The Metropolis-Hastings Algorithm

Angelos Korakitis

National and Kapodistrian University of Athens

14 November 2024

Outline of talk

Ising Model

Markov Chain Fundamentals

Metropolis-Hastings Algorithm

Table of Contents

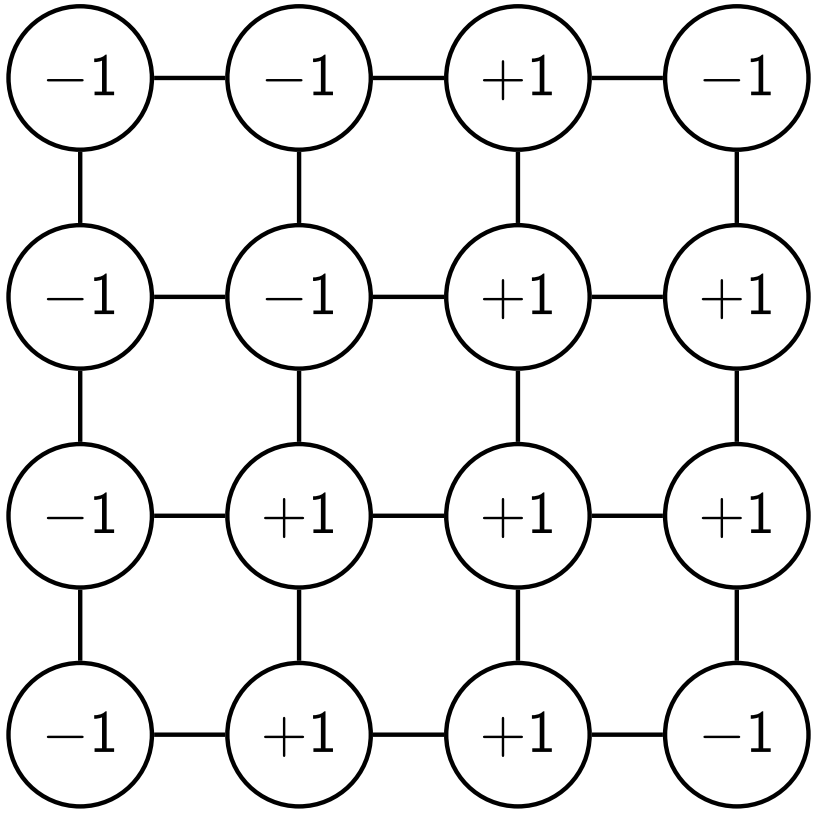
Ising Model

The Ising model:

- Standard model from statistical physics
- Models ferromagnetism
- Shows how microscopic particles create macroscopic magnetic properties

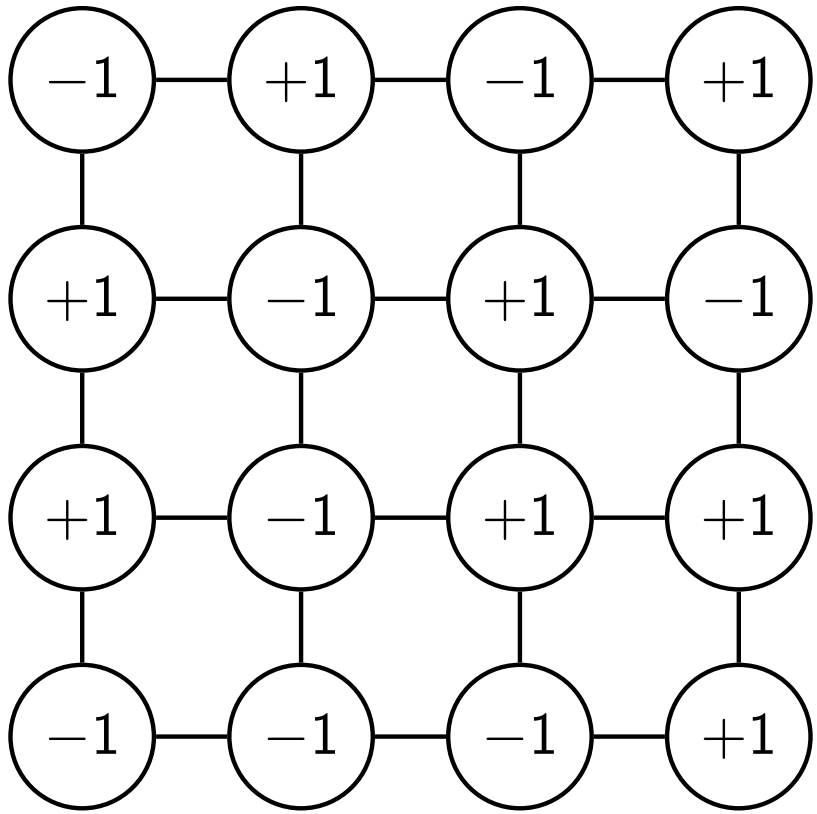
Ising Model

- The lattice Λ is structured as an $L \times L$ grid.
- Each lattice node contains a particle with a spin, either "up" (+1) or "down" (-1): $\sigma : \Lambda \to \{\pm 1\}$
- Interactions between neighboring spins affect the energy of the system: $H(\sigma) = -\sum_{i \sim j} \sigma(i)\sigma(j)$



Hamiltonian

- Interactions between neighboring spins affect the energy of the system: $H(\sigma) = -\sum_{i \sim j} \sigma(i)\sigma(j)$



Observe that energy maximazes when we have complete disorder and minimizes when all 1 or -1

Gibbs Measure

Probability of a configuration:

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}$$

where:

- $\pi_u(\sigma) = e^{-\beta H(\sigma)}$
- $Z(\beta) = \sum_\sigma e^{-\beta H(\sigma)}$

Gibbs Measure

Probability of a configuration:

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}$$

where:

- $\pi_u(\sigma) = e^{-\beta H(\sigma)}$
- $Z(\beta) = \sum_\sigma e^{-\beta H(\sigma)}$
- $\beta = \frac{1}{T}$ (inverse temperature)

Gibbs Measure

Probability of a configuration:

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}$$

where:

- $\pi_u(\sigma) = e^{-\beta H(\sigma)}$
- $Z(\beta) = \sum_\sigma e^{-\beta H(\sigma)}$
- $\beta = \frac{1}{T}$ (inverse temperature)
- Observe that because $\beta > 0$, Gibbs measure assigns bigger probability at states with lower energy. That is, states for which neighboring particles have the same spin.

Gibbs Measure

Probability of a configuration:

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}$$

where:

- $\pi_u(\sigma) = e^{-\beta H(\sigma)}$
- $Z(\beta) = \sum_\sigma e^{-\beta H(\sigma)}$
- $\beta = \frac{1}{T}$ (inverse temperature)
- Observe that because $\beta > 0$, Gibbs measure assigns bigger probability at states with lower energy. That is, states for which neighboring particles have the same spin.
- *As expected from our natural understanding, at low temperature the magnet has his ferromagnetic properties*

To get a better understanding...

Consider the edge cases:

- $T \to 0$ ($\beta \to \infty$):
  - Gibbs measure tends to concetrate on the states with the lowest possible energy
  - Spins align, all $+1$ or all -1
- $T \to \infty$ ($\beta \to 0$):
  - All configurations have the same probability
  - Disorder dominates

To get a better understanding...

Consider the edge cases:

- $T \to 0$ ($\beta \to \infty$):
  - Gibbs measure tends to concetrate on the states with the lowest possible energy
  - Spins align, all $+1$ or all -1
- $T \to \infty$ ($\beta \to 0$):
  - All configurations have the same probability
  - Disorder dominates

Remark

As we heat the system from temperature $T = 0$ to temperature $T = \infty$, the system transitions from a state of complete organization (all spins aligned), to a state of complete disorder, where all spin configurations have the same probability.

The Sampling Challenge

- Number of configurations: $2^{|\Lambda|}$

The Sampling Challenge

- Number of configurations: $2^{|\Lambda|}$
- Computing Z is intractable
  - Even small lattices have enormous state spaces
  - E.g., 100×100 lattice has $2^{10000} > 10^{3000}$ states
  - Atoms in the universe are estimated to be around 10^{80} ...

The Sampling Challenge

Question

Can we sample π?

The Sampling Challenge

Question

Can we sample π?

Question

Can we sample π without computing Z?

Table of Contents

Markov chain

Definition (Markov Chain)

A (discrete-time) Markov chain on Ω is a sequence of random variables $\{X_{i=1}^{\infty}\}$ taking values in Ω satisfying the Markov property:

$$Pr[X_t = x_t | X_0 = x_0, \ldots, X_{t-1} = x_{t-1}] = Pr[X_t = x_t | X_{t-1} = x_{t-1}]$$

In other words, the distribution of the next state is independent of the history given the present. E.g., boarding games.

Markov chain

We can describe a Markov chain using two parameters:

- First, we have the *transition (probability) matrix* $P \in \mathbb{R}^{\Omega \times \Omega}$. The entries $P(x, y)$ specify the transition probabilities

$$Pr[X_t = x | X_{t-1} = y]$$

for all $x, y \in \Omega$.

Markov chain

We can describe a Markov chain using two parameters:

- First, we have the *transition (probability) matrix* $P \in \mathcal{R}^{\Omega \times \Omega}$. The entries $P(x, y)$ specify the transition probabilities

$$Pr[X_t = x | X_{t-1} = y]$$

  for all $x, y \in \Omega$.
- Second, we have an initial distribution π_0

Markov chain

We can describe a Markov chain using two parameters:

- First, we have the *transition (probability) matrix* $P \in \mathcal{R}^{\Omega \times \Omega}$. The entries $P(x, y)$ specify the transition probabilities

$$Pr[X_t = x | X_{t-1} = y]$$

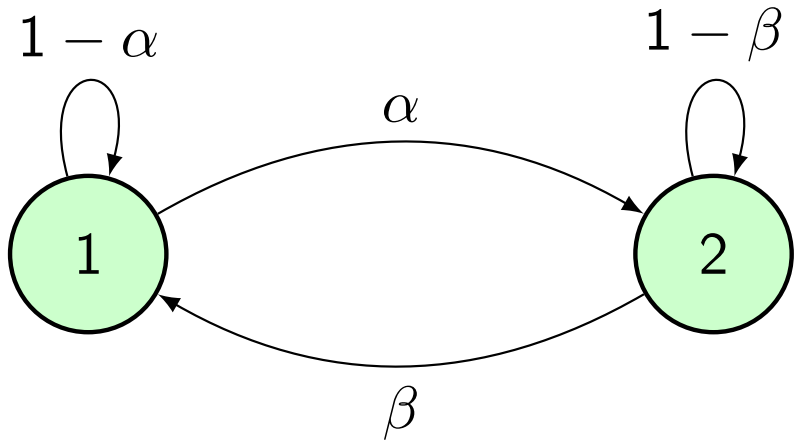
  for all $x, y \in \Omega$.
- Second, we have an initial distribution π_0

The distribution π_t of X_t over Ω is given by

$$\pi_t = \pi_0 P^t$$

Markov Chain

One should conceptually imagine the Markov chain as a random walk on $\Omega = \{1, 2\}$.



$$\pi_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

Stationary Distribution

Definition (Stationary Distribution)

A probability measure π on Ω is stationary w.r.t. a Markov chain P if

$$\pi P = \pi \qquad \pi(x) = \sum_{y \in \Omega} \pi(y) P(y, x), \forall x \in \Omega.$$

Lemma

Every Markov chain P has at least one stationary distribution.

Proof

Proof.

Consider a vector v s.t. $vP = v$. This vector might have negative entries, but we'll use this v to construct our distribution π. Since P has rows summing to 1, $P1 = 1$, in particular P has eigenvalue 1. Since P and P^T have the same eigenvalues (same characteristic polynomials), it follows that there exists a v s.t. $vP = v$. $\square$

Proof (cont'd)

Proof.

Now we define π via $\pi(x) \propto |v(x)|$. We claim that $\pi P = \pi$. Observe that

$$|v(x)| = \left| \sum_{y \in \Omega} v(y) P(y, x) \right| \leq \sum_{y \in \Omega} |v(y)| \, P(y, x), \forall x \in \Omega.$$

We claim that the above inequality is actually an equality. Suppose the contrary, then,

$$\begin{aligned} \sum_{x \in \Omega} |v(x)| &< \sum_{x \in \Omega} \sum_{y \in \Omega} |v(y)| P(y, x) \\ &= \sum_{y \in \Omega} |v(y)| \sum_{x \in \Omega} P(y, x) = \sum_{y \in \Omega} |v(y)|. \end{aligned}$$

which is a contradiction. Hence $|v(x)| = \sum_{y \in \Omega} |v(y)| \, P(y, x), \forall x \in \Omega$, and because $\pi(x) \propto |v(x)|$, $\pi P = \pi$ $\square$

Back to the Ising model

We would like to create a stochastic process that has $\pi(\sigma)$ as its stationary distribution.

Ergodicity

Definition (Ergodicity)

Consider a Markov chain on a finite state space Ω. We say that P is ergodic if P satisfies the following properties:

- *Irreducibility*: P is irreducible if for all $x, y \in \Omega$ there exists a $t \geq 0$ s.t. $P^t(x, y) > 0$. *In other words, the underlying weighted graph of P is strongly connected.*
- *Aperiodicity*: The period of a state $x \in \Omega$ under P is defined as the gcd of $\{t \geq 1 : P^t(x, x) > 0\}$. P is aperiod if all states have period 1.

Ergodicity

Remark

Ergodicity is actually a weak and easy-to-satisfy property.

Ergodicity

- *Irreducibility*: We just need connectivity of Ω under the transitions of P.

Ergodicity

- *Irreducibility*: We just need connectivity of Ω under the transitions of P.
- *Aperiodicity*: For every P we can make an 'equivalent' aperiodic Markov chain by replacing P with

$$\frac{I + P}{2}$$

That is, at every step we toss a coin for staying in the same state or transitioning.

Ergodicity

- *Irreducibility*: We just need connectivity of Ω under the transitions of P.
- *Aperiodicity*: For every P we can make an 'equivalent' aperiodic Markov chain by replacing P with

$$\frac{I+P}{2}$$

That is, at every step we toss a coin for staying in the same state or transitioning.

$$\pi^T\left(\frac{I+P}{2}\right) = \frac{1}{2}\pi^T I + \frac{1}{2}\pi^T P = \frac{1}{2}\pi^T + \frac{1}{2}\pi^T = \pi^T$$

Fundamental Theorem of Markov Chains

Theorem

*Let P be an **ergodic** Markov chain on a state space Ω. Then P has a **unique** stationary distribution π. Furthermore, for every initial distribution π_0, the distribution $\pi_t = \pi_0 P^t$ converges to π.*

Fundamental Theorem of Markov Chains

Remark

Ergodicity is an analogue of the law of large numbers for stochastic processes. In some settings, the law of large numbers holds, even when the sequence of random variables is not i.i.d.

This indicates a connection to Markov chains and sampling.

Question

- *How can we determine the stationary distribution, given a Markov chain?*
- *How to construct a Markov chain making sure that it has π as a stationary distribution?*

Question

How can we determine the stationary distribution, given a Markov chain?

Reversibility/ Detailed Balance

Definition

Reversibility We say a Markov chain P is reversible w.r.t. a distribution π if together they satisfy the detailed balance condition:

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

Time does not play any role.

Reversibility/ Detailed Balance

Definition

Reversibility We say a Markov chain P is reversible w.r.t. a distribution π if together they satisfy the detailed balance condition:

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

Proof.

$$\begin{aligned}
\pi P(x) &= \sum_{y\in\Omega} \pi(y)P(y,x) \\
&= \sum_{y\in\Omega} \pi(x)P(x,y) && (\textit{reversibility}) \\
&= \pi(x)\sum_{y\in\Omega} P(x,y) = \pi(x).
\end{aligned}$$

Question

How to construct a Markov chain making sure that it has π as a stationary distribution?

Question

How to construct a Markov chain making sure that it has π as a stationary distribution?

Wait, that looks like our goal for the Ising model!

Table of Contents

Metropolis-Hastings Algorithm

Require: Initial state X_0, Proposal mechanism $Q(x, y)$,

1: **for** $n = 0, 1, \ldots$ **do**

2: Given X_n, choose y with probability $Q(X_n, y)$

3: Calculate acceptance ratio:

$$A(X_n, y) := \min\left(1, \frac{\pi_u(y)Q(y, X_n)}{\pi_u(x)Q(X_n, y)}\right)$$

4: Set next state:

$$X_{n+1} = \begin{cases} y, & \text{w.p. } A(X_n, y) \\ X_n, & \text{otherwise} \end{cases}$$

5: **end for**

Intuition Behind Metropolis-Hastings

- Use proposal distribution Q that's easy to sample from
- Accept/reject proposals to "correct" for difference between Q and π

Metropolis–Hastings Algorithm

Theorem

The stationary distribution of the Metropolis–Hastings (chain) is π.

Proof.

Let P be a Markov chain with proposal distribution Q and target distribution π. We prove π is stationary by showing detailed balance:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \text{ for all } x, y \in \Omega$$

For $x \neq y$, transition probability is:

$$P(x,y) = Q(x,y)\alpha(x,y)$$

where

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right)$$

□

Proof

Proof.

When $\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)} < 1$:

$$\begin{aligned}\pi(x)P(x,y) &= \pi(x)Q(x,y) \cdot \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)} \\ &= \pi(y)Q(y,x) \\ &= \pi(y)P(y,x)\end{aligned}$$

Because $\alpha(y,x) = 1$.

For $x = y$, detailed balance holds trivially.

Therefore detailed balance holds and π is the stationary distribution. $\square$

Metropolis-Hastings for Ising Model: Proposal Mechanism

- States differ by single spin flip
- Symmetric proposal distribution:

$$Q(x, y) = \begin{cases} \frac{1}{L^2} & \text{if } x, y \text{ differ by one spin} \\ 0 & \text{otherwise} \end{cases}$$

- For single spin flip, energy change depends only on neighbors:

$$\Delta H = 2 s_i \sum_{j \in \mathcal{N}(i)} s_j$$

where $\mathcal{N}(i)$ are nearest neighbors of site i

Metropolis-Hastings for Ising Model: Algorithm

1. Select site i uniformly at random ($\frac{1}{L^2}$ probability)
2. Calculate local energy change using neighbors:

$$\Delta H = 2s_i \sum_{j \in \mathcal{N}(i)} s_j$$

3. Accept flip with probability:

$$\min(1, e^{-\beta \Delta H}) = \begin{cases} 1 & \Delta H \leq 0 \\ e^{-\beta \Delta H} & \Delta H > 0 \end{cases}$$

Note: Only local energy difference needed, not total energy