Markov Chain Monte Carlo and the Metropolis Hastings Algorithm

Bob Zhao

July 11, 2025

Motivation

- Many modern problems involve complex, high-dimensional distributions.
- Classical sampling methods fail to scale.
- We need general-purpose tools for inference.

Why Sampling is Hard

- We often only know $\pi(x) \propto f(x)$.
- Computing the normalizing constant is intractable.
- But we still want samples from π.

Example:

$$\pi(x) \propto f(x) = e^{-x^4+3x^2}$$

$$Z = \int_{-\infty}^{\infty} e^{-x^4+3x^2}\, dx$$

Markov Property

Definition

A stochastic process $(X_t)_{t \geq 0}$ with state space X satisfies the Markov property if

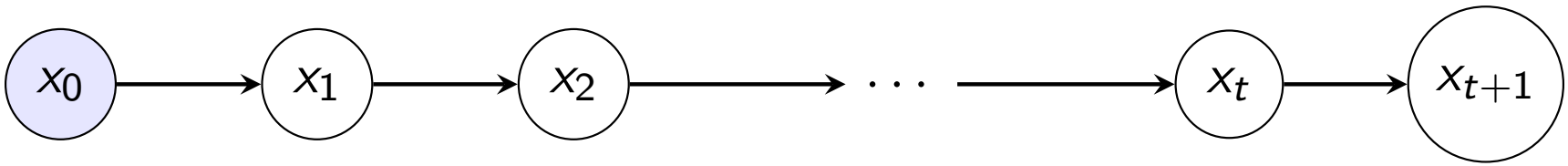
$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t).$$

Markov Property

Definition

A stochastic process $(X_t)_{t \geq 0}$ with state space X satisfies the Markov property if

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t).$$



Markov Chains are **memoryless**

Transition Kernel

Definition

Let X be a countable state space and $P(x \to x')$ denote the transition kernel. Then,

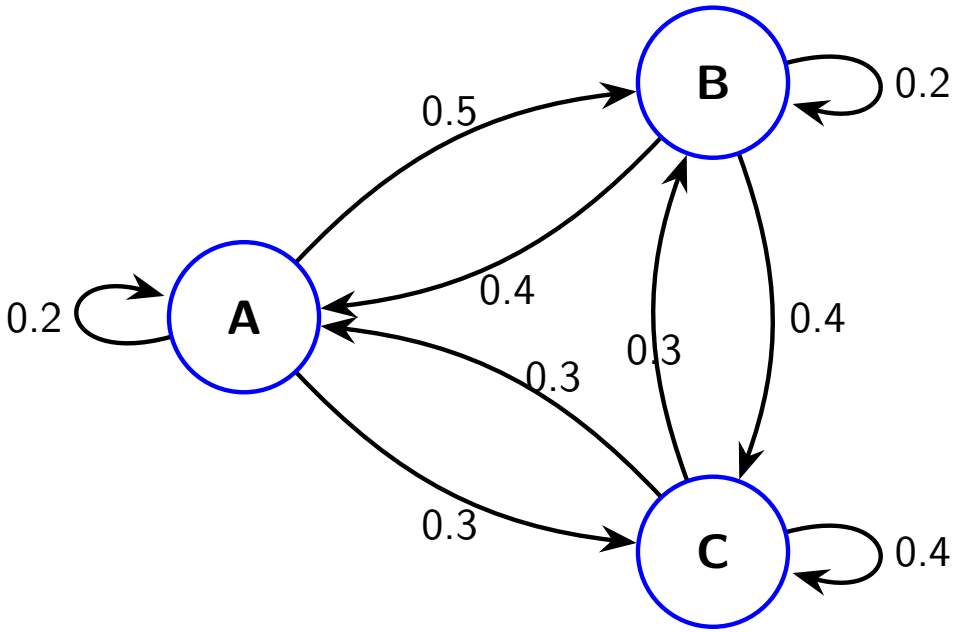
$$\sum_{x'} P(x \to x') = 1, \quad P(x \to x') \geq 0 \quad \forall x, x'.$$

Transition Kernel

Definition

Let X be a countable state space and $P(x \to x')$ denote the transition kernel. Then,

$$\sum_{x'} P(x \to x') = 1, \quad P(x \to x') \geq 0 \quad \forall x, x'.$$



Stationary Distribution

Definition

$$\sum_{x \in X} \pi(x) P(x \to x') = \pi(x') \quad \text{for all } x' \in X.$$

Stationary Distribution

Definition

$$\sum_{x \in X} \pi(x) P(x \to x') = \pi(x') \quad \text{for all } x' \in X.$$

Example:

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$$

$$\pi P = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.2 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix} = \pi$$

Detailed Balance

Definition

$$\pi(x)P(x \to x') = \pi(x')P(x' \to x) \quad \text{for all } x, x' \in X.$$

Detailed Balance

Definition

$$\pi(x)P(x \to x') = \pi(x')P(x' \to x) \quad \text{for all } x, x' \in X.$$

Example of Detailed Balance

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \end{bmatrix}, \quad \pi = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

$$\pi(1)P(1 \to 2) = \frac{1}{4} \cdot 1 = \frac{1}{4}, \quad \pi(2)P(2 \to 1) = \frac{1}{2} \cdot 0.5 = \frac{1}{4}$$

$$\pi(2)P(2 \to 3) = \frac{1}{2} \cdot 0.5 = \frac{1}{4}, \quad \pi(3)P(3 \to 2) = \frac{1}{4} \cdot 1 = \frac{1}{4}$$

Irreducibility

Definition

A Markov chain with state space X is **irreducible** if for any $x, x' \in X$, there exists $t \in \mathbb{N}$ such that

$$P^t(x \to x') > 0.$$

Irreducibility

Definition

A Markov chain with state space X is **irreducible** if for any $x, x' \in X$, there exists $t \in \mathbb{N}$ such that

$$P^t(x \to x') > 0.$$

Example Transition Matrix

$$P = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.6 & 0.4 \\ 0.3 & 0.0 & 0.7 \end{bmatrix}$$

Aperiodicity

Definition

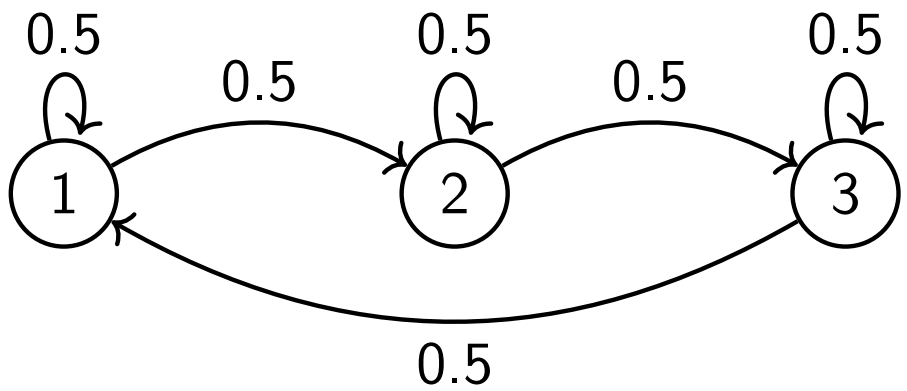
A state $x \in X$ has period d if

$$d = \gcd\{t \in \mathbb{N} P^t(x \to x) > 0\}.$$

Example: Aperiodicity

- Consider a 3-state Markov chain:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$



- Each state can return to itself in multiple steps (2, 3, 4...).
  - So the period is $\gcd(2, 3, 4, \dots) = 1$: chain is aperiodic.

Why These Matter

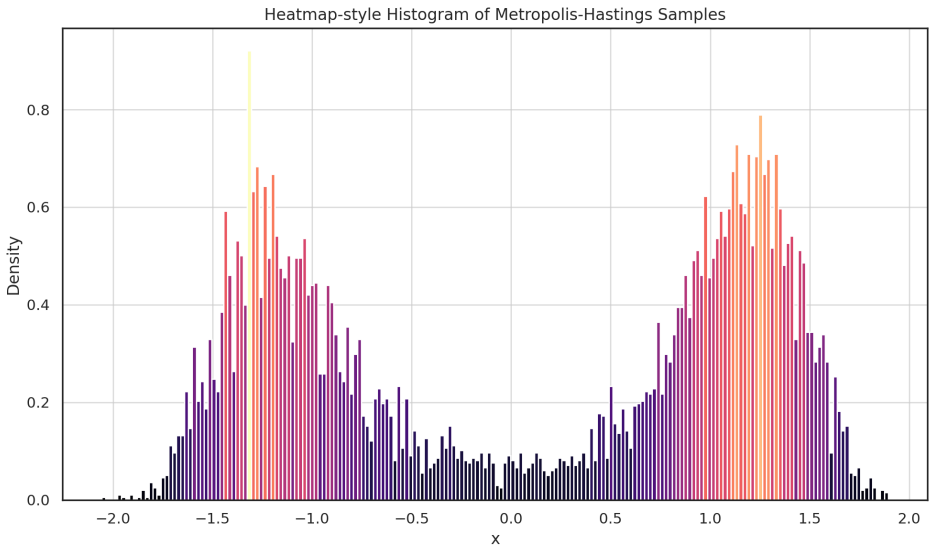
- **Irreducibility**: The chain can explore the entire state space.
- **Aperiodicity**: The chain does not get locked into cyclic patterns.
- Together with a stationary distribution, they ensure convergence of the chain to that distribution regardless of starting point.

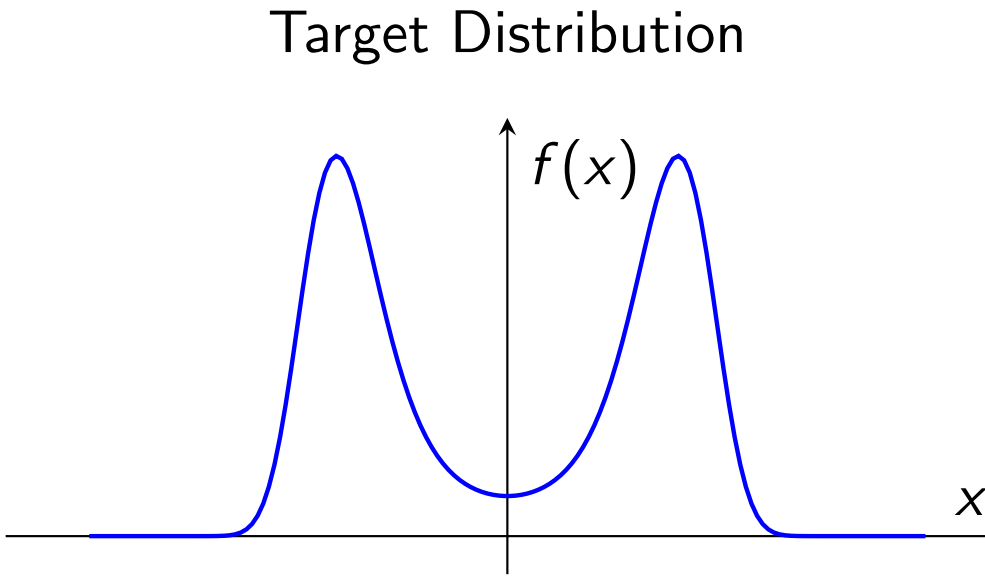
MCMC Overview

- MCMC "reverse engineers" a markov chain.
- Goal: Long-run distribution of the chain equals π.

MCMC Overview

- MCMC "reverse engineers" a markov chain.
- Goal: Long-run distribution of the chain equals π.



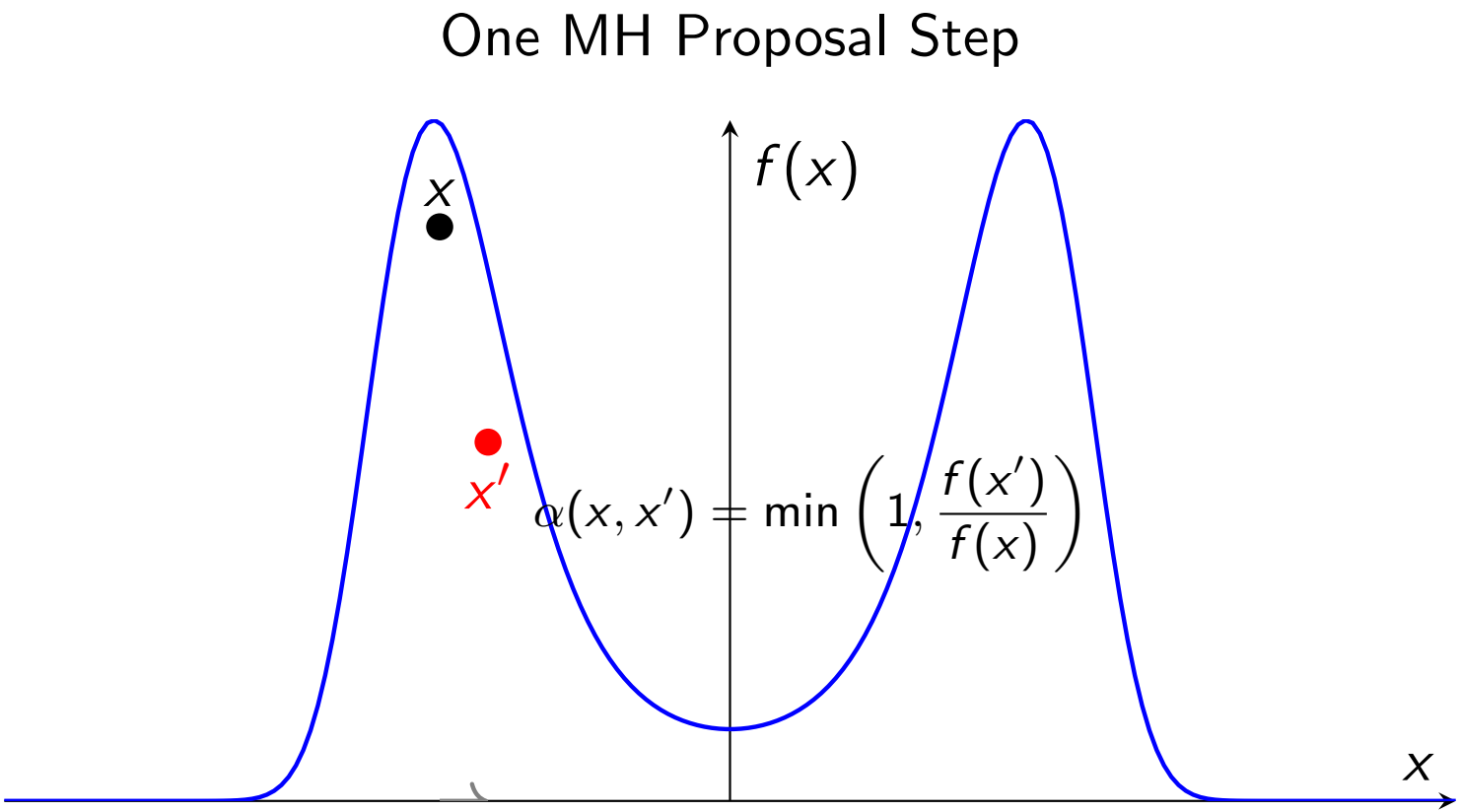


Metropolis-Hastings

- Propose $x' \sim q(x \to x')$.
- Accept with probability:

$$\alpha(x, x') = \min\left(1, \frac{\pi(x')q(x' \to x)}{\pi(x)q(x \to x')}\right)$$

Metropolis-Hastings: Proposal and Acceptance



Total Variation Distance

Definition

The total variation distance between two distributions μ and π over a discrete state space X is:

$$\|\mu - \pi\|_{\mathrm{TV}} := \frac{1}{2} \sum_{x \in X} |\mu(x) - \pi(x)|$$

Total Variation Distance

Definition

The total variation distance between two distributions μ and π over a discrete state space X is:

$$\|\mu - \pi\|_{\text{TV}} := \frac{1}{2} \sum_{x \in X} |\mu(x) - \pi(x)|$$

- Measures how far apart two distributions are.
- Ranges from 0 (identical) to 1 (completely disjoint).

Mixing Time

Definition

The **mixing time** $t_{\text{mix}}(\varepsilon)$ is the smallest time t such that the distribution of the chain is within ε of the stationary distribution π, for all starting states:

$$t_{\text{mix}}(\varepsilon) := \min \left\{ t : \max_{x \in X} \| P^t(x, \cdot) - \pi \|_{\text{TV}} \leq \varepsilon \right\}.$$

Mixing Time Example

- Markov chain with transition matrix:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \quad \text{Stationary distribution: } \pi = \left[\frac{5}{6}, \frac{1}{6}\right]$$

Mixing Time Example

- Markov chain with transition matrix:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \quad \text{Stationary distribution: } \pi = \left[\frac{5}{6}, \frac{1}{6}\right]$$

- Start with $\mu_0 = [0, 1]$ (start from state 2)

t	μ_t	$\|\mu_t - \pi\|_{\text{TV}}$
0	$[0, 1]$	$5/6$
1	$[0.5, 0.5]$	$1/3$
2	$[0.7, 0.3]$	$1/15$
3	$[0.78, 0.22]$	≈ 0.018
∞	$[\frac{5}{6},\ 1/6]$	0

Expected Squared Jump Distance (ESJD)

Definition

The **Expected Squared Jump Distance (ESJD)** is defined as:

$$\text{ESJD} = \mathbb{E}[\epsilon^2 \cdot \alpha(x, x + \epsilon)],$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\alpha(x, x + \epsilon)$ is the Metropolis-Hastings acceptance probability.

Variants

- Adaptive MH

Variants

- Adaptive MH
- MALA (uses gradients)

Variants

- Adaptive MH
- MALA (uses gradients)
- Hamiltonian Monte Carlo (HMC)

Adaptive Metropolis-Hastings (AMH)

- Proposal updates over time:

Adaptive Proposal at Step n

$$q_n(x' \mid x) = \mathcal{N}(x, \sigma^2 \Sigma_n)$$

$$\Sigma_n = \text{Cov}(x_1, \ldots, x_n) + \epsilon I$$

Metropolis-Adjusted Langevin Algorithm (MALA)

- Incorporates gradient of log-density into proposal.
- Moves towards higher-density regions.

Proposal Step

$$x' = x + \frac{\epsilon^2}{2} \nabla \log f(x) + \epsilon Z, \quad Z \sim \mathcal{N}(0, I)$$

Hamiltonian Monte Carlo (HMC)

- Simulates Hamiltonian dynamics with position x and momentum p.
- Avoids random walk behavior by using gradients to propose distant points with high acceptance.

Hamiltonian

$$H(x, p) = -\log f(x) + \frac{1}{2}\|p\|^2$$

Acknowledgments

- Dr. Simon Rubinstein-Salzedo, for organizing and guiding the program.
- Rachana Madhukara, for her invaluable help as a TA.
- My fellow students at Euler Circle, for sharing all of the fun math topics.