribution over $[0, 14]$ by generating a random walk over a cycle of length 15.

The autocorrelation plot below shows the correlation between samples at different lags i.e. X_i and X_{i+k} for different values of k. We can see that the correlation decreases as k increases and stabilizes around $k = 40$.

We can either discard the first 40 samples or select every 40-th sample to reduce the correlation. If independence is important, then we should select every 40-th sample. However, this leads to a lot of wasted samples and it might not get rid of all the correlation. This method is not preferred in practice. In practice, it is better to generate a large number of samples than to "thin" the samples.





Reversible Markov Chains

A Markov Chain is **reversible** with respect to a distribution π if the following holds:

$$\pi_i P(i,j) = \pi_j P(j,i) \quad \text{for all } i,j. \tag{1}$$

This is saying that the probability of transitioning from x to y is the same as the probability of transitioning from y to x. Equation 1 is known as the **detailed balance equation**.

Theorem 0.1. ***(Detailed Balance Equation).*** *If a Markov Chain is reversible with respect to a distribution π, then π is a stationary distribution of the Markov Chain.*

Proof. Suppose the Markov Chain is reversible with respect to π. Then,

$$\begin{aligned}(\pi P)_i &= \sum_j \pi_j P(j,i) \\ &= \sum_j \pi_i P(i,j) \\ &= \pi_i \sum_j P(i,j) \\ &= \pi_i.\end{aligned}$$

Thus, π is the stationary distribution of the Markov Chain.

□

Equation 1 is a sufficient but not necessary condition for π to be the stationary distribution of the Markov Chain. You can have a Markov Chain with a stationary distribution that is not reversible.

Note that we did not use any properties of the Markov Chain in the proof of the theorem, except that the row sum of the transition matrix is 1. A better way to phrase this theorem would be to say that "if the row sum of the transition matrix is 1 and the detailed balance equation holds, then π is an eigenvector of the transition matrix with eigenvalue 1."

Example 0.2. Random Walks on Graphs. Consider a graph $G = (V, E)$ with vertices V and edges E. Let $P(i, j) = 1/d(i)$ if $(i, j) \in E$ and 0 otherwise, where $d(i)$ is the degree of vertex i. Then, the stationary distribution of the Markov Chain is $\pi_i = d(i)/2|E|$, where $|E|$ is the number of edges in the graph.

The Markov Chain is reversible with respect to π. Consider two vertices i and j. If $(i, j) \in E$, then

$$\begin{aligned}\pi_i P(i,j) &= \frac{d(i)}{2|E|} \cdot \frac{1}{d(i)} \\ &= \frac{1}{2|E|} \\ &= \frac{d(j)}{2|E|} \cdot \frac{1}{d(j)} \\ &= \pi_j P(j,i).\end{aligned}$$

If $(i, j) \notin E$, then $\pi_i P(i, j) = \pi_j P(j, i) = 0$.

Many Markov chains encountered in practice are reversible with respect to some distribution. It is much easier to check the detailed balance equation than to compute the stationary distribution directly. Moreover, reversible markov chains can be analyzed using spectral methods and we can find good bounds on their mixing time.

Reversibility and Symmetry

Theorem 0.2. *If a Markov Chain is reversible with respect to a distribution π, then the matrix*

$$Q = diag(\sqrt{\pi})\ P\ diag(\sqrt{\pi^{-1}})$$

is symmetric. In particular, P is similar to the symmetric matrix Q and hence has real eigenvalues.

Proof. This is because

$$\begin{aligned} Q(i,j) &= \sqrt{\pi_i} P(i,j) \sqrt{\pi_j^{-1}} \\ &= \sqrt{\pi_i} \frac{\pi_j P(j,i)}{\pi_i} \sqrt{\pi_j^{-1}} \\ &= \sqrt{\pi_j} P(j,i) \sqrt{\pi_i^{-1}} \\ &= Q(j,i). \end{aligned}$$

Similarly, $Q(j,i) = \pi_j P(j,i)$. As the Markov Chain is reversible with respect to π, we get $Q(i,j) = Q(j,i)$. Thus, Q is symmetric.

□

Now suppose $\mathbf{v}$ is a left eigenvector of Q with eigenvalue λ. Then,

$$\begin{aligned} \mathbf{v}Q &= \lambda \mathbf{v} \\ \mathbf{v}\text{diag}(\sqrt{\pi})\, P\, \text{diag}(\sqrt{\pi^{-1}}) &= \lambda \mathbf{v} \\ \implies \mathbf{v}\text{diag}(\sqrt{\pi})\, P &= \lambda \mathbf{v}\, \text{diag}(\sqrt{\pi}). \end{aligned}$$

Hence, $\text{diag}(\sqrt{\pi})\mathbf{v}$ will be an eigenvector of P with the same eigenvalue. As Q is symmetric, it has real eigenvalues and orthogonal eigenvectors. It is easier to do spectral analysis of Q and use that to deduce properties of P. For example, we can find the eigenvector corresponding to the largest eigenvalue of Q by solving the optimization problem

$$\mathbf{v} = \arg\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T Q \mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

Multiplying the above vector $\mathbf{v}$ by $\text{diag}(\sqrt{\pi})$ gives us the stationary distribution for P.

Transition Kernels

The sample spaces we encounter in MCMC methods are not discrete but continuous. Instead of a transition matrix, we use a **transition kernel** $K(x,y)$ that gives the "probability of transitioning from state x to state y". However, because the sample space is continuous, the probability of being in a particular state is zero. Instead, we use the **density** of the distribution at that point. The transition kernel satisfies the equation:

$$\mathbb{P}(X_{n+1} \in A \mid X_n = x) = \int_A K(x, y)\, dy.$$

All the properties of Markov Chains that we discussed earlier can be extended to transition kernels. The fundamental theorem of Markov Chains becomes

Theorem 0.3. ***(Fundamental Theorem of Markov Chains for Kernels).*** *If a Markov Chain with transition kernel K is*

1. ***Irreducible****: For all x, y, there exists n such that $K^n(x, y) > 0$.*
2. ***Positive recurrent****: The expected return time to a state is finite.*
3. ***Aperiodic****: For all x, $\gcd\{n : K^n(x, x) > 0\} = 1$.*

Then, the Markov Chain has a unique stationary distribution Π and for all initial distributions π_0,

$$\int \pi_0(x) K^n(x, y)\, dx \xrightarrow[TV]{} \Pi(y) \quad \text{as } n \to \infty.$$

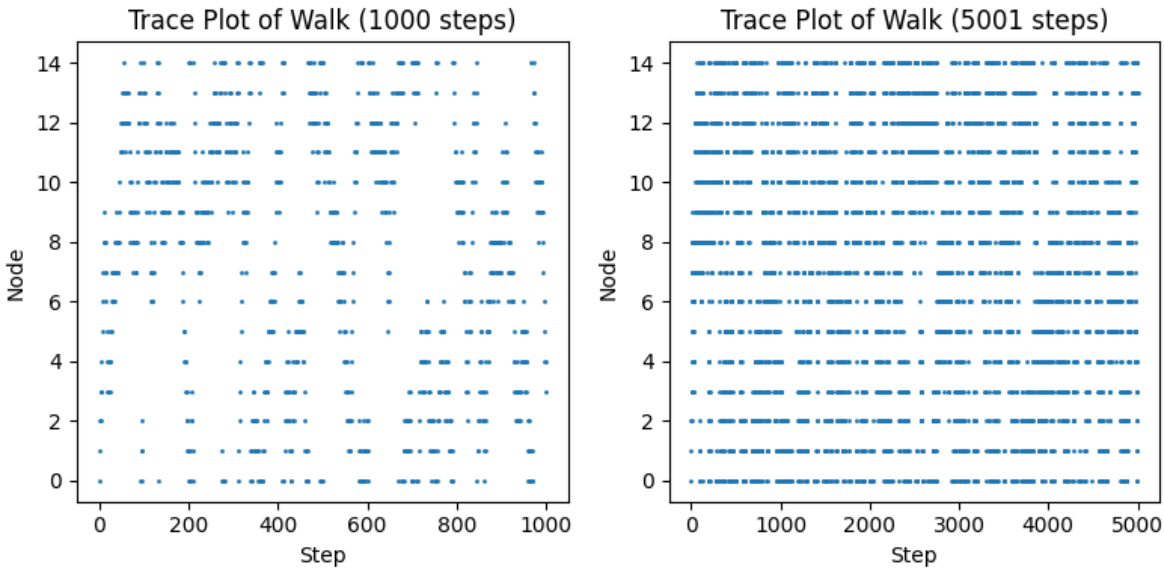
Analyzing Convergence of Markov Chains

In practice we need to decide how many samples to burn and how many samples to generate. We use various heuristics to decide this.

Trace Plots

A trace plot is simply a scatter plot of the samples generated by the Markov Chain. It is useful to see if the Markov Chain has converged. If the Markov Chain has converged, the trace plot should look like a cloud of points. If the Markov Chain has not converged, the trace plot will show a trend.

Below are the trace plots for Example 0.1. We can see that the chain does not look uniform even after a 1000 samples but after 5000 samples it is starting to look uniform.

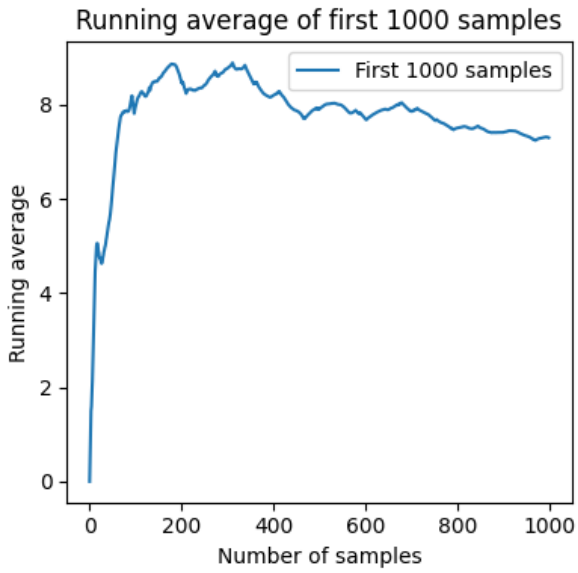


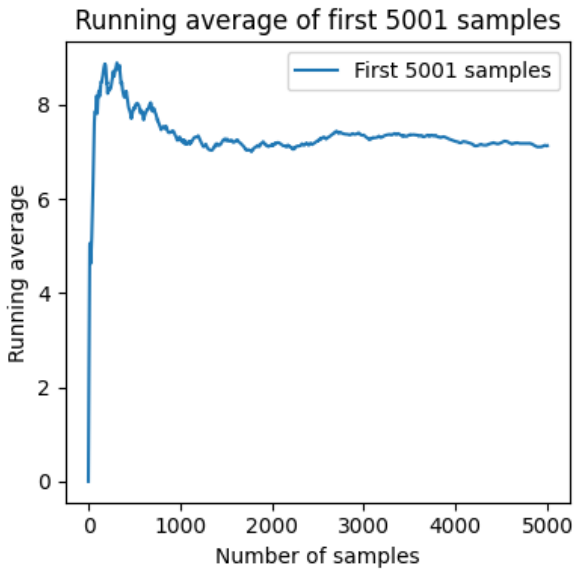
Running Average

The running average is the average of the first n samples. It is useful to see if the Markov Chain has converged. If the Markov Chain has converged, the running average should stabilize around the true mean. If the Markov Chain has not converged, the running average will show a trend.

Below is the running average plot for Example 0.1. We can see that the running average is stabilizing around the true mean around 3000 samples.

Running averages can be deceptive and show stability even when the Markov Chain has not converged. It is better to use multiple diagnostics to check for convergence. They are better at telling when the Markov Chain has **not** converged than when it has converged.





Autocorrelation Function

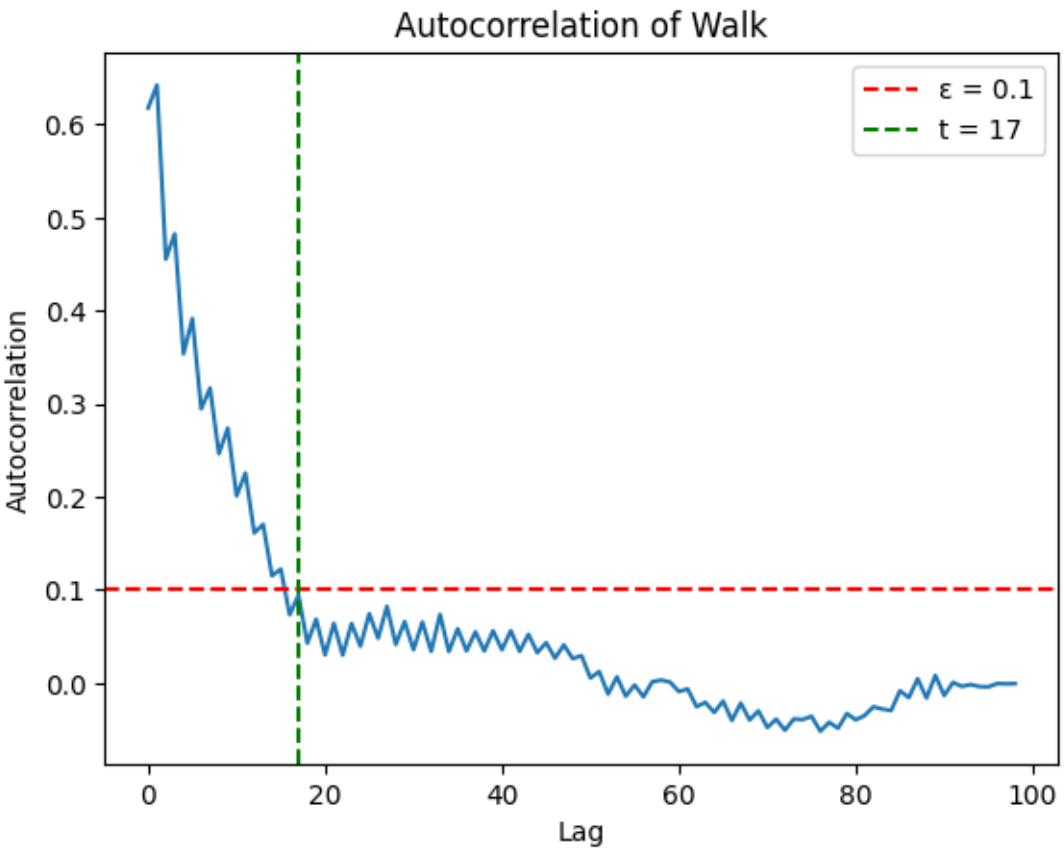
One method for finding the burn-in period is to use the autocorrelation function. The **autocorrelation function** of a sequence of numbers $x = (x_0, x_1, \dots, x_n)$ at lag k is defined as

$$\text{ACF}(k) = \text{Corr}(x[k:], x[:-k])$$

where by $x[k:]$ we mean the subsequence $x_k, x_{k+1}, \dots, x_n$ and by $x[:-k]$ we mean the subsequence $x_0, x_1, \dots, x_{n-k}$. It is the correlation between the sequence x and the same sequence shifted by k. One way to choose a burn-in period is to pick a threshold ϵ and find the smallest T such that $\text{ACF}(T) < \epsilon$. Then to be conservative choose a burn-in period of $2T$ or $3T$.

Below is the autocorrelation plot for Example 0.1. We can see that the correlation decreases as k increases and drops below 0.1 around 17. So a burn-in period of 40 is a good conservative choice.

This is just one way to choose the burn-in period. There are many other methods and the choice depends on the application. We will stick to this method for the rest of the course.


Autocorrelation of Walk
ε = 0.1
t = 17
Autocorrelation
Lag
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
20
40
60
80
100