Markov Chains and MCMC

CompSci 590.04
Instructor: AshwinMachanavajjhala
Announcement

• First assignment has been posted
  – Please work on it in groups of 2 or 3
  – Involves accessing Twitter for information
  – Only allowed a restricted number of API calls to Twitter a day
  – So do not delay the assignment till the last minute.

• Due date: Friday Sep 11, 11:59 pm
Recap: Monte Carlo Method

• If U is a universe of items, and G is a subset satisfying some property, we want to estimate |G|
  
  – Either intractable or inefficient to count exactly

For i = 1 to N
• Choose u ∈ U, uniformly at random
• Check whether u ∈ G?
• Let $X_i = 1$ if $u ∈ G$, $X_i = 0$ otherwise

Return $\hat{C} = |U| \cdot \frac{\sum_i X_i}{N}$

Variance: $|U| \frac{\mu(1 - \mu)}{\sqrt{N}}$, where $\mu = \frac{|G|}{|U|}$
Recap: Monte Carlo Method

When is this method an FPRAS?

• $|U|$ is known and easy to uniformly sample from $U$.
• Easy to check whether sample is in $G$
• $|U|/|G|$ is small ... (polynomial in the size of the input)

**Theorem:**

$$\forall 0 < \varepsilon < 1.5, 0 < \delta < 1, \text{if } N > \frac{|U|}{|G|} \cdot \frac{3}{\varepsilon^2} \cdot \ln \frac{2}{\delta}$$

then, $P\left[(1 - \varepsilon)|G| \leq \hat{C} \leq (1 + \varepsilon)|G|\right] \geq 1 - \delta$
Recap: Importance Sampling

• In certain case \( |G| \ll |U| \), hence the number of samples is not small.

• Suppose \( q(x) \) is the density of interest, sample from a different approximate density \( p(x) \)

\[
\int f(x)q(x)\,dx = \int f(x)\left(\frac{q(x)}{p(x)}\right)p(x)\,dx
\]

\[
= E_{p(x)} \left[ f(x) \frac{q(x)}{p(x)} \right]
\]

Hence, \( \int f(x)q(x)\,dx \approx \frac{1}{N} \sum_{i=0}^{N} f(X_i) \frac{q(X_i)}{p(X_i)} \),

where \( X_i \) are sampled from \( p(x) \)
Today’s Class

• Markov Chains

• Markov Chain Monte Carlo sampling
  – Standard technique for probabilistic inference in machine learning, when the probability distribution is hard to compute exactly
Markov Chains

• Consider a time varying random process which takes the value $X_t$ at time $t$
  \- Values of $X_t$ are drawn from a finite (more generally countable) set of states $\Omega$.

• $\{X_0 \ldots X_t \ldots X_n\}$ is a Markov Chain if the value of $X_t$ only depends on $X_{t-1}$
Transition Probabilities

- \( \Pr[X_{t+1} = s_j \mid X_t = s_i] \), denoted by \( P(i, j) \), is called the transition probability
  - Can be represented as a \(|\Omega| \times |\Omega|\) matrix \( P \).
  - \( P(i, j) \) is the probability that the chain moves from state \( i \) to state \( j \)

Let \( \pi_i(t) = \Pr[X_t = s_i] \) denote the probability of reaching state \( i \) at time \( t \)

\[
\pi_j(t) = \Pr[X_t = s_j] = \sum_i \Pr[X_t = s_j \mid X_{t-1} = s_i] \Pr[X_{t-1} = s_i] \\
= \sum_i P(i, j) \cdot \Pr[X_{t-1} = s_i] = \sum_i P(i, j) \pi_i(t - 1)
\]
Transition Probabilities

• Pr[\(X_{t+1} = s_j \mid X_t = s_i\)], denoted by P(i,j), is called the transition probability
  – Can be represented as a \(|\Omega| \times |\Omega|\) matrix P.
  – \(P(i,j)\) is the probability that the chain moves from state i to state j

• If \(\pi(t)\) denotes the \(1 \times |\Omega|\) vector of probabilities of reaching all the states at time t,

\[
\pi(t) = \pi(t - 1)P
\]
Example

• Suppose $\Omega = \{\text{Rainy}, \text{Sunny}, \text{Cloudy}\}$
• Tomorrow’s weather only depends on today’s weather.
  – Markov process

\[
P = \begin{bmatrix}
0.5 & 0.25 & 0.25 \\
0.5 & 0 & 0.5 \\
0.25 & 0.25 & 0.5 \\
\end{bmatrix}
\]

$\Pr[X_{t+1} = \text{Sunny} \mid X_t = \text{Rainy}] = 0.25$

$\Pr[X_{t+1} = \text{Sunny} \mid X_t = \text{Sunny}] = 0$

No 2 consecutive days of sun (Seattle?)
Example

• Suppose $\Omega = \{\text{Rainy, Sunny, Cloudy}\}$
• Tomorrow’s weather only depends on today’s weather.
  – Markov process

\[
P = \begin{bmatrix}
0.5 & 0.25 & 0.25 \\
0.5 & 0 & 0.5 \\
0.25 & 0.25 & 0.5
\end{bmatrix}
\]

• Suppose today is Sunny. \( \pi(0) = [0 \ 1 \ 0] \)
• What is the weather 2 days from now?

\[
\pi(2) = \pi(0)P^2 = [0.375 \ 0.25 \ 0.375]
\]
Example

- Suppose $\Omega = \{\text{Rainy}, \text{Sunny}, \text{Cloudy}\}$
- Tomorrow’s weather only depends on today’s weather.
  - Markov process

$$P = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

- Suppose today is Sunny. $\pi(0) = [0 \ 1 \ 0]$
- What is the weather 7 days from now?

$$\pi(7) = \pi(0)P^7 = [0.4 \ 0.2 \ 0.4]$$
Example

• Suppose $\Omega = \{\text{Rainy, Sunny, Cloudy}\}$
• Tomorrow’s weather only depends on today’s weather.
  – Markov process

\[
P = \begin{bmatrix}
0.5 & 0.25 & 0.25 \\
0.5 & 0 & 0.5 \\
0.25 & 0.25 & 0.5 
\end{bmatrix}
\]

• Suppose today is Rainy. $\pi(0) = [1 \ 0 \ 0]$
• What is the weather 2 days from now?
  $\pi(2) = \pi(0)P^2 = [0.4375 \ 0.1875 \ 0.375]$
• Weather 7 days from now?
  $\pi(7) = \pi(0)P^7 = [0.4 \ 0.2 \ 0.4]$
Example

\[ P = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \]

\[ \pi(0) = [0 \ 1 \ 0] \]

\[ \pi(0) = [1 \ 0 \ 0] \]

\[ \pi(7) = \pi(0)P^7 = [0.4 \ 0.2 \ 0.4] \]

\[ \pi(7) = \pi(0)P^7 = [0.4 \ 0.2 \ 0.4] \]

- After sufficient amount of time the expected weather distribution is independent of the starting value.
- Moreover, \( \pi(7) = \pi(8) = \pi(9) = \ldots = [0.4 \ 0.2 \ 0.4] \)
- This is called the stationary distribution.
Stationary Distribution

• \( \pi \) is called a \textit{stationary distribution} of the Markov Chain if

\[
\pi = \pi P
\]

• That is, once the stationary distribution is reached, every subsequent \( X_i \) is a sample from the distribution \( \pi \)

How to use Markov Chains:

• Suppose you want to sample from a set \(|\Omega|\), according to distribution \( \pi \)
• Construct a Markov Chain (\( P \)) such that \( \pi \) is the stationary distribution
• \textit{Once stationary distribution is achieved, we get samples from the correct distribution.}
Conditions for a Stationary Distribution

A Markov chain is **ergodic** if it is:

- **Irreducible**: A state \( j \) can be reached from any state \( i \) in some finite number of steps.

\[
P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.25 & 0.75 \end{bmatrix}
\]
A Markov chain is **ergodic** if it is:

- **Irreducible**: A state $j$ can be reached from any state $i$ in some finite number of steps.

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.25 & 0.75 \end{bmatrix}$$

- **Aperiodic**: A chain is not forced into cycles of fixed length between certain states.

$$P = \begin{bmatrix} 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{bmatrix}$$
Conditions for a Stationary Distribution

A Markov chain is **ergodic** if it is:

- **Irreducible**: A state \( j \) can be reached from any state \( i \) in some finite number of steps.

- **Aperiodic**: A chain is not forced into cycles of fixed length between certain states

**Theorem**: For every ergodic Markov chain, there is a unique vector \( \pi \) such that for all initial probability vectors \( \pi(0) \),

\[
\lim_{t \to \infty} \pi(t) = \lim_{t \to \infty} \pi(0) P^t = \pi
\]
Sufficient Condition: Detailed Balance

• In a stationary walk, for any pair of states $j$, $k$, the Markov Chain is as likely to move from $j$ to $k$ as from $k$ to $j$.

$$\pi_j P(j, k) = \pi_k P(k, j)$$

• Also called reversibility condition.
Example: Random Walks

- Consider a graph $G = (V,E)$, with weights on edges $(w(e))$

Random Walk:
- Start at some node $u$ in the graph $G(V,E)$
- Move from node $u$ to node $v$ with probability proportional to $w(u,v)$.

Random walk is a Markov chain
- State space $= V$
- $P(u,v) = \frac{w(u,v)}{\sum w(u,v')}$ if $(u,v) \in E$
  
  $= 0$ if $(u,v)$ is not in $E$
Example: Random Walk

Random walk is ergodic if:

- **Irreducible**: A state \( j \) can be reached from any state \( i \) in some finite number of steps.

  If \( G \) is connected.

- **Aperiodic**: A chain is not forced into cycles of fixed length between certain states

  If \( G \) is not bipartite
Example: Random Walk

Uniform random walk:
• Suppose all weights on the graph are 1
• \( P(u,v) = 1/\deg(u) \) (or 0)

Theorem: If \( G \) is connected and not bipartite, then the stationary distribution of the random walk is

\[
\pi_u = \frac{\deg(u)}{2|E|}
\]
Example: Random Walk

Symmetric random walk:
• Suppose $P(u,v) = P(v,u)$

Theorem: If $G$ is connected and not bipartite, then the stationary distribution of the random walk is

$$\pi_u = \frac{1}{|V|}$$
Stationary Distribution

• $\pi$ is called a *stationary distribution* of the Markov Chain if

\[
\pi = \pi P
\]

• That is, once the stationary distribution is reached, every subsequent $X_i$ is a sample from the distribution $\pi$

How to use Markov Chains:

• Suppose you want to sample from a set $|\Omega|$, according to distribution $\pi$
• Construct a Markov Chain ($P$) such that $\pi$ is the stationary distribution
• *Once stationary distribution is achieved*, we get samples from the correct distribution.
Metropolis-Hastings Algorithm (MCMC)

• Suppose we want to sample from a complex distribution \( f(x) = \frac{p(x)}{K} \), where \( K \) is unknown or hard to compute.

• Example: Bayesian Inference.
Metropolis-Hastings Algorithm

• Start with any initial value \( x_0 \), such that \( p(x_0) > 0 \)

• Using current value \( x_{t-1} \), sample a new point according some proposal distribution \( q(x_t | x_{t-1}) \)

• Compute \[ \alpha(x_t|x_{t-1}) = \min \left( 1, \frac{p(x_t)}{p(x_{t-1})} \frac{q(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \]

• With probability \( \alpha \) accept the move to \( x_t \), otherwise reject \( x_t \)
Why does Metropolis-Hastings work?

• Metropolis-Hastings describes a Markov chain with transition probabilities:

\[ P(x, y) = q(y | x) \min \left( 1, \frac{p(y) q(x | y)}{p(x) q(y | x)} \right) \]

• We want to show that \( f(x) = \frac{p(x)}{K} \) is the stationary distribution

• Recall sufficient condition for stationary distribution:

\[ \pi_j P(j, k) = \pi_k P(k, j) \]
Why does Metropolis-Hastings work?

• Metropolis-Hastings describes a Markov chain with transition probabilities:

\[ P(x, y) = q(y | x) \min \left( 1, \frac{p(y) q(x | y)}{p(x) q(y | x)} \right) \]

• Sufficient to show: \( p(x)P(x, y) = p(y)P(y, x) \)
Proof: Case 1

\[ P(x, y) = q(y | x) \min \left( 1, \frac{p(y) q(x | y)}{p(x) q(y | x)} \right) \]

- Suppose \( p(y) q(x | y) = p(x) q(y | x) \)

- Then, \( P(x, y) = q(y | x) \)

- Therefore
  \[ P(x, y)p(x) = q(y | x) p(x) = p(y) q(x | y) = P(y, x) p(y) \]
Proof: Case 2

\[ P(x, y) = q(y | x) \min\left(1, \frac{p(y) q(x | y)}{p(x) q(y | x)}\right) \]

Suppose, \( p(y)q(x | y) > p(x) q(y | x) \)

Then, \( \alpha(y | x) = 1, \quad \alpha(x | y) = \frac{p(x)q(y | x)}{p(y)q(x | y)} \)

\[
P(y, x)p(y) = q(x | y)\alpha(x | y)p(y) \\
= q(x | y)\frac{p(x)q(y | x)}{p(y)q(x | y)}p(y) = p(x)q(y | x) \\
= p(x)q(y | x)\alpha(y | x) = p(x)P(x, y)
\]

• Proof of Case 3 is identical.
When is stationary distribution reached?

• Next class ...