Markov Chain Monte Carlo Method

AI Friends Seminar

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Introduction

1. Monte Carlo Simulation
   1. We want to estimate $E[h(X)]$ for a function $h(x)$ where X is a random variable with distribution function $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$ (or $f(x)$).
   2. Use the Strong Law of Large Numbers to estimate

$$E[h(X)] \sim \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

2. What is a Markov Chain Monte Carlo (MCMC) sampling?
   1. It generates a sample from some complex probability distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$ (or $f(x)$).
   2. To this end, the MCMC sampling uses a discrete time Markov chain with the stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$.

Discrete Time Markov Chain

Definition 1

The sequence of R.V.s $X_0, X_1, X_2, \cdots$ with a countable state space S is said to be a discrete time Markov chain (DTMC) if it satisfies the Markov Property:

$$\begin{aligned} & P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} \\ = \ & P\{X_{n+1} = j | X_n = i\} \end{aligned}$$

for any $i_k \in S, k = 0, 1, \cdots, n-1$ and $i, j \in S$.

Time homogeneous DTMC : $P\{X_{n+1} = j | X_n = i\}$ is independent of n.
From now on we consider a time homogeneous DTMC.

State Transition Probability Matrix

One step transition probability matrix $\mathbf{P}$ of a DTMC

- $\mathbf{P} = (p_{ij})$ where $p_{ij} = P\{X_{n+1} = j | X_n = i\}$.
- The matrix $\mathbf{P}$ is nonnegative and stochastic, i.e., $p_{ij} \geq 0$ and $\sum_{j \in S} p_{ij} = 1$.

n step transition probability matrix $\mathbf{P}$ of a DTMC $\mathbf{P}^{(n)}$

$$p_{ij}^{(n)} = P\{X_n = j | X_0 = i\}$$

Theorem 1

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)}.$$

$$\begin{aligned}
p_{ij}^{(n+m)} &= P\{X_{n+m} = j | X_0 = i\} \\
&= \sum_{k \in S} P\{X_{n+m} = j, X_n = k | X_0 = i\} \\
&= \sum_{k \in S} P\{X_{n+m} = j | X_n = k, X_0 = i\} P\{X_n = k | X_0 = i\} \\
&= \sum_{k \in S} P\{X_{n+m} = j | X_n = k\} P\{X_n = k | X_0 = i\} \\
&= \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)}
\end{aligned}$$

Let $\mathbf{P}^{(n)} = (p_{ij}^{(n)})$, i.e., the n step transition matrix. Then by Theorem 1,

$$\mathbf{P}^{(n)} = \mathbf{P}^{(n-1)}\mathbf{P} = \cdots = \mathbf{P}^n,$$

i.e., the n-th power of the matrix $\mathbf{P}$ is, in fact, the n step transition matrix.

For a Markov chain $\{X_n\}$ with transition probability matrix $\mathbf{P} = (p_{ij})$, what is the distribution of X_n?

$$
\begin{aligned}
P\{X_n = j\} &= \sum_{i \in S} P\{X_n = j | X_0 = i\} P\{X_0 = i\} \\
&= \sum_{i \in S} p_{ij}^{(n)} P\{X_0 = i\}.
\end{aligned}
$$

Let

$$\pi_j^{(n)} = P\{X_n = j\}, \qquad \boldsymbol{\pi}^{(n)} = (\pi_0^{(n)}, \pi_1^{(n)}, \cdots), n \geq 0.$$

We then have

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \mathbf{P}^n.$$

Sample path

When a communication system can be modeled by a DTMC with $\mathbf{P}$ and $S = 0, 1, 2$, what happens?

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

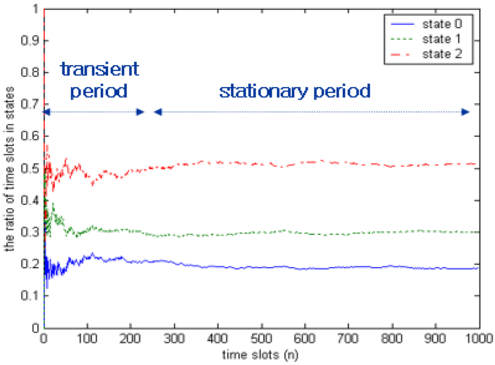

Figure: A sample path of a DTMC

Stationary Probabilities

Let $S = \{0, 1, 2, \cdots\}$ be the state space.

The stationary probability vector (distribtuion) $\boldsymbol{\pi}$ if it satisfies

- $0 \leq \pi_i < \infty$,
- $\sum_{i=0}^{\infty} \pi_i = 1$, and
- $\sum_{j \in S} \pi_j p_{ji} = \pi_i$ for any $i \in S$, i.e., $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.

The meaning of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$

- $\boldsymbol{\pi}$: the probability distribution of the Markov chain at the present time
- $\boldsymbol{\pi}\mathbf{P}$: the probability distribution of the Markov chain at the next time

The stationary distribution does not always exist. To check the existence of the stationary distribution we need to classify DTMCs according to its probabilistic properties such as

- irreducibility
- recurrence
- positive/null recurrence

In fact, if a DTMC is irrecudible and positive recurrent, the stationary distribution always exists and is unique.

Let $\{X_n\}$ be a discrete time Markov chain with transition probability matrix $\mathbf{P} = (p_{ij})$ and stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$. $\{X_n\}$ satisfies detailed balance (or is time-reversible) if

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

Suppose that we have a vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$. If we construct an irreducible, aperiodic, finite state DTMC with $\mathbf{P} = (p_{ij})$ that satisfies $\pi_i p_{ij} = \pi_j p_{ji}$, then the vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$ is the stationary distribution of the DTMC because

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j.$$

Gibbs Sampling

Gibbs sampling is a special case of the Metropolis–Hastings algorithm to be discussed later. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginal distribution by integrating over a joint distribution.

Suppose we want to obtain k samples of $\mathbf{x} = (x_1, \cdots, x_n)$ from a joint distribution $p(x_1, \cdots, x_n)$. Denote the i-th sample by $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \cdots, x_n^{(i)}\right)$.

Gibbs Sampling

We proceed as follows:

- We begin with some initial value $\mathbf{x}^{(i)}$.
- We generate a new sample $\mathbf{x}^{(i+1)}$ by sampling each component in turn. More formally, to sample $x_j^{(i+1)}$, we update it according to the distribution specified by $p\left(x_j^{(i+1)}|x_1^{(i+1)},\ldots,x_{j-1}^{(i+1)},x_{j+1}^{(i)},\ldots,x_n^{(i)}\right)$.
- Repeat the above step k times.

Gibbs Sampling

Note that

$$
\begin{aligned}
& p(x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_j^{(i)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}) \\
& \qquad \times p\left(x_j^{(i+1)} | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right) \\
& \quad = p(x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_j^{(i)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}) \\
& \qquad \times \frac{p\left(x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_j^{(i+1)}, x_j^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right)}{p\left(x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right)} \\
& \quad = p\left(x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_j^{(i+1)}, x_j^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right) \\
& \qquad \times p\left(x_j^{(i)} | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right)
\end{aligned}
$$

The Metropolis Algorithm

Suppose we know the distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots)$.

Algorithm

1. start with any initial value i with $\pi_i > 0$.
2. sample a candidate value j by using the proposal distribution q_{ij} (that is symmetric).
3. compute $\alpha_{ij} = \min\{\frac{\pi_j}{\pi_i}, 1\}$.
4. accept the new value j with probability α_{ij}, else reject the new value j and return to step 2.
5. start with the accepted value j and return to step 2.

The accepted values are denoted by $X_0 = i, X_1, X_2, \cdots$.

Analysis of The Metroplis Algorithm

The transition probability p_{ij} from state i to state j of the Markov chain: $p_{ij} = q_{ij}\alpha_{ij}$.

Let's check if it satisfies detailed balance, i.e., $\pi_i p_{ij} = \pi_j p_{ji}$.

If $\frac{\pi_j}{\pi_i} > 1$, then $\alpha_{ij} = 1$ and $\alpha_{ji} = \frac{\pi_i}{\pi_j}$. So

$$\pi_i p_{ij} = \pi_i q_{ij} = \pi_i q_{ji} = \pi_j q_{ji} \frac{\pi_i}{\pi_j} = \pi_j q_{ji} \alpha_{ji} = \pi_j p_{ji}.$$

If $\frac{\pi_j}{\pi_i} \le 1$, then $\alpha_{ij} = \frac{\pi_j}{\pi_i}$ and $\alpha_{ji} = 1$. So

$$\pi_i p_{ij} = \pi_i q_{ij} \frac{\pi_j}{\pi_i} = \pi_j q_{ij} = \pi_j q_{ji} = \pi_j p_{ji}.$$

Suppose that we know $\theta_i, i \geq 0$ that satisfy

$$\pi_i = c\theta_i$$

for some constant c that we don't know.

In this case we can still use the Metropolis algorithm because we use the ratio $\frac{\pi_j}{\pi_i}$ and

$$\frac{\pi_j}{\pi_i} = \frac{c\theta_j}{c\theta_i} = \frac{\theta_j}{\theta_i}.$$

The Metropolis-Hastings Algorithm

We drop the symmetry of the proposal distribution and use an arbitrary proposal distribution.

Algorithm

1. start with any initial value i with $\pi_i > 0$.
2. sample a candidate value j by using the proposal distribution q_{ij}.
3. compute $\alpha_{ij} = \min\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\}$.
4. accept the new value j with probability α_{ij}, else reject the new value j and return to step 2.
5. start with the accepted value j and return to step 2.

The accepted values are denoted by $X_0 = i, X_1, X_2, \cdots$.

Analysis of The M-H Algorithm

The transition probability p_{ij} from state i to state j of the Markov chain:

$$p_{ij} = q_{ij}\alpha_{ij}$$

.

If $\pi_i q_{ij} = \pi_j q_{ji}$, then $\alpha_{ij} = \alpha_{ji} = 1$. So

$$\pi_i p_{ij} = \pi_i q_{ij} = \pi_j q_{ji} = \pi_j p_{ji}.$$

If $\pi_i q_{ij} > \pi_j q_{ji}$, then $\alpha_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$ and $\alpha_{ji} = 1$. So

$$\pi_i p_{ij} = \pi_i q_{ij}\alpha_{ij} = \pi_i q_{ij}\frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \pi_j q_{ji} = \pi_j p_{ji}.$$

Choosing the proposal distribution

There are two general approaches.

- The first one is the random walk approach.

  When i is given, a new value j is obtained by $j = i + z$ where z has the pdf $g(z)$ with $g(z) = g(-z)$. In this case, $q_{ij} = g(j - i)$ and hence $q_{ij} = q_{ji}$. The Metropolis algorithm can be used.

- The second one is the independent chain approach.

  The proposal distribution q_{ij} satisfies $q_{ij} = g(j)$ for some distribution $g(j)$ regardless of the value i. Since $q_{ij} \neq q_{ji}$, the Metropolis-Hastings algorithm should be used.

Convergence to the steady state

Recall the behavior of a DTMC.

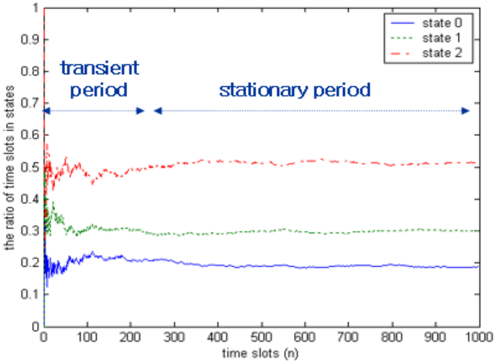

Figure: A sample path of a DTMC

Convergence to the steady state

Note that, starting from an initial value, the DTMC needs time (the transient or burn-in period) to enter the steady state.

The test for convergence: The Geweke test

1. Remove the burn-in period from the sample
2. Consider the first 10% and the last 50% and compute their sample means.
3. Test the difference between two sample means.

There are other tests for stationarity.

Independent sampling

Obviously, the sequence from the MH algorithm has correlation. However, we need a sample where the values are sampled independently.

A simple way to get an independent sample is thinning the sequence. That is, we sample only every m-th value in the sequence after the burn-in period.

- The sampled values are close to independent ones
- Save memory since we store a fraction of values
- Shown to increase the variance of the estimates

By Ergodic Theorem, we use all the sampled values after removing the burn-in period.

Theorem 2

(Ergodic Theorem) Let $\{X_n\}$ *be an irreducible and positive recurrent DTMC, and* $\boldsymbol{\pi}$ *be the stationary probability vector. Let* $f : \mathcal{S} \to \mathbb{R}$ *such that* $\sum_{i \in \mathcal{S}} |f(i)|\pi_i < \infty$. *Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{i\in\mathcal{S}} f(i)\pi_i \text{ a.s.}.$$