The Metropolis-Hastings Algorithm
(Chapter 11)

Michael Tsiang

Stats 102C: Introduction to Monte Carlo Methods

UCLA

Acknowledgements: Qing Zhou

Outline

Introduction

The objective of Markov Chain Monte Carlo (MCMC) methods (just like classical Monte Carlo methods) is to simulate random variables. We are interested in a target distribution $\pi(x)$.

Main Goals:

$\pi(x) =$ limiting distribution

- To generate a sequence of correlated samples from $\pi(x)$.
- To estimate $E_\pi[h(X)] = \int h(x)\pi(x)\,\mathrm{d}x$.

Note: MCMC methods are often used to sample from multivariate distributions, so x could represent a vector in $\mathbb{R}^d$, for $d \geq 1$.

Stationary and Limiting Distributions

From the previous chapter:

- Let $\{X_0, X_1, X_2, \ldots,\}$ be a Markov chain with state space $\{0, 1, 2, \ldots, N\}$ and transition probabilities

$$P_{ij} = P(X_{n+1} = j | X_n = i).$$ (MCMC)

- If $\pi = (\pi_0, \pi_1, \pi_2, \ldots, \pi_N)$ is a stationary distribution of the Markov chain, then it satisfies the equations

$$\pi_j = \sum_{k=0}^{N} \pi_k P_{kj}.$$

has solution

- If the Markov chain (determined by the transition probabilities) is irreducible and aperiodic, then π is the limiting distribution of the Markov chain:

$$\lim_{n\to\infty} P(X_n = j | X_0 = i) = \pi_j, = \sum_{k=0}^{N} \pi_k P_{kj}$$

for every i and j.

Transition Kernel

- We will now generalize the state space of the Markov chain to be discrete or continuous.

n- steps / times

- For discrete-state Markov chains, the one-step transition probabilities are described by the transition matrix $\mathbb{P} = [P_{ij}]$.

one step in finite state

- Each row of $\mathbb{P}$, i.e., $P_{i\cdot} := P(X_{t+1}|X_t = i)$, can be thought of as the conditional distribution of $X_{t+1}|X_t = i$.

one step in infinite states

- For continuous-state Markov chains, the conditional distribution of $X_{t+1}|X_t = x$ is called the transition kernel.

$P(X_{t+1}|X_t = x)$

Definition

The **transition kernel** for a Markov chain $\{X_t : t = 0, 1, 2, \ldots\}$ is the conditional density $K(x, y)$ of $Y|X = x$.

Example 1: Uniform Transition Kernel

$X_0 = 0 \begin{cases} 0+1 \rightarrow x_1 \\ 0-1 \rightarrow x_1 \end{cases}$

- Define the initial state $X_0 = 0$ and transition kernel

$$K(x, y) \sim \text{Unif}(x - \delta, x + \delta), \quad \text{with } \delta = 1.$$

- We want to simulate $n = 10000$ steps of the Markov chain with this transition kernel.

- For $t = 1, 2, \ldots, n$: Generate $X_t \sim K(X_{t-1}, y)$.

Example 1: Uniform Transition Kernel

R Code to simulate a Markov chain with uniform transition kernel:

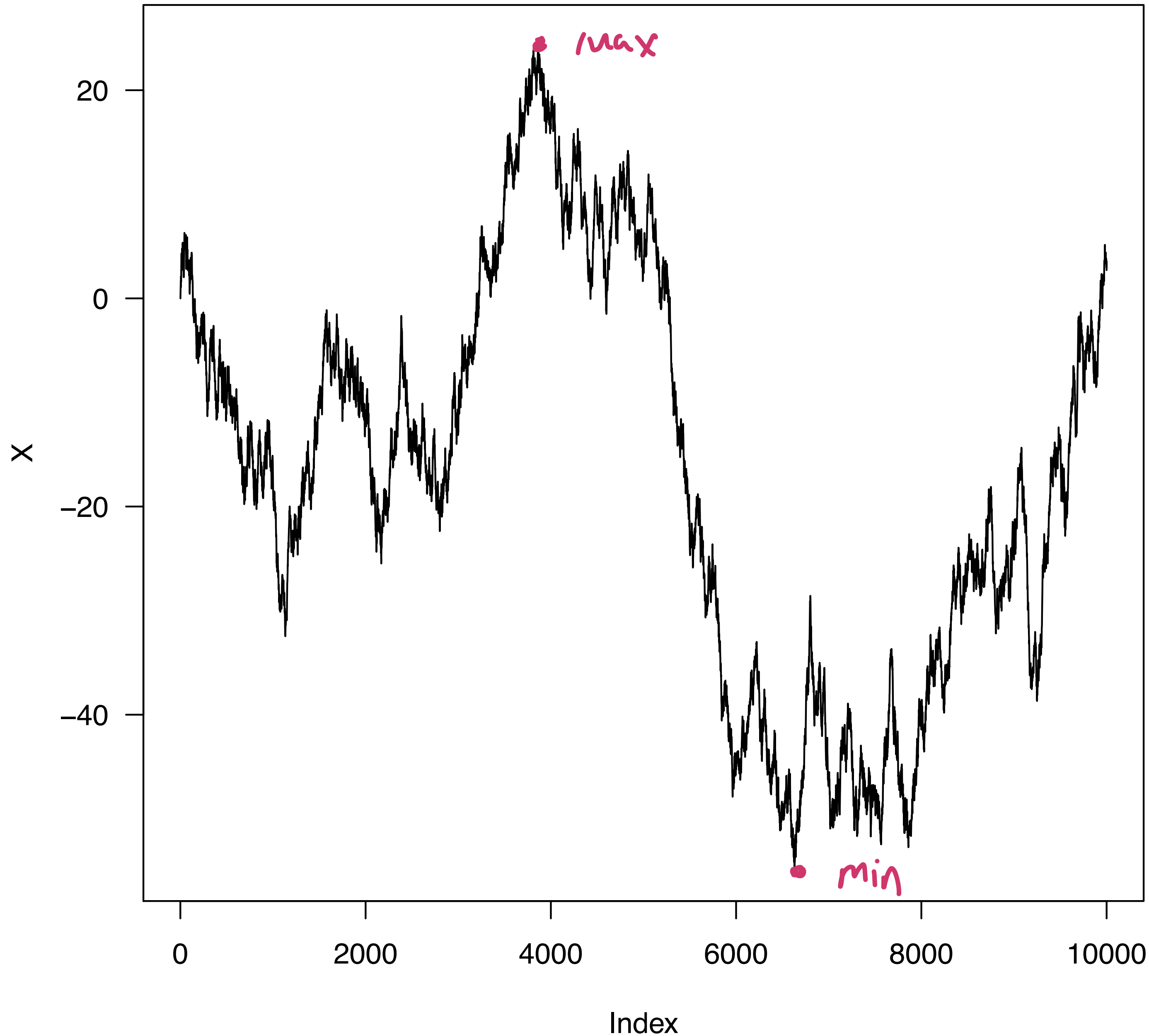
```
> set.seed(9999) # for reproduceability

> X <- 0 # specify initial state
> n <- 10000 # specify length of chain
> delta <- 1 # specify width parameter

> for (t in 2:n) {
+     # Take next step of Markov chain
+     X[t] <- runif(1, X[t - 1] - delta, X[t - 1] + delta)
+ }

> # Plot the Markov chain over time
> plot(X, type = "l", las = 1)
```

Example 1: Uniform Transition Kernel



Definition

Let $\{X_0, X_1, X_2, \ldots, \}$ be a Markov chain with state space $\mathcal{X}$ and transition kernel $K(\overset{x+1}{x}, \overset{x-1}{y})$. A distribution $\pi(x)$ on $\mathcal{X}$ is a stationary distribution of the Markov chain if it satisfies the **global balance** equations

p(y|x)

$$\pi(y) = \int \pi(x) K(x, y)\, \mathrm{d}x, \quad \text{for all } y \in \mathcal{X}.$$ (continuous)

If $\mathcal{X}$ is discrete, the global balance equations can be written as

(discrete)

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) K(x, y), \quad \text{for all } y \in \mathcal{X}.$$

uniform outcome

- The discrete case coincides with the equations from the previous chapter ($\pi = \pi\mathbb{P}$).
- The global balance equations imply that the total flow out of each state is balanced with the total flow into each state.

Introduction

MCMC methods use Markov chain theory in reverse:

- The stationary distribution of the Markov chain is the target distribution $\pi(x)$ from which we want to sample.
- We want to find a transition kernel $K(x, y)$ such that the corresponding Markov chain has $\pi(x)$ as its stationary distribution.
- Once we find a suitable transition kernel, we can simulate the Markov chain for a large number of steps (the **burn-in period**) until (approximate) convergence. The simulated observations after convergence are approximately from $\pi(x)$.

Outline

Generalization of Rejection Sampling

- Consider a target distribution $\pi(x)$ that we cannot sample from directly. We want to use MCMC to generate samples from $\pi(x)$.
- Recall rejection sampling: We sampled from a trial distribution $g(x)$ and used a rejection criterion $r(x)$ to decide whether to accept or reject the sample as from the target distribution.
- The Metropolis-Hastings algorithm: Based on the previous state of a Markov chain $X_{t-1} = x$, we sample from a proposal distribution $q(x, y)$ and use a rejection criterion $r(x, y)$ to decide whether to accept or reject the sample as the next step in the Markov chain.

The Metropolis-Hastings Algorithm

Goal: Generate $X \sim \pi(x)$.

Given a random initial state $x^{(0)}$ and a **proposal distribution** $q(x, y)$, then:

For $t = 1, 2, \ldots, n$:

1. Generate $y \sim q(x^{(t-1)}, y)$.
2. Compute the **Metropolis-Hastings ratio**
$$r(x^{(t-1)}, y) = \min\left[1, \frac{\pi(y)}{\pi(x^{(t-1)})} \cdot \frac{q(y, x^{(t-1)})}{q(x^{(t-1)}, y)}\right].$$
3. Generate $U \sim \mathrm{Unif}(0, 1)$ and update
$$x^{(t)} = \begin{cases} y & \text{if } U \leq r(x^{(t-1)}, y) \\ x^{(t-1)} & \text{otherwise.} \end{cases}$$

The Metropolis Algorithm (Symmetric Proposal)

A or B

- The algorithm was first proposed in Metropolis et al. 1953[1] for symmetric proposal distributions: $q(x, y) = q(y, x)$. A or B

- The Metropolis-Hastings algorithm with symmetric proposal is sometimes called the **Metropolis algorithm**.

- Example: For $\delta > 0$, the proposal $q(x, y) \sim \text{Unif}(x - \delta, x + \delta)$ is symmetric, since

$\frac{1}{x+\delta-(x-\delta)} = \frac{1}{2\delta}$

$$\boxed{q(y, x) = \frac{1}{2\delta} = q(x, y).}$$

- When $q(x, y) = q(y, x)$, then the Metropolis-Hastings ratio becomes

MH ratio

$$r(x, y) = \min\left[1, \frac{\pi(y)}{\pi(x)}\right] = \begin{cases} 1 & \text{if } \pi(y) \geq \pi(x), \\ \frac{\pi(y)}{\pi(x)} & \text{if } \pi(y) < \pi(x). \end{cases}$$

[1] http://dx.doi.org/10.1063/1.1699114

Example 2: Standard Normal Distribution

trail → uniform

- The standard normal distribution $\mathcal{N}(0,1)$ has PDF

f(x)

$$\pi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{for } x \in (-\infty, \infty).$$

$\pi_0 =$

- Consider the proposal distribution $\mathrm{Unif}(x-\delta, x+\delta)$, with $\delta = 1$, so

→ q(x)

$\frac{1}{2\delta} = \frac{1}{2}$

$$q(x,y) = \begin{cases} \frac{1}{2} & \text{for } y \in (x-1, x+1) \\ 0 & \text{otherwise.} \end{cases}$$

- We want to implement the Metropolis-Hastings algorithm to sample from $\pi(x) \sim \mathcal{N}(0,1)$.

$$r(x^{(t-1)}, y) = \min\left[1, \frac{\pi(y)}{\pi(x^{(t-1)})} \cdot \frac{q(y, x^{(t-1)})}{q(x^{(t-1)}, y)}\right.$$

$= \frac{1}{2}$

$= \frac{1}{2}$

Example 2: Standard Normal Distribution

- The proposal is symmetric, so the Metropolis-Hastings ratio is

$q(x,y)=q(y,x)$

$$
\begin{aligned}
r(x,y) &= \min\left[1, \frac{\pi(y)}{\pi(x)}\right] \\
&= \min\left[1, \frac{\frac{1}{\sqrt{2\pi}}e^{-y^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}\right] \\
&= \min\left[1, \frac{e^{-y^2/2}}{e^{-x^2/2}}\right] \\
&= \min\left[1, e^{-\frac{1}{2}(y^2-x^2)}\right].
\end{aligned}
$$

$x = x^{t-1}$

- Notice that $\pi(x)$ can be unnormalized, since the normalizing constant cancels out.

Example 2: Standard Normal Distribution

Metropolis algorithm to sample from $\pi(x) \sim \mathcal{N}(0,1)$:

Start from a random initial state $x^{(0)}$.

For $t = 1, 2, \ldots, n$:

1. Generate $y \sim \mathrm{Unif}(x^{(t-1)} - 1, x^{(t-1)} + 1)$.
2. Compute the Metropolis-Hastings ratio
$$r(x^{(t-1)}, y) = \min\left[1, e^{-\frac{1}{2}[y^2 - (x^{(t-1)})^2]}\right].$$
3. Generate $U \sim \mathrm{Unif}(0,1)$ and update
$$x^{(t)} = \begin{cases} y & \text{if } U \le r(x^{(t-1)}, y) \\ x^{(t-1)} & \text{otherwise.} \end{cases}$$

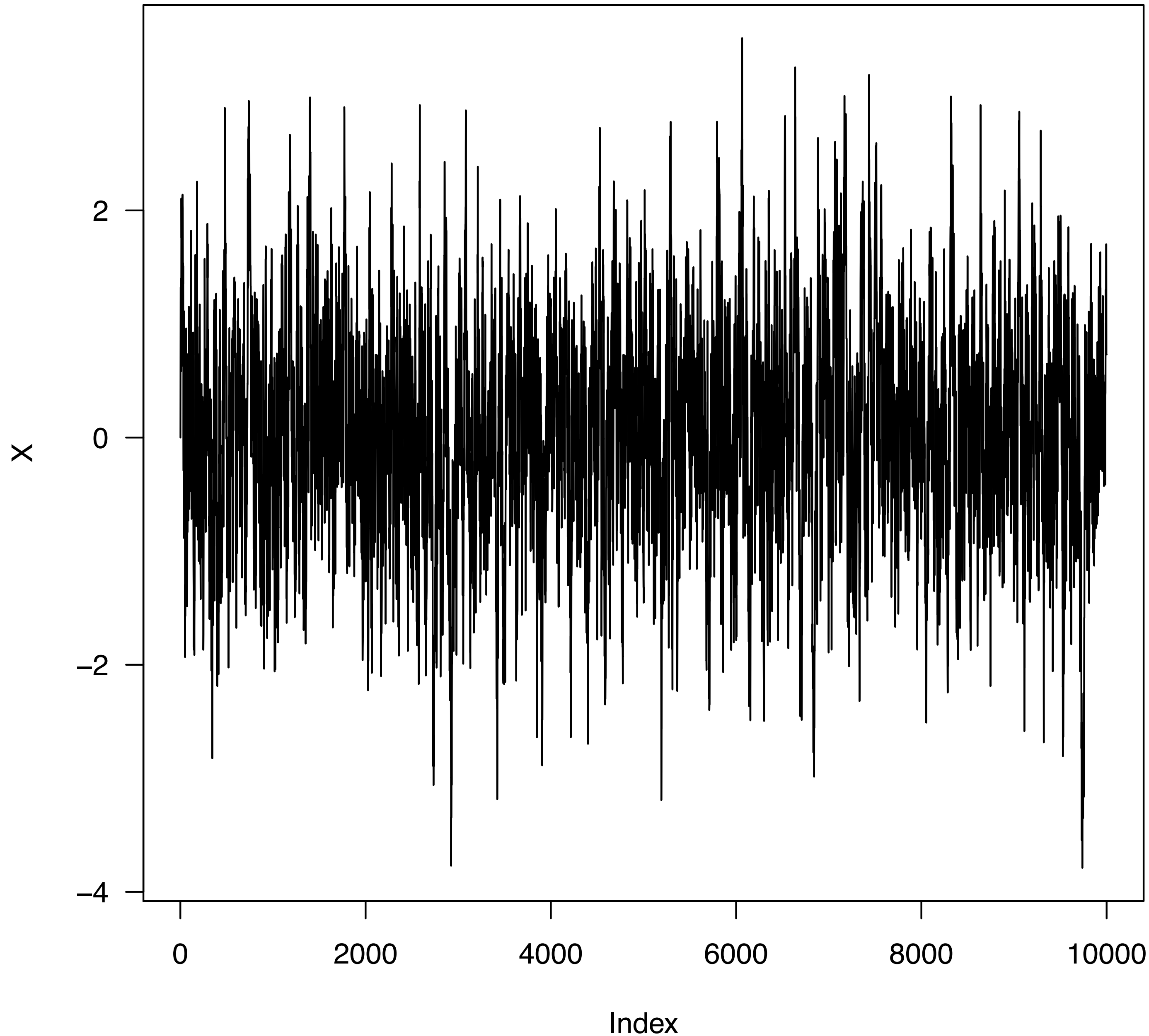
Example 2: Standard Normal Distribution

R Code for Metropolis algorithm to sample from $\mathcal{N}(0, 1)$:

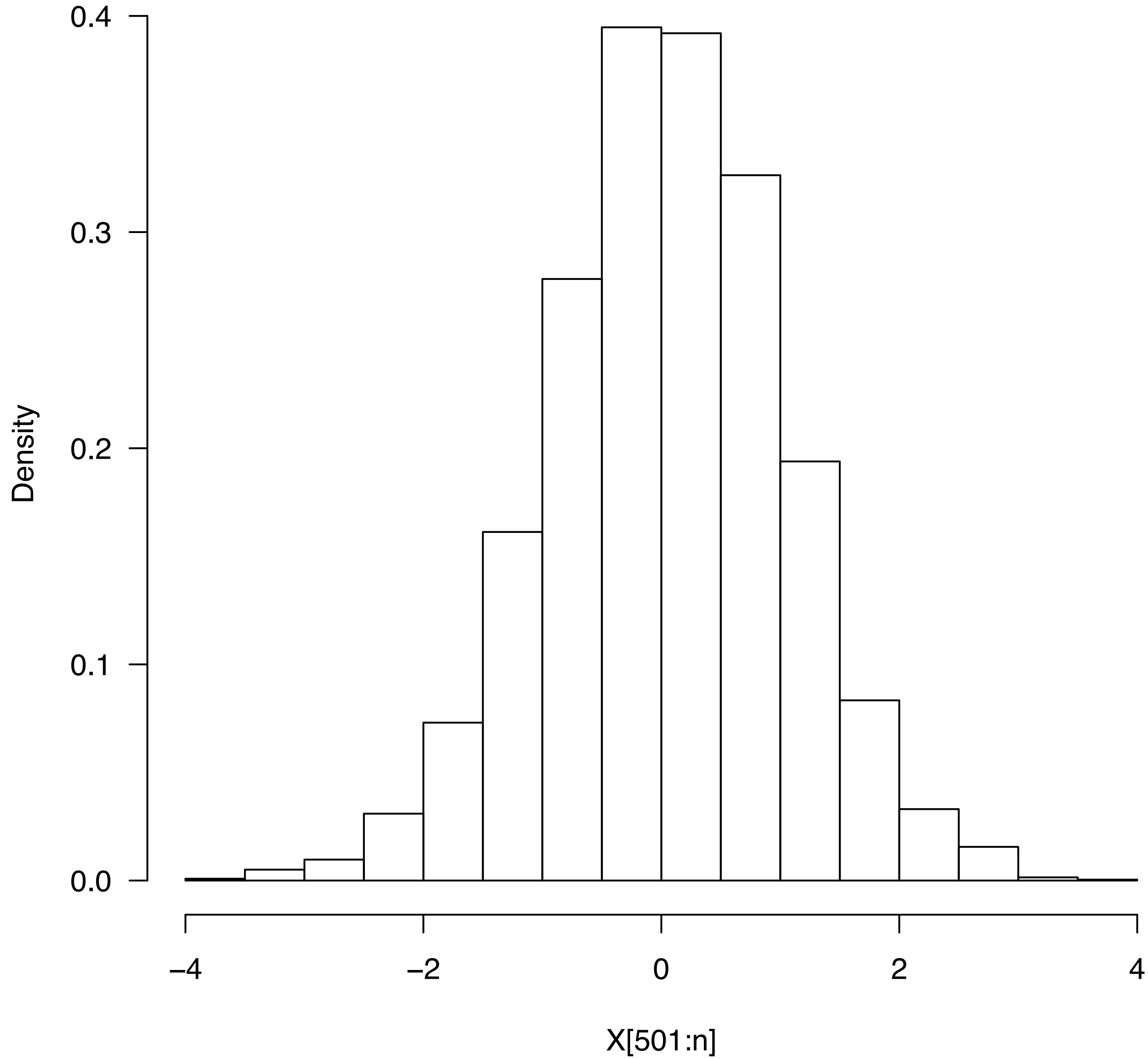
```
> set.seed(9999) # for reproduceability
> n <- 10000 # specifiy length of chain
> X <- 0 # initialize chain

> for (t in 2:n) {
+     # Generate Y from proposal
+     Y <- runif(1, X[t - 1] - 1, X[t - 1] + 1)
+
+     # Compute MH ratio
+     r <- min(1, exp(-0.5 * (Y^2 - X[t - 1]^2)))
+
+     U <- runif(1, 0, 1) # Generate U from Unif(0,1)
+     if (U <= r) {
+         X[t] <- Y # Move to Y if U <= r
+     } else{
+         X[t] <- X[t - 1] # Stay at X[t - 1] if U > r
+     }
+ }
```

Example 2: Standard Normal Distribution



Example 2: Standard Normal Distribution



Example 2: Standard Normal Distribution

R Code for the plots:

```
> # Plot the Markov chain over time
> plot(X, type = "l", las = 1)

> # Plot histogram after 500 burn-in iterations
> hist(X[501:n], prob = TRUE, las = 1, main = "")
```

Example 3: The Poisson Distribution

- The Poisson distribution with mean parameter λ has PDF

uniform : trail

$$\text{target}: \pi(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots. \quad y \geq 0$$

$x \geq 0 \rightarrow y \geq 0$

- Consider the proposal distribution $\sigma = 1$

$$q(x, y): \begin{cases} \text{If } x \geq 1, \quad y = \begin{cases} x+1, & \text{with probability } \frac{1}{2} \\ x-1, & \text{with probability } \frac{1}{2} \end{cases} \\ \text{If } x = 0, \quad \underline{y = x + 1,} \quad \text{with probability } 1. \end{cases}$$

$y \geq 0$

$x-1$ (×)

- We want to find the form of $r(x, y)$ for all x, y.

- **Exercise**: Implement the Metropolis-Hastings algorithm to sample from $\pi(x) \sim \text{Pois}(\lambda = 5)$.

$$r(x^{(t-1)}, y) = \min\left[1, \frac{\pi(y)}{\pi(x^{(t-1)})} \cdot \frac{q(y, x^{(t-1)})}{q(x^{(t-1)}, y)}\right]$$

Example 3: The Poisson Distribution

$\pi(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

There are three cases:

1. If $x \geq 1$ and $y \geq 1$, then $q(x, y) = q(y, x)$, so

MH Ratio
$$\frac{\pi(y)}{\pi(x)} \cdot \frac{\overbrace{q(y,x)}^{=q(x,y)}}{q(x,y)} = \frac{\pi(y)}{\pi(x)} = \frac{\frac{e^{-\lambda}\lambda^{y}}{y!}}{\frac{e^{-\lambda}\lambda^{x}}{x!}} = \frac{\lambda^{y}}{\lambda^{x}} \cdot \frac{x!}{y!}.$$

$r(x^{t-1}, y) = \min\left[1, 5^{y-x}\frac{x!}{y!}\right]$

2. If $x = 1$ and $y \overset{-1}{=} 0$, then $p(X_n=1 \mid X_{n-1}=0)=1$

$$\frac{\pi(y)}{\pi(x)} \cdot \frac{q(y,x)}{q(x,y)} = \frac{\pi(0)}{\pi(1)} \cdot \frac{q(0,1)}{q(1,0)} = \frac{\frac{e^{-\lambda}\lambda^{0}}{0!}}{\frac{e^{-\lambda}\lambda^{1}}{1!}} \cdot \frac{1}{\frac{1}{2}} = \frac{2}{\lambda}. = \frac{2}{5} \to r = \frac{2}{5}$$

$\frac{1}{2} = p(X_n=0 \mid X_{n-1}=1)$

$r(x^{t-1}, y) = \min\left[1, \frac{2}{\lambda}\right]$

3. If $x = 0$ and $y \overset{+1}{=} 1$, then

$$\frac{\pi(y)}{\pi(x)} \cdot \frac{q(y,x)}{q(x,y)} = \frac{\pi(1)}{\pi(0)} \cdot \frac{q(1,0)}{q(0,1)} = \frac{\frac{e^{-\lambda}\lambda^{1}}{1!}}{\frac{e^{-\lambda}\lambda^{0}}{0!}} \cdot \frac{\frac{1}{2}}{1} = \frac{\lambda}{2}. = \frac{5}{2},\ r=1$$

$\frac{1}{2}$

1

$r(x^{t-1}, y) = \min\left[1, \frac{\lambda}{2}\right)$

Outline

generate $r_1 = \min\left[1, 5^{(y-x)} \frac{x!}{y!}\right]$ generate $U \sim \text{unifr}(0,1)$

$r_2 = \frac{2}{5}$

Why Does Metropolis-Hastings Work?

Question: Why does the Metropolis-Hastings algorithm work?

- Finding a transition kernel $K(x, y)$ that satisfies the global balance equations with respect to $\pi(x)$ can be difficult.

- Why does the Metropolis-Hastings algorithm produce a Markov chain with stationary distribution $\pi(x)$?

Detailed Balance

global : $\pi_y = \int \pi_x K(x,y)$

Definition

Let $\{X_0, X_1, X_2, \ldots,\}$ be a Markov chain with state space $\mathcal{X}$ and transition kernel $K(x,y)$. The Markov chain is **time reversible** if there is a distribution $\pi(x)$ on $\mathcal{X}$ such that $K(x,y)$ and $\pi(x)$ satisfy the **detailed balance** (or **time reversibility**) condition

$$\underbrace{\pi(x)K(x,y)}_{\text{flow from } x \to y} = \underbrace{\pi(y)K(y,x)}_{\text{flow } y \to x}, \quad \text{for all } x, y \in \mathcal{X}.$$

- The detailed balance condition is more strict than the global balance condition, since it is a condition on the probability flow for every pair of states, not just the total probability flow.
- The detailed balance condition means (roughly speaking) that, on average, the Markov chain moves from x to y as often as it moves from y to x.

Detailed Balance

Theorem (Detailed Balance Implies Global Balance)

If $K(x, y)$ and $\pi(x)$ satisfy the detailed balance condition, then $\pi(x)$ is a stationary distribution of the Markov chain with $K(x, y)$ as the one-step transition kernel.

Proof.

From the detailed balance condition,

$$\pi(x)K(x, y) = \pi(y)K(y, x), \quad \text{for all } x, y \in \mathcal{X}.$$

Integrating both sides with respect to x, we have, for all $y \in \mathcal{X}$,

$$\begin{aligned} \int \pi(x)K(x, y)\,\mathrm{d}x &= \int \pi(y)K(y, x)\,\mathrm{d}x \\ \pi(x)\int K(x,y)\,dx \quad &= \pi(y) \int \underbrace{K(y, x)\,\mathrm{d}x}_{\to 1} \\ &= \pi(y). \end{aligned}$$

Thus the detailed balance condition implies global balance. $\square$

$\pi(y) = \int \pi(x)\, K(x,y)\, dx$ (global)

Detailed Balance

$\pi(y) = \int \pi(x) P(x,y)\, dx$ (global)

DBC

- Detailed balance implies global balance, so detailed balance is a sufficient (but not necessary) condition for stationarity.
- If the proposal distribution $q(x, y)$ satisfies the detailed balance condition, then the corresponding Markov chain with transition kernel $K(x, y) = q(x, y)$ will have $\pi(x)$ as its stationary distribution. if it's not DBC $\rightarrow K(x,y) = q(x,y)\, r(x,y)$
- However, it is generally difficult to find a proposal distribution that satisfies the detailed balance condition.

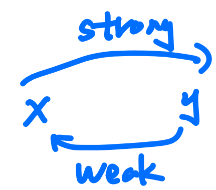
if $q(x) = $ DBC $\rightarrow K(x,y) = q(x,y)$

$q(x) \neq$ DBC $\longrightarrow \pi(x) q(x,y) \neq \pi(y) q(y,x)$

Detailed Balance

- Usually, for some x, y, greater

$$\pi(x)\underset{\text{more often to }y}{\overset{x\to y}{q(x,y)}} > \pi(y)\overset{y\to x}{q(y,x)}.$$



  Roughly speaking, the Markov chain moves, on average, from x to y too often and from y to x too rarely.

- One way to correct for this imbalance is to reduce the number of moves from x to y by introducing a probability $r(x, y) < 1$ of moving from x to y.

- For a suitable choice of $r(x, y)$, the corrected transition kernel

if it's DBC: $K(x,y) = q(x,y)$

$$\underset{p(y|x)}{K(x,y)} = q(x,y)\boxed{r(x,y).}$$

correction vector

  will satisfy the detailed balance conditions for the target distribution $\pi(x)$.

$\pi(x)\, K(x,y) = \pi x\, q(x,y)\, r(x,y)$

Detailed Balance

Theorem (The Metropolis-Hastings Algorithm)

The Metropolis-Hastings algorithm induces a transition kernel with respect to which $\pi(x)$ is a stationary distribution.

Proof (Metropolis-Hastings Algorithm, Part 1).

The transition kernel induced from the Metropolis-Hastings algorithm is

$$K(x, y) = q(x, y)r(x, y) = q(x, y)\min\left[1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right].$$

It suffices to show that the detailed balance condition holds:

$$\pi(x)K(x, y) = \pi(y)K(y, x) \text{ for all } x, y.$$

If $x = y$, the condition is trivially satisfied. We need to show the condition holds for $x \neq$

$$r(x^{(t-1)}, y) = \min\left[1, \frac{\pi(y)}{\pi(x^{(t-1)})} \cdot \frac{q(y, x^{(t-1)})}{q(x^{(t-1)}, y)}\right]$$

Detailed Balance

Proof (Metropolis-Hastings Algorithm, Part 2).

If $x \neq y$, then

$$\begin{aligned}
\pi(x)K(x,y) &= \pi(x)q(x,y)\overset{r(x,y)}{\min\left[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right]} \\
&= \min\left[\pi(x)q(x,y), \pi(y)q(y,x)\right] \\
&= \min\left[\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}, 1\right]\pi(y)q(y,x) \\
&= \pi(y)q(y,x)r(y,x) \\
&= \pi(y)K(y,x).
\end{aligned}$$

So $K(x,y)$ satisfies the detailed balance condition with respect to $\pi(x)$, which implies that the Markov chain with transition kernel $K(x,y)$ will have $\pi(x)$ as a stationary distribution. □

Outline

Choosing the Proposal Distribution

$q(x, y)$

The Metropolis-Hastings algorithm induces a Markov chain that converges to the stationary (limiting) distribution $\pi(x)$ if the Markov chain satisfies certain regularity conditions:

- Irreducibility
  (All states communicate with each other.)
- Aperiodicity
  (Every state has a period of 1.)
- Positive recurrence sure to come back
  (Expected time until the chain returns to any state is finite.)

Positive recurrence is trivially satisfied for irreducible and aperiodic Markov chains with finite state spaces.

Choosing the Proposal Distribution

Sufficient (but not necessary) conditions to check:

- The induced Markov chain is irreducible if

$$q(x, y) > 0 \quad \text{for all } x, y \in \mathrm{Supp}(\pi).$$

  In other words, every state can be reached in a single transition.

- The induced Markov chain is aperiodic if

$$P(x^{(t)} = x^{(t-1)}) > 0. \quad \neq 1 \quad \text{must go out}$$

  In other words, there is positive probability that the chain remains in the current state.

Choosing the Proposal Distribution

The spread (variance) of the proposal distribution affects the behavior of the Markov chain in two main ways:

- Acceptance rate
  (How often the proposal is accepted.)

- Mixing rate
  (How long it takes to move through the state space.)

A Markov chain that moves quickly through the state space is said to have **good mixing behavior**.

Choosing the Proposal Distribution

- If the spread of the proposal is too large:
  - The proposal moves through the state space quickly (good mixing behavior).
  - The proposed state may be far from the current state, and the probability of acceptance will be low.
- If the spread of the proposal is too small:
  - The probability of acceptance will be high.
  - The chain will take a long time to move through the state space, and low density regions will be undersampled.
- Both situations above will exhibit high **autocorrelation**: Correlation between subsequent values in the chain.
- An ideal proposal distribution will balance between good mixing behavior and a high acceptance rate.

Outline

Ergodic Theorem (Law of Large Numbers)

Suppose $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ is a sample from a Markov chain with stationary distribution $\pi(x)$. Let $h(x)$ be a function, and

$$E_\pi[h(X)] = \int h(x)\pi(x)\,\mathrm{d}x.$$

Define

$$\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X^{(i)}).$$

If the Markov chain is irreducible, aperiodic, and positive recurrent, then

$$\bar{h}_n \xrightarrow{\text{a.s.}} E_\pi[h(X)],$$

for any initial state $X^{(0)}$.

Autocorrelation and Efficiency

Assume that $X^{(0)} \sim \pi(x)$. Then

$$\begin{aligned} n\mathrm{Var}(\bar{h}_n) &= \sigma^2 \left[1 + 2\sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \rho_j \right] \\ (n \to \infty) &\approx \sigma^2 \left[1 + 2\sum_{j=1}^{\infty} \rho_j \right], \end{aligned}$$

where $\sigma^2 = \mathrm{Var}[h(X)]$ and

$$\rho_j = \mathrm{Corr}[h(X^{(1)}), h(X^{(j+1)})]$$

is the **j-step** (or **lag j**) **autocorrelation**.

Autocorrelation and Efficiency

- The **integrated autocorrelation time** is defined as

$$\tau_{\text{int}}(h) = 1 + 2\sum_{j=1}^{\infty} \rho_j,$$

so $\operatorname{Var}(\bar{h}_n) \approx \dfrac{\tau_{\text{int}}(h)\sigma^2}{n}$.

$Var(h(x)) = \sigma^2 \cdot \Big[\underbrace{1 + \sum_{1}^{\infty} \rho_j}_{\tau_{int}(h)}\Big]$

- For independent samples, $\operatorname{Var}(\bar{h}_n) = \dfrac{\sigma^2}{n}$.

- The **effective sample size** of $\bar{h}_n$ is then $\dfrac{n}{\tau_{\text{int}}(h)}$.

- Therefore, if the Markov chain has high autocorrelation, the effective sample size can be much smaller than n.

Estimation by Markov Chain Monte Carlo

Central Limit Theorem for Markov Chains

Suppose $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ is a sample from a Markov chain with stationary distribution $\pi(x)$. Let $h(x)$ be a function, and

$$E_\pi[h(X)] = \int h(x)\pi(x)\,\mathrm{d}x.$$

Define

$$\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X^{(i)}).$$

If the Markov chain is irreducible, aperiodic, and positive recurrent, then

$$\frac{\bar{h}_n - E_\pi[h(X)]}{\sqrt{\frac{1}{n}\mathrm{Var}[h(X)]}} \xrightarrow{d} \mathcal{N}(0, \overset{\sigma^2}{\tau_{\mathrm{int}(h)}}),$$

for any initial state $X^{(0)}$.

Example 4: Ising Model

- The **Ising model** is a mathematical model of ferromagnetism. We will focus on the one-dimensional Ising model.
- Consider the spin (either positive or negative) of d magnetic particles in a system. We represent the spins of these particles by $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, where $x_i = \pm 1$ for $i = 1, 2, \ldots, d$.
- Define the **energy** of the particles in state $\boldsymbol{x}$ by

$$h(\boldsymbol{x}) = -\sum_{i=1}^{d-1} x_i x_{i+1}.$$

- At a fixed temperature T, the **Boltzmann distribution** (the probability that the particles are in state $\boldsymbol{x}$) is given by

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{kT} h(\boldsymbol{x})\right) = \exp\left(\frac{1}{kT}\sum_{i=1}^{d-1} x_i x_{i+1}\right),$$

where k is the **Boltzmann constant**.

Example 4: Ising Model

- We want to use the Metropolis-Hastings algorithm to sample from the Boltzmann distribution $\pi(\boldsymbol{x})$ and estimate the average (mean) energy in the system: $E_\pi[h(\boldsymbol{x})]$.
- Denote the current configuration of the particles by
$$\boldsymbol{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_d^{(t)}).$$
- Given the current configuration $\boldsymbol{x}^{(t)}$, we propose a new configuration by randomly selecting a spin j and flipping its orientation from $x_j^{(t)}$ to $-x_j^{(t)}$.
- The proposed state is then
$$\boldsymbol{y} = (x_1^{(t)}, \ldots, -x_j^{(t)}, \ldots, x_d^{(t)}) \text{ for some } j, \text{ with prob. } \frac{1}{d}.$$

- The proposal distribution is symmetric:

$$q(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = q(\boldsymbol{y}, \boldsymbol{x}^{(t)}) = \frac{1}{d}.$$

- The Metropolis-Hastings ratio is then

$$r(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})}\right],$$

where

$$\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})} = \begin{cases} \exp\left(-\frac{2}{kT} x_j^{(t)} x_{j+1}^{(t)}\right) & \text{if } j = 1, \\ \exp\left(-\frac{2}{kT} x_{j-1}^{(t)} x_j^{(t)}\right) & \text{if } j = d, \\ \exp\left(-\frac{2}{kT} x_j^{(t)} (x_{j-1}^{(t)} + x_{j+1}^{(t)})\right) & \text{otherwise.} \end{cases}$$