Unit 2: Introduction to Metropolis–Hastings Algorithms

2.1 Markov Chains

Let $X_0, X_1, \ldots$ be measurable mappings (i.e., random variables) from an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to some measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra. Let $\mathcal{F}_t = \sigma(X_0, X_1, \ldots, X_t)$ for each t. We say $(X_t)_{t\geq 0}$ is a (homogeneous) Markov chain with transition kernel $P\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$, if for every t and $B \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}(X_t \in B \,|\, \mathcal{F}_{t-1}) = \mathbb{P}(X_t \in B \,|\, X_{t-1}) = P(X_{t-1}, B), \text{ a.s.}$$

In other words, $P(x, B)$ is the probability of moving to the set B in the next step given that the current state is x. Note that for every x, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We will not consider non-homogeneous Markov chains in this course, whose transition kernels may change over time.

Notations that are often used in the literature include:

$$\begin{aligned}
\nu(f) &\coloneqq \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x). \\
(\nu P)(B) &\coloneqq \int_{\mathcal{X}} P(x, B)\nu(\mathrm{d}x). \\
(Pf)(x) &\coloneqq \mathbb{E}[f(X_1) \,|\, X_0 = x] = \int_{\mathcal{X}} f(y)P(x, \mathrm{d}y). \\
P^t(x, B) &\coloneqq \mathbb{P}(X_t \in B \,|\, X_0 = x) = \int_{\mathcal{X}} P(y, B)P^{t-1}(x, \mathrm{d}y). \qquad (1)
\end{aligned}$$

In the above definitions, ν is any measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, f is any real-valued measurable function, and B is any set in $\mathcal{B}(\mathcal{X})$. When ν is a probability measure, we can interpret $(\nu P)(B)$ as the probability of $X_1 \in B$ when we draw the initial value X_0 from ν. The second equality in (1) is known as Chapman–Kolmogorov equation. Define the total variation distribution between two probability measures π and ν by

$$\|\nu - \pi\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathcal{X})} |\nu(A) - \pi(A)|.$$

2.2 Stationary Distributions of Markov Chains

Definition 2.1. We say a probability measure π is a stationary (or invariant) distribution of the transition kernel P if $(\pi P)(B) = \pi(B)$ for every $B \in \mathcal{B}(\mathcal{X})$.

If we initialize a Markov chain with transition kernel P by drawing X_0 from the stationary distribution π, then X_t has marginal distribution π for every t; in this case, we say $(X_t)_{t\geq 0}$ is a stationary process. More importantly, under some conditions on P, π is unique and the distribution of X_t will converge to π in total variation distance, as $t \to \infty$, regardless of the initial distribution (that is, $\|\nu P^t - \pi\|_{\mathrm{TV}} \to 0$ for any probability distribution ν). So, if direct

sampling from π is difficult, we may consider simulating a Markov chain with stationary distribution π. For now, we choose not to worry about those technical conditions that ensure the convergence, since they are satisfied for almost every MCMC algorithm used in practice. A more important question is how to verify π is the stationary distribution of P, which is the first (and perhaps most important) step in the development of MCMC algorithms. One approach is to verify a stronger condition known as reversibility.

Definition 2.2. We say P is reversible with respect to a probability measure π, if

$$\int_{y\in B}\int_{x\in A}\pi(\mathrm{d}x)P(x,\mathrm{d}y)=\int_{y\in A}\int_{x\in B}\pi(\mathrm{d}x)P(x,\mathrm{d}y), \tag{2}$$

for any $A, B \in \mathcal{B}(\mathcal{X})$.

Lemma 2.1. *If P is reversible with respect to π, then π is a stationary measure of P.*

Proof. Letting $A = \mathcal{X}$ in (2), we get $(\pi P)(B) = \pi(B)$. □

Let $X_0 \sim \pi$ and $X_1 \sim P(X_0, \cdot)$. Condition (2) can be equivalently expressed as

$$\mathbb{P}(X_0 \in A, X_1 \in B) = \mathbb{P}(X_0 \in B, X_1 \in A).$$

So the joint distribution of (X_0, X_1) is the same as that of (X_1, X_0). In other words, reversing the Markov chain $(X_t)_{t\geq 1}$ does not change its distribution, which explains why we say P is reversible. We will see that the majority of MCMC algorithms are reversible.

Definition 2.3. Let π be the stationary distribution of P. Suppose π has a density with respect to a dominating measure μ; denote it by $\pi(x) = (\mathrm{d}\pi/\mathrm{d}\mu)(x)$. Suppose P also has a density p with respect to μ; that is, $P(x, B) = \int_B p(x, y)\mu(\mathrm{d}y)$ for any $x \in \mathcal{X}, B \in \mathcal{B}(\mathcal{X})$. We say P satisfies a detailed balance condition, if for any $x, y \in \mathcal{X}$,

$$\pi(x)p(x, y) = \pi(y)p(y, x). \tag{3}$$

Lemma 2.2. *If* (3) *holds, then P is reversible with respect to π.*

Proof. This directly follows from the definition. □

Exercise 2.1. Let t be a positive integer. Clearly, P^t is also a transition kernel. Prove:

(a) If π is a stationary distribution of P, then it is also a stationary distribution of P^t.

(b) If π is a stationary distribution of P^t, then P also has a stationary distribution.

2.3 Construction of Metropolis–Hastings Algorithms

Let the state space $\mathcal{X}$ and target distribution π be given. We now consider how to construct a Markov chain that is easy to simulate and has stationary distribution π. To begin with, let us fix a"reference" transition kernel Q. We can interpret Q as a Markov chain moving "randomly" on the space $\mathcal{X}$, and in most cases, Q is chosen such that each step of this chain is small (with high probability). For example, if $\mathcal{X} = \mathbb{R}$, we can let $Q(x, \cdot)$ be a normal distribution with mean x and variance σ^2. If $\mathcal{X}$ is the node set of an undirected graph, we can let $Q(x, \cdot)$ be the uniform distribution on the set of nodes connected to x. The choice of Q is almost arbitrary; in particular, $Q(x, \cdot)$ can depend on π. But to be able to implement the sampling algorithm we will develop, for each x, $Q(x, \cdot)$ needs to be a distribution that we know how to sample from (that is, we know how to simulate a Markov chain with kernel Q).

Of course, Q probably does not have π as the stationary distribution. So let's modify the dynamics of this chain using the idea of rejection sampling. If the current state is $X_t = x$, we draw $Y \sim Q(x, \cdot)$ but do not necessarily "accept" this proposal. Instead, we calculate an acceptance probability, denoted by $\alpha(x, y)$, where y is the realized value of Y. We set $X_{t+1} = y$ only with probability $\alpha(x, y)$, and we set $X_{t+1} = x$ with probability $1 - \alpha(x, y)$ (i.e., stay at the previous state). Denote the resulting transition kernel by P. For any set B such that $x \notin B$, we have

$$P(x, B) = \int_{y \in B} \int_{u \in [0,1]} \mathbb{1}_{[0,\alpha(x,y)]}(u) \mathrm{d}u \, Q(x, \mathrm{d}y) = \int_B \alpha(x, y) Q(x, \mathrm{d}y).$$

Hence, for any $x \neq y$, we can write $P(x, \mathrm{d}y) = \alpha(x, y)Q(x, \mathrm{d}y)$. If B may contain the state x, we can write

$$P(x, B) = \int_B \alpha(x, y) Q(x, \mathrm{d}y) + \mathbb{1}_B(x) \int_{\mathcal{X}} (1 - \alpha(x, y)) Q(x, \mathrm{d}y).$$

From now on, we assume that $Q(x, \mathrm{d}y) = q(x, y)\mu(\mathrm{d}y)$ and $\pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x)$. Then $P(x, \cdot)$ has a density with respect to $\mu + \delta_x$ (where δ_x denotes the Dirac measure assigning probability one to x), and we can write

$$P(x, \mathrm{d}y) = \alpha(x, y)q(x, y)\mu(\mathrm{d}y) + \left\{ \int_{\mathcal{X}} (1 - \alpha(x, z))q(x, z)\mu(\mathrm{d}z) \right\} \delta_x(\mathrm{d}y). \tag{4}$$

For $x \neq y$, we have transition density $p(x, y) = \alpha(x, y)q(x, y)$. By Lemmas 2.1 and 2.2, if α is chosen such that

$$\pi(x)\alpha(x, y)q(x, y) = \pi(y)\alpha(y, x)q(y, x), \quad \forall x, y \in \mathcal{X}, \tag{5}$$

then P has π as a stationary distribution. Recall that we interpret α as the acceptance probability, so we have one more constraint that α has to be always in $[0, 1]$. Still, there are infinitely many choices of α. Two simple choices that have been often considered in the literature are

$$\alpha^*(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right\}, \quad \alpha_{\mathrm{B}}(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)}. \tag{6}$$

The subscript B in the second choice stands for Barker, since the resulting Metropolis–Hastings algorithm is also known as Barker's dynamics [1].

We summarize this derivation in the following theorem. We say P in Theorem 2.1 is the transition kernel of a Metropolis–Hastings algorithm with proposal Q.

Theorem 2.1. *Let Q be a transition kernel on $\mathcal{X}$ with density q (with respect to the dominating measure μ). Let P be a transition kernel defined by (4), where α is some function taking values in $[0,1]$ and satisfies (5); in particular, α can be one of the two choices given in (6). Then P is reversible with respect to π and thus has π has a stationary density.*

Example 2.1. We now illustrate the use of Metropolis–Hastings algorithms using a toy example. Let $\mathcal{X} = \{1, 2, \ldots, p\}$ with $p = 10$, and define $\pi(x) \propto 1/x$ for each x. Since $\mathcal{X}$ is discrete, we will always take counting measure as the dominating measure whenever talking about densities. Let the density of the proposal kernel Q be given by

$$q(x, p \wedge (x+1)) = q(x, 1 \vee (x-1)) = \frac{1}{2}.$$

(All other moves have proposal probability zero.) To run the Metropolis–Hastings algorithm, we simulate a Markov chain $(X_t)_{t\geq 0}$ with kernel P as described in Theorem 2.1 and $\alpha(x, y) = 1 \wedge (\pi(y)/\pi(x))$. Note that the proposal probabilities are always canceled out.

First, let's verify that the distribution of X_t converges to π in total variation distance, i.e., $\lim_{t\to\infty} \|\delta_{x_0} P^t - \pi\|_{\mathrm{TV}} = 0$, where x_0 denotes the initial value. We choose $x_0 = p$ and run the algorithm 10^4 times. Then we numerically calculate $\|\hat{P}^t(x_0, \cdot) - \pi\|_{\mathrm{TV}}$, where $\hat{P}^t(x_0, \cdot)$ is the empirical distribution of X_t out of the 10^4 replicates. The result is shown in the left panel of Figure 1 (note that the y-axis is shown on log scale). The blue line in the plot is obtained from linear regression, with t as the predictor and $\log\|\hat{P}^t(x_0, \cdot) - \pi\|_{\mathrm{TV}}$ as the response. It is clear that the total variation distance goes to zero at an exponential rate. Second, we check that the empirical distribution of $(X_1, X_2, \ldots, X_t)$ also converges to π as $t \to \infty$ (think about why). Again, we let $x_0 = p$ and run the algorithm only once for 10^3 iterations. The decay of the total variation distance between π and the distribution of $(X_1, X_2, \ldots, X_t)$ is shown in the right panel of Figure 1.

2.4 Asymptotic Variances and Peskun Theorem

Let $(X_t)_{t\geq 0}$ be a Markov chain with stationary distribution π. To estimate $\pi(f)$, we can use

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i). \tag{7}$$

Intuitively, the variance of this estimator reflects how fast the chain converges to π. The next definition formalizes this idea.

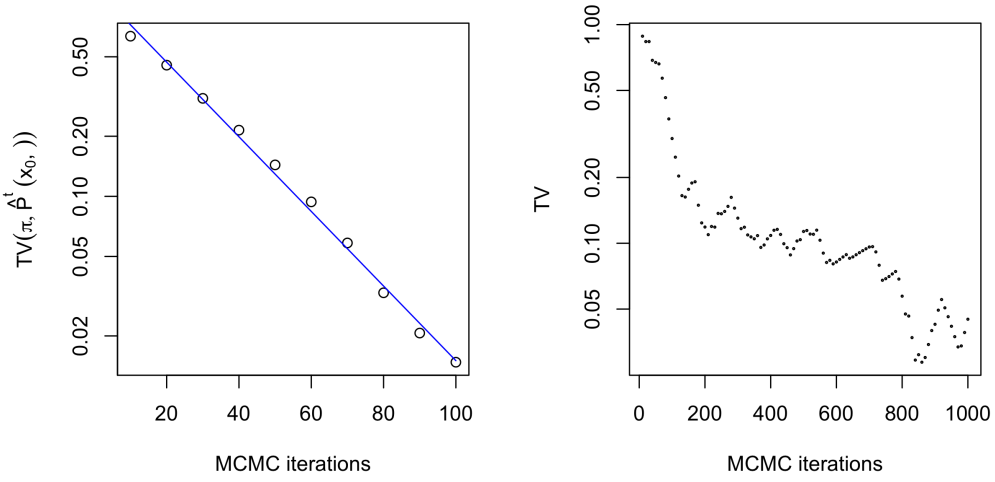

Figure 1: Convergence to π in Example 2.1.

Definition 2.4. Let P be a transition kernel reversible with respect to π, and let $f\colon \mathcal{X} \to \mathbb{R}$ be such that $\pi(f) = 0$ and $\pi(f^2) < \infty$. Define

$$\sigma_f^2(P) = \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{i=1}^{n} f(X_i) \right),$$

where $(X_t)_{t\geq 0}$ is a Markov chain with kernel P and $X_0 \sim \pi$. We say $\sigma_f^2(P)$ is the asymptotic variance of $\hat{\pi}_n(f)$ defined in (7).

Remark 2.1. Under mild conditions on P and f, we have the CLT: $\sqrt{n}\hat{\pi}_n(f)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma_f^2(P)$; see, e.g. [4, 2] for technical details.

Definition 2.5. Let P_1, P_2 be two transition kernels reversible with respect to π. We write $P_1 \succeq P_2$ if $\sigma_f^2(P_1) \leq \sigma_f^2(P_2)$ for any f such that $\pi(f) = 0$ and $\pi(f^2) < \infty$.

We now state a very important result due to Peskun and Tierney [5, 6]; it is often known as Peskun ordering of Markov chains.

Theorem 2.2. *Let P_1, P_2 be transition kernels reversible with respect to π. Then, $P_1 \succeq P_2$ if*

$$P_1(x, B \setminus \{x\}) \geq P_2(x, B \setminus \{x\}), \quad \forall x \in \mathcal{X},\ B \in \mathcal{B}(\mathcal{X}).$$

In Section 2.3, we have seen that the acceptance probability $\alpha(x, y)$ in Metropolis–Hastings schemes can take many forms, and now Theorem 2.2 tells us which one to use.

Exercise 2.2. Fix the stationary distribution π and proposal kernel Q. Let P_α denote the Metropolis–Hastings kernel defined by (4). Prove that $P_{\alpha^*} \succeq P_{\alpha'}$ where α^* is as given in (6), and $\alpha'(x, y)$ is any function that takes values in $[0, 1]$ and satisfies (5).

Example 2.2. Let's compare the two choices of α given in (6) for Example 2.1. Let $\theta = \sum_{i=1}^{p} x\,\pi(x)$ (for $p = 10$, $\theta = 3.414$), and we can estimate it using $\hat{\theta}_t = t^{-1}\sum_{i=1}^{t} X_i$. Denote by $\hat{\theta}_t^*$ and $\hat{\theta}_t^{\mathrm{B}}$ the estimators obtained from the Metropolis–Hastings algorithm with acceptance probability α^* and that with acceptance probability α_{B}, respectively. This time we initialize $X_0 \sim \pi$ and still run the algorithm 10^4 times. Then we numerically calculate the standard deviation of $\hat{\theta}_t$ across 10^4 replicates, and we plot it against t in Figure 2.

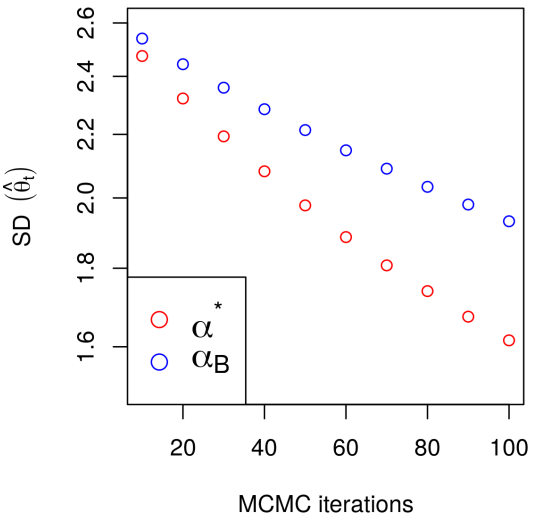

Figure 2: Standard deviation of the estimator $\hat{\theta}_t$ in Example 2.2.

References

[1] Anthony Alfred Barker. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

[2] Olle Häggström and Jeffrey Rosenthal. On variance conditions for Markov chain CLTs. *Electronic Communications in Probability*, 12:454–464, 2007.

[3] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[4] Galin L Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299, 2004.

[5] Peter H Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

[6] Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of applied probability*, pages 1–9, 1998.