Lecture 3: Markov Chain Monte Carlo

Tianpei Xie

Sep. 28th., 2022

Contents

1 Basic Concept of Markov Chain

- ***Markov Chain*** $(X_t)_t$ is a ***probabilistic graphical model*** over a chain graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_C)$, where each random variable X_t only has exactly one children X_{t+1} and one parent X_{t-1}. Denote the index of variable t as the time. Markov chain $(X_t)_t$ is also a ***stochastic process***.

- By Markov property,

$$P(X_{t+1}|X_t, X_{t-1}, \ldots, X_1) = P(X_{t+1}|X_t).$$

  It is seen that the transition probability does not depend on the time t, i.e. Markov chain is **time-invariant**.

- Define the ***kernel*** of Markov Chain as the *time-invariant* ***transition probability***

$$K(x, y) = p(x, y) := P(X_{t+1} = y|X_t = x). \tag{1}$$

  Then the m***-step transition probability*** is defined as

$$K^m(x, y) = P(X_{t+m} = y|X_t = x). \tag{2}$$

- For $X_t \in \mathcal{X} := \{1, \ldots, n\}$ as discrete random variable with $|\mathcal{X}| = n$, we can define the ***transition matrix***

$$\boldsymbol{K} = [K(i, j)]_{n\times n} \tag{3}$$

- We have the ***Chapman-Kolmogorov equation*** [Ross, 2014]:

$$\begin{aligned} K^{m+n}(x, y) &= \sum_z K^m(x, z)K^n(z, y), \quad \forall x, y \in \mathcal{X} \\ \Rightarrow \boldsymbol{K}^{m+n} &= \boldsymbol{K}^m\boldsymbol{K}^n \end{aligned} \tag{4}$$

  That is, we split the $(m + n)$-step path from $x \to y$ into all possible combination of a m-step path from $x \to z$ and a n-step path from $z \to y$ for some intermediate state z.

- Define $T_j = \min\{t \geq 1 : X_t = j\}$ as the time steps for Markov Chain $(X_t)_t$ to *hit* state j for the ***first time***. T_j is called the state j's ***first hitting time***.

  Denote $f_{i,j}$ be the ***probability of ever hitting state*** j (***within finite time***) ***starting from state*** i. That is

$$f_{i,j} := P(T_j < \infty|X_0 = i) \tag{5}$$

  Denote $f_{i,j}^{(m)}$ be the *probability of hitting at state* j *at time* m starting from state i

$$f_{i,j}^{(m)} := P(T_j = m|X_0 = i) \tag{6}$$

- Define $N(y) := \sum_{t=0}^{\infty} \mathbb{1}\{X_t = y\}$ is the ***total number of times hitting the state*** y.

$$P(N(y) \geq 1|\, X_0 = x) = P(T_y < \infty|X_0 = x) = f_{x,y} \tag{7}$$

$$P(N(y) \geq m|\, X_0 = x) = P(N(y) \geq 1|\, X_0 = x)P^{m-1}(N(y) \geq 1|\, X_0 = y) \tag{8}$$

$$= f_{x,y}\, f_{y,y}^{m-1}$$

Note that in order to visit y at least m times, we need to visit y first time and stating from y recurrently visit y $(m-1)$ times.

The random variable $N(x)|X_0 = x$ follows a ***geometric distribution*** with mean $1/(1-f_{x,x})$.

$$P(N(x) = m|\, X_0 = x) = (1 - f_{x,x}) f_{x,x}^{m-1} \tag{9}$$

- For any pair $x, y \in \mathcal{X}$, if there exists $n \in \mathbb{N}_+$ so that $K^n(x,y) > 0$, then the state y is ***accessible*** from state x. This is equivalent to say that the probability of hitting time of y being finite starting from x is above zero, i.e. $f_{x,y} > 0$.
- If x is accessible from y, and y is accessible from x, then we say that x and y ***coummunicate***, $x \leftrightarrow y$. It is easy to check that this is an ***equivalence relation***.
- A Markov Chain is ***irreducible*** if it has **only one equivalence class**, i.e. all states in $\mathcal{X}$ communicate to each other.
- Based on hitting time, we can categorize states into two groups:
  - A state i is ***recurrent*** if and only if $f_{i,i} = P(T_i < \infty | X_0 = i) = 1$, i.e. the Markov Chain will definitely revisit the state i after stating from i.
  - A recurrent state i is ***positive recurrent*** if the *expected returning time* $\mathbb{E}\left[T_i | X_0 = i\right] < \infty$; otherwise we say it is ***null recurrent***.
  - A state i is called ***transient*** if $f_{i,i} < 1$.
- **Proposition 1.1** *If i is recurrent, and $i \to j$, then j is also recurrent. Therefore, in any* ***equivalent class, either all states are recurrent*** *or* ***all are transient****. In particular, if the chain is* **irreducible**, *then either all states are recurrent or all are transient.*

  Proposition 1.2 *An* ***irreducible finite state Markov chain*** *must be* ***positive recurrent****.*
- **Definition** The probability of states $\{\pi(x), \forall x \in \mathcal{X}\}$ is a ***stationary distribution*** if and only if

$$\pi(y) = \sum_{x \in X} K(x,y)\pi(x), \forall y \in \mathcal{X} \tag{10}$$

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \boldsymbol{K} \tag{11}$$

  That is, $\boldsymbol{\pi}$ is the eigenvector of stochastic matrix $\boldsymbol{K}$ corresponding to eigenvalue $\lambda_0 = 1$.

  The ***stationary distribution does not change over time***.
- **Proposition 1.3** *Suppose that the* ***limit distribution*** $\lim_{t\to\infty} P(X_t = y)$ *exists, and*

$$\lim_{t\to\infty} K^t(x,y) = \pi(y), \quad \forall x, y \in \mathcal{X}$$

  which is independent of where it starts from, then the Markov Chain has a ***unique stationary distribution*** *and*

$$\lim_{t\to\infty} P(X_t = y) = \pi(y), \quad \forall y \in \mathcal{X} \tag{12}$$

  i.e. the limit distribution is stationary distribution.

- **Proposition 1.4** *(**Global Balance Equation**)*
  The stationary distribution $\{\pi(x), \forall x \in \mathcal{X}\}$ *satisfies the following* ***global balance equation****:*

$$\sum_{j \in \mathcal{X}} \pi(i)K(i,j) = \sum_{j \in \mathcal{X}} \pi(j)K(j,i). \tag{13}$$

  This means the total flow out of i *(LHS) is equal to the total flow into* i *(RHS) in steady state.*

- **Proposition 1.5** *(**Detailed Balance Equation**)*
  For distribution $\{\pi(x), \forall x \in \mathcal{X}\}$*, if the following* ***detailed balance equation*** *is satisfied*

$$\pi(i)K(i,j) = \pi(j)K(j,i), \quad \forall i, j \in \mathcal{X} \tag{14}$$

  then $\{\pi(x), \forall x \in \mathcal{X}\}$ *is a stationary distribution.*

- **Theorem 1.6** *(Stationary distribution for transient and null recurrent states)*
  Let $\{\pi(x), \forall x \in \mathcal{X}\}$ *be stationary distribution. If* $x \in \mathcal{X}$ *is* ***transient*** *or* ***null recurrent*** *state, then*

$$\pi(x) = 0.$$

- **Theorem 1.7** *[Ross, 2014]*
  An ***irreducible recurrent*** *Markov Chain has a* ***unique stationary distribution*** $\{\pi(x)\}$*, given*

$$\pi(x) = \frac{1}{\mu_x}, \quad \forall x \in \mathcal{X} \tag{15}$$

  where $\mu_x := \mathbb{E}\left[T_x | X_0 = x\right]$ *is the* ***expected first return time*** *of state* x*.*

  It implies that as $n \to \infty$, for any state $x \in \mathcal{X}$, the fraction of time that Markov Chain stays at x is unchanged and is the reciprocal of the expected first return time.

- **Definition** The ***periodicity*** of a state $x \in \mathcal{X}$ is defined as

$$d(x) = \text{g.c.d.}\left\{t \geq 0 : K^t(x,x) > 0\right\} \tag{16}$$

  where g.c.d. is the ***greatest common divisor***.

- **Definition** If $d(x) \geq 2$, then state x is ***periodic***. If $d(x) = 1$, then state x is ***aperiodic***

  The periodicity property is *closed* under the equivalence class C.

- **Definition** A Markov Chain is ***irreducible***, ***positive recurrent*** and ***aperiodic***, then it is called ***ergodic***.

- **Theorem 1.8** *A Markov Chain is* ***ergodic*** *having stationary distribution* $\boldsymbol{\pi}$*, then*

$$\lim_{t \to \infty} K^t(x,y) = \pi(y), \quad \forall x, y \in \mathcal{X} \tag{17}$$

  That is, when a Markov Chain is ergodic, its marginal state distribution will converge to the stationary distribution.

- **Definition** A Markov Chain $(X_t)_t$ is called ***time-reversible***, if it has stationary distribution $\boldsymbol{\pi}$ and the detailed balance equation is satisfied:

$$\pi(i)K(i,j) = \pi(j)K(j,i), \quad \forall i,j \in \mathcal{X} \tag{18}$$

- The reversed process $(Y_k)_k := (X_{t-k})_k$ is a Markov Chain and its transition probability

$$Q(i,j) = \frac{\pi(j)K(j,i)}{\pi(i)} \tag{19}$$

  Note that $(Y_k)_k$ and $(X_t)_t$ are statistically equivalent since $Q(i,j) = K(i,j)$.

- **Theorem 1.9** *An ergodic Markov Chain $(X_t)_t$ for which $K(i,j) = 0$ whenever $K(j,i) = 0$ is **time-reversible** if and only if starting from any state i, any path back to i has the **same probability** as its reverse path. That is, for path $i \to i_1 \to i_2 \to \ldots \to i_k \to i$ and its reverse path $i \leftarrow i_1 \leftarrow i_2 \leftarrow \ldots \leftarrow i_k \leftarrow i$*

$$K(i,i_1)\,K(i_1,i_2)\,\ldots\,K(i_k,i) = K(i,i_k)\,\ldots\,K(i_2,i_1)\,K(i_1,i), \quad \forall i,i_1,\ldots,i_k \in \mathcal{X} \tag{20}$$

- **Theorem 1.10** ***(Reversal Test)***
  Let $\boldsymbol{K}$ be a stochastic matrix indexed by a countable set $\mathcal{X}$ and let $\boldsymbol{\pi}$ be a probability distribution on $\mathcal{X}$. Let $\boldsymbol{Q}$ be a stochastic matrix indexed by $\mathcal{X}$ such that

$$\pi(i)Q(i,j) = \pi(j)K(j,i), \quad \forall i,j \in \mathcal{X}. \tag{21}$$

  Then $\boldsymbol{\pi}$ is a stationary distribution of $\boldsymbol{K}$

- **Theorem 1.11** ***(Ergodic Theorem)*** *[Robert and Casella, 1999]*
  If $(X_t)_t$ is Harris recurrent with a σ-finite invariant measure π, then for any $f,g \in L_1(\pi)$ with $\mathbb{E}_\pi[g] \neq 0$,

$$\lim_{T\to\infty} \frac{S_T(f)}{S_T(g)} = \frac{\mathbb{E}_\pi[f]}{\mathbb{E}_\pi[g]} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)} \tag{22}$$

  It can be shown that if $(X_t)_t$ is ***Harris positive*** with **stationary distribution** π and if $S_T(h)$ *converges* μ_0-almost surely (μ_0 a.s.) to $\mathbb{E}_\pi[h]$, for an initial distribution μ_0, this convergence occurs for ***every initial distribution*** $\mu\cdot$

- **Theorem 1.12** ***(Central Limit Theorem for reversible chains)****[Robert and Casella, 1999]*
  *If $(X_t)_t$ is **aperiodic, irreducible, and reversible** with stationary distribution $\boldsymbol{\pi}$, the **Central Limit Theorem** applies when*

$$0 < \sigma^2 = \mathbb{E}_{\boldsymbol{\pi}}\left[g^2(X_t)\right] + 2\sum_{s=1}^{\infty} \mathbb{E}_{\boldsymbol{\pi}}\left[g(X_t)\,g(X_{t+s})\right] < \infty. \tag{23}$$

2 Markov Chain Monte Carlo

2.1 From Vanilla Monte Carlo to MCMC

- **Definition** ***A Markov Chain Monte Carlo (MCMC) method*** for the simulation of a distribution f is any method producing an ergodic Markov chain $(X_t)_t$ whose stationary distribution is f.

- Compared to vanilla Monte Carlo (e.g. inverse transform, reject sampling, importance sampling), MCMC has the following **characteristics**:
  - Unlike vanilla Monte Carlo methods, which reply on i.i.d samples, ***Markov Chain Monte Carlo (MCMC) methods*** generate ***dependent*** **samples** via Markov chain.
  - The MCMC updates ***preserve the probability measure*** π when convergence is attained. That is, *when the Markov chain* ***converges***, the distribution of X_t is the same as the distribution of $X_{t+1}, X_{t+2}, \ldots$. Thus we have obtained a sequence of ***identically distributed (but dependent) samples***. When Markov chain converges (***mixing***), we can use samples the same way as we did in vanilla Monte Carlo to approximate the expectation. In particular, an MCMC estimator is

$$J_T = \widehat{\mathbb{E}}_{\boldsymbol{\pi}}\left[h(X)\right] = \frac{1}{T}\sum_{t=0}^{T} h(X_t). \tag{24}$$

    The ergodic theorem guarantees the (almost sure) convergence of the empirical average to $\mathbb{E}_{\boldsymbol{\pi}}\left[h(X)\right]$ where $\boldsymbol{\pi}$ is the stationary distribution. A sequence $(X_t)_t$ produced by a Markov chain Monte Carlo algorithm can thus be employed just as an i.i.d sample.
  - Similar to importance sampling, we can approximate the expectation $\mathbb{E}_f\left[h(X)\right]$ using an alternative proposal distribution g which is the stationary distribution of an ergodic Markov chain. This is the idea behind *Metropolis-Hastings algorithm*.
- Markov Chain Monte Carlo (MCMC) methods are ***preferred*** in following situations:
  - The target distribution is ***high dimensional***. Due to *the curse of dimensionality*, the variance, which is a function of dimension d, will grow exponentially as the dimensionality increases. Moreover, many high dimensional joint distributions are usually not represented in explicit function form due to their complicated partition functions. In this situation, finding a proposal distribution that is close to the target distribution in high dimensional space is also very challenging.
  - Some stochastic optimization algorithms **naturally produce Markov chain structures**. It is a general fact that the use of Markov chains allows for a greater scope than the methods presented in vanilla Monte Carlo.
  - Vanilla Monte Carlo and MCMC algorithms both satisfy the $O(1/\sqrt{n})$ convergence requirement for the approximation of J. There are thus many instances where a specific MCMC algorithm dominates, variance-wise, the corresponding Monte Carlo proposal.

2.2 Metropolis-Hastings Algorithm

- Consider the following energy-based model

$$\pi(\boldsymbol{x}) = \frac{1}{Z}\exp(-h(\boldsymbol{x})),$$
$$\text{where } Z = \int_{\mathcal{X}} \exp(-h(\boldsymbol{x}))d\boldsymbol{x}$$

  is the partition function in high dimensional space.
- The basic idea for Metropolis algorithm is to simulate π using the stationary distribution g from a Markov chain. Compare to analysis of Markov chain itself, which often starts from a

known transition kernel, the Metropolis algorithm starts from a known stationary distribution g and is interested in how to prescribe an efficient transition kernel to reach the equilibrium.

- Starting from an initial configuration $\boldsymbol{X}_0$, the ***Metropolis Algorithm*** iteratively repeats the following steps [Liu, 2001]:

  1. Propose a random unbiased "perturbation" of $\boldsymbol{X}_t$ so as to generate a new configuration $\boldsymbol{X}^{'}$. $\boldsymbol{X}^{'}$ is seen as generated by a *symmetric* probability transition kernel (or ***proposal function***) $K(\boldsymbol{X}_t\,, \boldsymbol{X}^{'})$, i.e. $K(x, y) = K(y, x)$. Then calculate the change $\Delta h := h(\boldsymbol{X}^{'}) - h(\boldsymbol{X}_t)$;

  2. Generate a uniform random variable $U \in \mathcal{U}[0, 1]$. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}^{'}$ if

  $$U \leq \pi(\boldsymbol{X}^{'})/\pi(\boldsymbol{X}_t) := \exp\left(-\Delta h\right);$$

  Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

- Heuristically, the Metropolis Algorithm is based on a *"trial-and-error" strategy*: at each iteration, a random perturbation of $\boldsymbol{X}_t$ is generated by Markov chain. Then the gain is computed for this new sample. If the gain is large enough, it will be accepted by high probability. Otherwise, we keep using the old sample $\boldsymbol{X}_t$.

- The requirement for symmetric transition kernel means that the chance of getting $\boldsymbol{X}^{'}$ from $\boldsymbol{X}_t$ and the chance of getting $\boldsymbol{X}_t$ from $\boldsymbol{X}^{'}$ are equal. This could avoid the "trend bias" at the proposal stage.

- Hastings generalize the idea of Metropolis Algorithm by removing the requirement for the symmetric proposal function K. In Hastings' generalization, it only requires that $K(x, y) > 0 \Leftrightarrow K(y, x) > 0$.

- The ***Metropolis-Hastings Algorithm*** is described as below:

  1. Given current configuration $\boldsymbol{X}_t$, draw $\boldsymbol{Y}$ from the proposal function $K(\boldsymbol{X}_t\,, \boldsymbol{Y})$.

  2. Compute the ***Hastings ratio*** (*acceptance function*):

  $$r(\boldsymbol{X}_t\,, \boldsymbol{Y}) := \frac{\pi(\boldsymbol{Y})\, K(\boldsymbol{Y}, \boldsymbol{X}_t)}{\pi(\boldsymbol{X}_t)\, K(\boldsymbol{X}_t\,, \boldsymbol{Y})} \tag{25}$$

  3. (***Metropolis Rejection***) Generate a uniform random variable $U \in \mathcal{U}[0, 1]$.
  Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$ if

  $$U \leq \alpha(\boldsymbol{X}_t\,, \boldsymbol{Y}),$$

  where

  $$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\left\{1,\ r(\boldsymbol{X}_t\,, \boldsymbol{Y})\right\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

  Clearly if K is symmetric $K(\boldsymbol{Y}\,, \boldsymbol{X}_t) = K(\boldsymbol{X}_t\,, \boldsymbol{Y})$, the ratio (**??**) is equal to $\exp(-\Delta h)$. The Metropolis Algorithm is a special case of Metropolis-Hastings Algorithm.

- The intuition behind the ratio $K(\boldsymbol{y}, \boldsymbol{x}_t)/K(\boldsymbol{x}_t, \boldsymbol{y})$ is that it compensates the "flow bias" from the proposal distribution.

- **Theorem 2.1** *Let $(\boldsymbol{X}_t)_t$ be the chain produced by the Metropolis-Hastings algorithm. For every conditional distribution $q(\boldsymbol{y}|\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{y})$, whose support includes $\mathcal{X}$,*

  1. *the kernel K of the chain satisfies the* ***detailed balance condition*** *with π;*
  2. *π is a stationary distribution of the chain.*

  Proof: For (1), we see that the *transition probability* of $(\boldsymbol{X}_t)_t$ is computed as

$$\begin{aligned} A(\boldsymbol{x}\,, \boldsymbol{y}) &= K(\boldsymbol{x}\,, \boldsymbol{y})\, \alpha(\boldsymbol{x}\,, \boldsymbol{y}) \\ &= K(\boldsymbol{x}\,, \boldsymbol{y})\ \min\left\{1,\ \frac{\pi(\boldsymbol{y})\, K(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})\, K(\boldsymbol{x}\,, \boldsymbol{y})}\right\} \\ &= \frac{\min\left\{\pi(\boldsymbol{x})\, K(\boldsymbol{x}\,, \boldsymbol{y})\,, \pi(\boldsymbol{y})\, K(\boldsymbol{y}, \boldsymbol{x})\right\}}{\pi(\boldsymbol{x})} := \frac{\delta(\boldsymbol{x}, \boldsymbol{y})}{\pi(\boldsymbol{x})}. \\ &= \pi(\boldsymbol{y}) \min\left\{\frac{K(\boldsymbol{x}\,, \boldsymbol{y})}{\pi(\boldsymbol{y})}\,, \frac{K(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})}\right\} \end{aligned} \tag{26}$$

  We can see that since $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{y}, \boldsymbol{x})$ the *detailed balance equation* holds:

$$\pi(\boldsymbol{x}) A(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) A(\boldsymbol{y}, \boldsymbol{x}). \quad \blacksquare \tag{27}$$

  For (2), it follows that when detailed balance equation is satisfied, then π is a stationary distribution for Markov chain $(\boldsymbol{X}_t)_t$. $\blacksquare$

  The stationarity of π is therefore established for *almost any conditional distribution* q, a fact which indicates the ***universality*** of Metropolis-Hastings algorithms.

- We can check on the properties of the Markov chain $(\boldsymbol{X}_t)_t$ from Metropolis-Hastings algorithm:
  - (***Irreducibility***): To make sure the *irreducibility* to hold, the domain $\mathcal{X}$ of target distribution π need to be ***connected***. Otherwise, the Markov chain will not be able to reach other connected components. In other words, the stationary distribution π is ***not multimodal distribution***.
  - (***Positive recurrent***): A sufficient condition for both the *irreducibility* and *positive recurrence* to hold is the positivity of transition kernel:

$$q(\boldsymbol{y}|\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{y}) > 0 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{28}$$

    This makes sure that every state can be visited in one step. Thus the Markov chain is positive recurrent. In fact, it can be shown in [Robert and Casella, 1999] that if $(\boldsymbol{X}_t)_t$ is π-*irreducible*, then it is *Harris recurrent*.

    Lemma 2.2 *[Robert and Casella, 1999]*
    If $(\boldsymbol{X}_t)_t$ from MH algorithm is π-irreducible, then it is Harris recurrent.

  - (***Aperiodic***): Unlike the Rejection Sampling, the Metropolis Rejection does not simply regenerate a new proposal repeatedly until some of them is accepted. Instead, it reuse the value from the old state. An attempt to make Metropolis Rejection like the Rejection sampling will destroy the property that **this update will preserve the distribution** π

A *sufficient condition* for ***aperiodic*** property to hold is to have $P\{\boldsymbol{X}_{t+1} = \boldsymbol{X}_t\} > 0$, and thus

$$P\{\pi(\boldsymbol{X}_t)\, K(\boldsymbol{X}_t, \boldsymbol{Y}) \leq \pi(\boldsymbol{Y})\, K(\boldsymbol{Y}, \boldsymbol{X}_t)\} < 1 \tag{29}$$

This is **critical** to guarantee the aperiodic property of the Markov chain, which is essential for *Ergodic theorem* to hold. Note that this condition also make sure K is *not the transition kernel* for the Markov chain induced by Metropolis-Hastings algorithm.

– (***Reversibility***): From the theorem, we see that $(\boldsymbol{X}_t)_t$ is ***time-reversible*** with π as its ***invariant distribution***.

- Note that due to the rejection rule, $A(\boldsymbol{x}, \boldsymbol{y}) \neq K(\boldsymbol{x}, \boldsymbol{y})$, i.e. K is not the transition kernel for the Markov chain induced by Metropolis-Hastings algorithm.

- In fact, as long as $A(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y})\delta(\boldsymbol{x}, \boldsymbol{y})$ for some symmetric function δ, the detailed balance equation will hold. The challenge, however, lies in the constraint on $\delta(\boldsymbol{x}, \boldsymbol{y})$ so that $\int A(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y} = \int \pi(\boldsymbol{y})\delta(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y} = 1$.

  For instance, We can replace the acceptance function (25) with a more general form

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{\delta(\boldsymbol{x}, \boldsymbol{y})}{\pi(\boldsymbol{x})\, K(\boldsymbol{x}, \boldsymbol{y})}$$

  where $\delta(\boldsymbol{x}, \boldsymbol{y})$ is any symmetric function that makes $r(\boldsymbol{x}, \boldsymbol{y}) \leq 1$. In this case, the probability of generating $\boldsymbol{y}$ given $\boldsymbol{x}$ is

$$A(\boldsymbol{x}, \boldsymbol{y}) := K(\boldsymbol{x}, \boldsymbol{y})r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\delta(\boldsymbol{x}, \boldsymbol{y})}{\pi(\boldsymbol{x})}.$$

  Since $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{y}, \boldsymbol{x})$, we have $\pi(\boldsymbol{x})A(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y})A(\boldsymbol{y}, \boldsymbol{x})$.

- Finally, we see that the convergence of estimator (24) based on ergodic theorem.

  Theorem 2.3 *[Robert and Casella, 1999]*
  Suppose that the Metropolis-Hastings Markov chain $(\boldsymbol{X}_t)_t$ is π-irreducible.

  1. *If $h \in L_1(\pi)$, then*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} h(\boldsymbol{X}_t) = \mathbb{E}_{\boldsymbol{\pi}}\left[h(\boldsymbol{X})\right] \tag{30}$$

  2. *If, in addition, $(\boldsymbol{X}_t)_t$ is **aperiodic**, then*

$$\lim_{t\to\infty} \left\| \int_{\mathcal{X}} K^t(\boldsymbol{x}, \cdot)d\mu(\boldsymbol{x}) - \pi \right\|_{TV} = 0 \tag{31}$$

  *for **every initial distribution** μ, where $K^t(\boldsymbol{x}, \cdot)$ denotes the kernel for t transitions.*

- **Corollary 2.4** *[Robert and Casella, 1999]*
  The conclusions of Theorem 2.3 hold if the Metropolis-Hastings Markov chain $(\boldsymbol{X}_t)_t$ has conditional density $q(\boldsymbol{y}|\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{y})$ that satisfies (28) *and* (29).

3 Some Special Metropolis-Hastings Algorithms

3.1 Random-walk Metropolis-Hastings

- A natural approach for the practical construction of a Metropolis-Hastings algorithm is to take into account the value previously simulated to generate the following value; that is, to consider a ***local exploration*** of the ***neighborhood*** of the current value of the Markov chain. This idea is already used in algorithms such as the *simulated annealing algorithm* and the *stochastic gradient method*.

- The *Random-walk Metropolis-Hastings algorithm* is very commonly used when ***no prior knowledge*** is available. In this case, the preferred proposal is just a *uniformly local* move, since it is ***uninformative*** and ***unbiased***.

- For Euclidean space $\mathcal{X} \subset \mathbb{R}^d$, the *random walk* model is [Robert and Casella, 1999]:

$$\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{\epsilon}_t, \tag{32}$$

  where the noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$ are independent from the starting position $\boldsymbol{X}_t$. The transition probability $q(\boldsymbol{y}|\boldsymbol{X}_t) = \mathcal{N}(\boldsymbol{X}_t, \sigma^2 \boldsymbol{I}_d) := g(\|\boldsymbol{y} - \boldsymbol{X}_t\|_2)$, where $g(\cdot)$ is a symmetric function $g(-r) = g(r)$.

- In general, define a ***symmetric distribution*** g_σ so that $q(\boldsymbol{y}|\boldsymbol{x}) = g_\sigma(\|\boldsymbol{y} - \boldsymbol{x}\|_2) > 0$ for all $\|\boldsymbol{y} - \boldsymbol{x}\|_2 < \delta$. Here σ controls the "***range***" of the exploration. The random walk with transition kernel as g_σ will produce an ergodic Markov chain.

  The most common distributions g in this setup are the ***uniform distributions on spheres*** centered at the origin or standard distributions like the ***normal*** and the ***Student's t distributions***. *All these distributions usually need to be* ***scaled***.

- The ***Random-walk Metropolis-Hastings*** is described as below:

  1. Draw $\boldsymbol{\epsilon}_t$ from $g_\sigma(\boldsymbol{\epsilon})$ and set $\boldsymbol{Y} := \boldsymbol{X}_t + \boldsymbol{\epsilon}_t$;
  2. Compute the ratio:

$$r(\boldsymbol{X}_t, \boldsymbol{Y}) := \frac{\pi(\boldsymbol{Y})}{\pi(\boldsymbol{X}_t)} \tag{33}$$

  3. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$ with probability

$$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\{1,\ r(\boldsymbol{X}_t, \boldsymbol{Y})\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

- Random-walk Metropolis is not only simple to implement, it also has a particularly nice **intuition**. The **proposol distribution** is biased towards ***large volumes*** ($g_\sigma(\cdot)$), and hence the *tails* of the target distribution, while the *Metropolis correction* **rejects** those proposals that jump into neighborhoods where ***the density is too small*** ($\pi(\boldsymbol{x}^{(t)})$). The combined procedure then preferentially selects out those proposals that fall into neighborhoods of high probability mass, concentrating towards the typical set as desired.

- **Theorem 3.1** *[Robert and Casella, 1999]*
  Consider a ***symmetric*** *density π which is α-****log-concave*** *with associated constant α, i.e.*

$\pi(\boldsymbol{x}) = \exp(-h(\boldsymbol{x}))$ *for* $h(\cdot)$ *is a convex function, and*

$$\log \pi(\boldsymbol{x}) - \log \pi(\boldsymbol{y}) \geq \alpha \|\boldsymbol{y} - \boldsymbol{x}\| \tag{34}$$

for $\|\boldsymbol{x}\|$ *large enough. If the density* g *is* ***positive*** *and* ***symmetric****, the chain* $(\boldsymbol{X}_t)_t$ *of Random-walk Metropolis-Hastings is* ***geometrically ergodic****. If* g *is not symmetric, a sufficient condition for geometric ergodicity is that* $g(r)$ *be bounded by* $b\exp(-\alpha\,|r|)$ *for a sufficiently large constant* b*.*

3.2 Independent Metropolis-Hastings (IMH) Algorithm

- ***Independent Metropolis-Hastings (IMH) Algorithm***:

  1. Generate $\boldsymbol{Y}_t$ from $g(\boldsymbol{y})$;
  2. Compute ratio:

  $$r(\boldsymbol{X}_t, \boldsymbol{Y}) := \frac{\pi(\boldsymbol{Y}_t)\, g(\boldsymbol{X}_t)}{\pi(\boldsymbol{X}_t)\, g(\boldsymbol{Y}_t)} = \frac{w(\boldsymbol{Y}_t)}{w(\boldsymbol{X}_t)} \tag{35}$$

  where $w(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{g(\boldsymbol{x})}$ is the *importance weight*.

  3. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$ with probability

  $$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\{1,\ r(\boldsymbol{X}_t, \boldsymbol{Y})\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

- *Independent Metropolis-Hastings (IMH)* choose the proposal transition function $K(\boldsymbol{x}, \boldsymbol{y})$ as an ***independent*** *density* $g(\boldsymbol{y})$. That is, the proposal move $\boldsymbol{Y}_t$ is generated ***independently*** from previous sample $\boldsymbol{X}_t$.

- This method is an **alternative** to the *Rejection Sampling* and *Importance Sampling*. As with Rejection sampling, the performance of *IMH* depends on how close the proposal distribution g to the target distribution π. It is also suggested that g has longer tail than π for robustness of performance.

  It can be shown that if the envelop condition $\pi(\boldsymbol{x}) \leq M\, g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ holds, then the Metropolis-Hastings Markov chain $(\boldsymbol{X}_t)_t$ is ergodic, thus the convergence to $\mathbb{E}_\pi[h]$ is guaranteed.

  Theorem 3.2 *The Independent Metropolis-Hastings algorithm produces a* ***uniformly ergodic*** *chain if there exists a constant* M *such that*

  $$\pi(\boldsymbol{x}) \leq M\, g(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{36}$$

  In this case,

  $$\left\|K^t(\boldsymbol{x}, \cdot) - \pi\right\|_{TV} \leq 2\left(1 - \frac{1}{M}\right)^t$$

  where $\|\cdot\|_{TV}$ *denotes the total variation norm. On the other hand, if for every* M*, there exists a set of positive measure where* (36) *does not hold,* $(\boldsymbol{X}_t)_t$ *is not even geometrically ergodic.*

- Compared to *Rejection Sampling* and *Importance Sampling*, the *Independent Metropolis-Hastings (IMH)* has ***larger expected acceptance probability*** (at least $1/M$ when the chain is stationary) [Robert and Casella, 1999]. Thus IMH is more ***sample efficient***.

3.3 Configurational bias Monte Carlo

- The *Configurational bias Monte Carlo (CBMC)* can be seen as a ***SIS-based IMH***. Assume that we have auxiliary distributions $\pi_s(\boldsymbol{x}_{1:s})$ where $\boldsymbol{x}_{1:s} = [x_1, \ldots, x_s]$ at each time s. $\boldsymbol{x} := \boldsymbol{x}_{1:d}$. The proposal sampling probability $g_s(x_s|\boldsymbol{x}_{1:(s-1)})$ for all s. Here the proposal joint distribution can be computed sequentially

$$g(\boldsymbol{x}) = g_1(x_1) \prod_{s=2}^{d} g_s(x_s|\boldsymbol{x}_{1:(s-1)}) \tag{37}$$

- The ***Configurational bias Monte Carlo (CBMC)*** is described as below:

  1. Generate $\boldsymbol{Y}$ according to $g(\boldsymbol{y})$ via the sequential process in (37).
  2. Compute the *importance weight*:

$$w(\boldsymbol{Y}) = \frac{\pi(\boldsymbol{Y})}{g(\boldsymbol{Y})} = \frac{\pi_1(Y_1)}{g_1(Y_1)} \prod_{s=2}^{d} \frac{\pi_s(\boldsymbol{Y}_{1:s})}{\pi_s(\boldsymbol{Y}_{1:(s-1)})\, g_s(Y_s \,|\, \boldsymbol{Y}_{1:(s-1)})} \tag{38}$$

     Similarly compute the importance weight for $\boldsymbol{X}_t$ as $w(\boldsymbol{X}_t)$.

  3. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$ with probability

$$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\left\{1,\ \frac{w(\boldsymbol{Y})}{w(\boldsymbol{X}_t)}\right\}$$

  4. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

- We can modify this process to make **stage-wise acceptance decision**. Suppose that, at time $t-1$, $\boldsymbol{X}_t = (X_1^{(t)}, \ldots, X_d^{(t)})$ is accepted. Then at s-th stage, we accept $\boldsymbol{Y}_{1:s}$ with probability

$$\alpha_s(\boldsymbol{X}_{1:s}^{(t)}, \boldsymbol{Y}_{1:s}) = \min\left\{1,\ \frac{u_s(\boldsymbol{Y}_{1:s})}{u_s(\boldsymbol{X}_{1:s}^{(t)})}\right\} \tag{39}$$

  where

$$u_s(\boldsymbol{Y}_{1:s}) = \frac{\pi_s(\boldsymbol{Y}_{1:s})}{\pi_s(\boldsymbol{Y}_{1:(s-1)})\, g_s(Y_s \,|\, \boldsymbol{Y}_{1:(s-1)})}.$$

  That is, the acceptance rate at each stage is equal to the *ratio of incremental weights* between proposed state and the old state.

  If ***rejected***, we *go back to the **first stage** to rebuild the whole configuration*.

  Note that it is **not necessary** to make accept-reject decision **for each stage**.

- Note that the **acceptance rate** for multi-stage CBMC is *less than* the original CBMC, since $\prod_s \min\{1, r_s\} \le \min\{1, \prod_s u_s\}$.

References

Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.