Sampling Methods

Nipun Batra

September 25, 2023

IIT Gandhinagar

Topics

1. Markov Chains
2. Markov Chain Monte Carlo (MCMC)

Main Goal

- We want to compute posterior predictive distribution (or something similar)

Main Goal

- We want to compute posterior predictive distribution (or something similar)
- We would typically use Monte Carlo methods to do this.

Main Goal

- We want to compute posterior predictive distribution (or something similar)
- We would typically use Monte Carlo methods to do this.
- $I = \int f(x)p(x)dx$ where $p(x)$ is the posterior distribution.

Main Goal

- We want to compute posterior predictive distribution (or something similar)
- We would typically use Monte Carlo methods to do this.
- $I = \int f(x)p(x)dx$ where $p(x)$ is the posterior distribution.
- We can approximate I by $\frac{1}{N}\sum_{i=1}^{N} f(x_i)$, where $x_i \sim p(x)$ are drawn **IID**.

Main Goal

- We want to compute posterior predictive distribution (or something similar)
- We would typically use Monte Carlo methods to do this.
- $I = \int f(x)p(x)dx$ where $p(x)$ is the posterior distribution.
- We can approximate I by $\frac{1}{N}\sum_{i=1}^{N} f(x_i)$, where $x_i \sim p(x)$ are drawn **IID**.
- Goal: sample from $p(x)$, usually using unnormalized density $\tilde{p}(x)$

Limitations of basic sampling methods

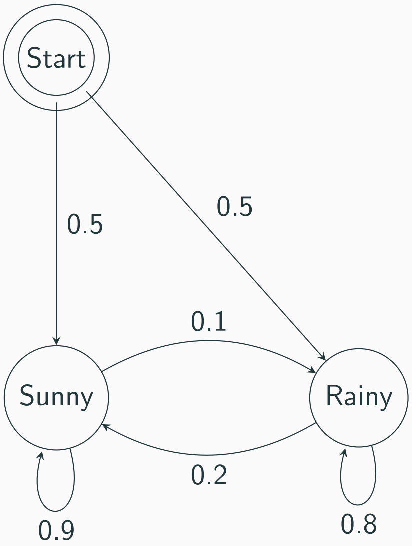
- *Transformation based methods*: Usually limited to drawing from standard distributions.
- *Rejection and Importance sampling*: Require selection of good proposal distirbutions.

In high dimensions, usually most of the density $p(x)$ is concentrated within a tiny subspace of x. Moreover, those subspaces are difficult to be known a priori.

A solution to these are Markov Chain Monte Carlo methods.

Markov Chains

Properties of Markov Chain: Stationarity



Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 1: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 2: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 1: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 2: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

What is the probability of it being sunny on day 0?

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 1: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 2: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

What is the probability of it being sunny on day 0?

0.5

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 3: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 4: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny on day 0?

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 3: Prior Probability (PI)

	$X_0 =$ Sunny	$X_0 =$ Rainy
PI	0.5	0.5

Table 4: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny on day 0?
- 0.5

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 5: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 6: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day 1?

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 5: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 6: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day 1?
- We can have two cases:
  - $X_0 = \text{Sunny}$: $P(X_1 = \text{Sunny}) = 0.9$
  - $X_0 = \text{Rainy}$: $P(X_1 = \text{Sunny}) = 0.2$
  - $P(X_1 = \text{Sunny}) = 0.5 \times 0.9 + 0.5 \times 0.2 = 0.55$
  - $P(X_1 = \text{Rainy}) = 0.5 \times 0.1 + 0.5 \times 0.8 = 0.45$

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 7: Prior Probability (PI)

	$X_0 =$ Sunny	$X_0 =$ Rainy
PI	0.5	0.5

Table 8: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day 2?

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 7: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 8: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day 2?
- We can have two cases:
  - $P(X_2 = \text{Sunny}) = 0.55 \times 0.9 + 0.45 \times 0.2 = 0.585$
  - $P(X_2 = \text{Rainy}) = 0.55 \times 0.1 + 0.45 \times 0.8 = 0.415$

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 9: Prior Probability (PI)

	$X_0 =$ Sunny	$X_0 =$ Rainy
PI	0.5	0.5

Table 10: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day T?

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 9: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 10: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day T?
- We can use matrix power to compute this.

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 9: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 10: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day T?
- We can use matrix power to compute this.
- Distribution of X_T is given by $\pi = \pi POWER(A, T)$.

Properties of Markov Chain: Stationarity

Let us consider the rainy-sunny example.

Table 9: Prior Probability (PI)

	$X_0 = \text{Sunny}$	$X_0 = \text{Rainy}$
PI	0.5	0.5

Table 10: Transition Matrix (A)

		X_{t+1}	
		Sunny	Rainy
X_t	Sunny	0.9	0.1
	Rainy	0.2	0.8

- What is the probability of it being sunny/rainy on day T?
- We can use matrix power to compute this.
- Distribution of X_T is given by $\pi = \pi POWER(A, T)$.
- At $T = 99$ and $T = 100$, $\pi = (0.67, 0.33)$.

Properties of Markov Chain: Stationarity

Notebook: markov-chain.ipynb

Questions:

- Does the distribution of X_T depend on initial distribution π?

Properties of Markov Chain: Stationarity

We can define statationary distribution as follows:

- A distribution π is said to be stationary for a Markov chain with transition matrix A if $\pi = \pi A$.
- For previous example,
    - $\pi = (\pi_1, \pi_2)$
    - $\pi_1 = 0.9\pi_1 + 0.2\pi_2$
    - $\pi_2 = 0.1\pi_1 + 0.8\pi_2$
    - $\pi_1 + \pi_2 = 1$
    - Solving, $\pi = (\frac{2}{3}, \frac{1}{3})$

Properties of Markov Chain: Stationarity

Can we have a Markov chain with multiple stationary distributions?

Properties of Markov Chain: Stationarity

Can we have a Markov chain with multiple stationary distributions?

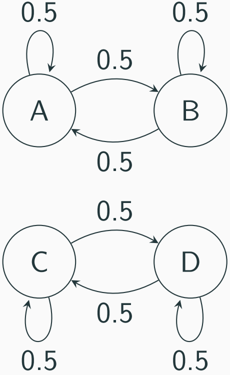


Table 11: Transition Matrix (A)

		X_{t+1}			
		A	B	C	D
X_t	A	0.5	0.5	0	0
	B	0.5	0.5	0	0
	C	0	0	0.5	0.5
	D	0	0	0.5	0.5

Properties of Markov Chain: Stationarity

Can we have a Markov chain with multiple stationary distributions?

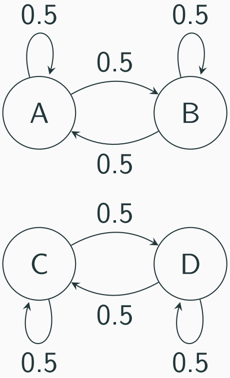


Table 11: Transition Matrix (A)

		X_{t+1}			
		A	B	C	D
X_t	A	0.5	0.5	0	0
	B	0.5	0.5	0	0
	C	0	0	0.5	0.5
	D	0	0	0.5	0.5

- If we start at A or B, the stationary distribution is $(0.5, 0.5, 0, 0)$.
- If we start at C or D, the stationary distribution is $(0, 0, 0.5, 0.5)$.

Properties of Markov Chain: Time Homogeneity

- A Markov chain is said to be **homogeneous** if the transition probabilities are independent of the time t.

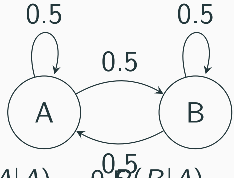
Properties of Markov Chain: Time Homogeneity

- A Markov chain is said to be **homogeneous** if the transition probabilities are independent of the time t.
- We have the same transition matrix A for all t.

Properties of Markov Chain: Irreducibility

- A Markov chain is said to be **irreducible** if every state is accessible from every other state.
- In other words, there is a non-zero probability of reaching any state from any other state.

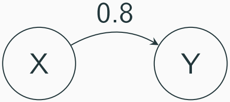
Irreducible Markov Chain (2 States)



$P(A|A) = 0.5$ $P(B|A) = 0.5$

$P(A|B) = 0.5$ $P(B|B) = 0.5$

Non-Irreducible Markov Chain (2 States)



$P(X|X) = 0.2$ $P(Y|X) = 0$

$P(X|Y) = 0.8$ $P(Y|Y) = 1$

Markov Chain Monte Carlo (MCMC)

MCMC main idea

- We then identify a way to construct a 'nice' Markov chain such that its stationary probability distribution is our target distribution $p(x)$.
- We then run the Markov chain for a long time and use the samples to estimate I.
- But, we thus far said: $x_i \sim p(x)$ are drawn **IID**.
- But, if we use a Markov chain to generate samples, then the samples are not i.i.d.
- But, we can still use the samples to estimate I using the **ergodic theorem**.

Ergodic Theorem for Markov Chains

Inspired from: MathematicalMonk's playlisty on MCMC.

- From Monte Carlo sampling, we know we can estimate $I = \int f(x)p(x)dx$ by $\frac{1}{N}\sum_{i=1}^{N} f(x_i)$, where $x_i \sim p(x)$.
- But, the samples are drawn i.i.d. from $p(x)$.
- But, if we use a Markov chain to generate samples, then the samples are not i.i.d.
- But, we can still use the samples to estimate I using the **ergodic theorem**.

Ergodic Theorem for Markov Chains

- Let $X_1, X_2, \ldots$ be a Markov chain with stationary distribution $p(x)$.
- Let f be a function such that $\mathbb{E}[|f(X)|] < \infty$.
- Then, $\frac{1}{N}\sum_{i=1}^{N} f(X_i) \to \mathbb{E}[f(X)]$ as $N \to \infty$.
- The proof is similar to the proof of the law of large numbers.
- The idea is that the estimates contain information about the shape of the target distribution p.

Markov Chain Properties: Stationarity

- A Markov chain is a sequence of random variables $X_1, X_2, \dots$ with the property that the distribution of X_{n+1} given $X_1, \dots, X_n$ depends only on X_n.
- A Markov chain is said to be **stationary** if the distribution of X_{n+1} given $X_1, \dots, X_n$ is the same as the distribution of X_{n+1} given X_n.
- A Markov chain is said to be **homogeneous** if the transition probabilities are independent of n.

Markov Chain

- **Markov Chain**: A joint distribution $p(X)$ over a sequence of random variables $X = \{X_1, X_2, \ldots, X_n\}$ is said to have the Markov property if

$$p(X_i | X_1, \ldots, X_{i-1}) = p(X_i | X_{i-1})$$

  The sequence is then called a Markov chain.
- The idea is that the estimates contain information about the shape of the target distribution p.

Metropolis Hastings

- The basic idea is propose to move to a new state x_{i+1} from the current state x_i with probability $q(x_{i+1}|x_i)$, where q is called the proposal distribution and our target density of interest is $p(=\frac{1}{Z}\tilde{p})$.
- The new state is accepted with probability $\alpha(x_i, x_{i+1})$.
    - If $p(x_{i+1}|x_i) = p(x_i|x_{i+1})$, then $\alpha(x_i, x_{i+1}) = \min(1, \frac{p(x_{i+1})}{p(x_i)})$.
    - If $p(x_{i+1}|x_i) \neq p(x_i|x_{i+1})$, then $\alpha(x_i, x_{i+1}) = \min(1, \frac{p(x_{i+1})q(x_i|x_{i+1})}{p(x_i)q(x_{i+1}|x_i)}) = \min(1, \frac{\tilde{p}(x_{i+1})q(x_i|x_{i+1})}{\tilde{p}(x_i)q(x_{i+1}|x_i)})$
- Evaluating α, we only need to know the target distribution up to a constant of proportionality or without normalization constant.

Algorithm: Metropolis Hastings

1. Initialize x_0.
2. for $i = 1, \ldots, N$ do:
3. Sample $x^* \sim q(x^*|x_{i-1})$.
4. Compute $\alpha = \min(1, \frac{\tilde{p}(x^*)q(x_{i-1}|x^*)}{\tilde{p}(x_{i-1})q(x^*|x_{i-1})})$
5. Sample $u \sim \mathcal{U}(0,1)$
6. if $u \leq \alpha$:

   $x_i = x^*$

   else:

   $x_i = x_{i-1}$

Pop Quiz

How do we choose the initial state x_0?

Pop Quiz

How do we choose the initial state x_0?

1. Start the Markov Chain at an initial x_0.
2. Using the proposal $q(x|x_i)$, run the chain long enough, say N_1 steps.
3. Discard the first $N_1 - 1$ samples (called 'burn-in' samples).
4. Treat x_{N_1} as first sample from $p(x)$.

MCMC demo

https://chi-feng.github.io/mcmc-demo/app.html