Unit 6: More Examples of MCMC Sampling Schemes

6.1 Core Principles

Consider a target probability distribution $\pi(x)$ defined on $\mathcal{X}$, which may be difficult to sample from for various reasons (e.g. multimodality, non-log-concavity, double intractability). The key idea underlying most of the more sophisticated MCMC sampling schemes is to augment the state space and consider another target probability distribution $\tilde{\pi}(x, y)$ defined on $\mathcal{X} \times \mathcal{Y}$. Let $(X_t, Y_t)_{t \geq 1}$ be a Markov chain with stationary distribution $\tilde{\pi}(x, y)$. If $\tilde{\pi}(x, y)$ satisfies any of the following three conditions, $(X_t, Y_t)_{t \geq 1}$ can be used to approximate $\pi(x)$:

(i) $\int \tilde{\pi}(x, y) \mathrm{d}y = \pi(x)$ for every $x \in \mathcal{X}$.

(ii) $\mathcal{Y}$ is finite, and $\tilde{\pi}(x, y_0) = c\,\pi(x)$ for some $y_0 \in \mathcal{Y}$, $c > 0$ and every $x \in \mathcal{X}$.

(iii) there exists another distribution $\bar{\pi}(x, y)$ such that (a) $\int \bar{\pi}(x, y) \mathrm{d}y = \pi(x)$ for every x, and (b) $\bar{\pi}(x, y)/\tilde{\pi}(x, y)$ is known or easy to evaluate up to a normalizing constant.

If $\tilde{\pi}$ satisfies (i), then we simply ignore the samples $(Y_t)_{t \geq 1}$ and use $(X_t)_{t \geq 1}$ to approximate $\pi(x)$. If $\tilde{\pi}$ satisfies (ii), then we only use $\{X_t \colon Y_t = y_0\}$. If $\tilde{\pi}$ satisfies (iii), then we can use importance sampling. In this unit, we focus on examples that satisfy (i) or (ii). In Remark 6.1, we will see one example of the third case, with more examples and detailed discussion to follow in the next unit.

6.2 Examples

Two examples we have already studied are pseudo-marginal MCMC and proximal sampling; both augment the state space and target some $\tilde{\pi}(x, y)$ satisfying condition (i) above. Below we give more examples.

Example 6.1 (slice sampling)**.** We assume that $\pi(x) = Cf(x)$ for some unknown constant C and known function f. Let $\mathcal{Y} = (0, \infty)$ and define

$$\tilde{\pi}(x, y) = C \mathbb{1}_{(0, f(x))}(y).$$

Then $\tilde{\pi}$ has $\pi(x)$ as a marginal distribution since

$$\int_0^\infty \tilde{\pi}(x, y) \mathrm{d}y = \int_0^\infty C \mathbb{1}_{(0, f(x))}(y) \mathrm{d}y = Cf(x).$$

Slice sampling proposed by [11] is an MCMC sampling scheme targeting $\tilde{\pi}(x, y)$. Its transition kernel can be written as $P = P_1 P_2$ where P_2 updates y from the conditional distribution $\tilde{\pi}_{Y \,|\, X}(\cdot \,|\, x)$, which is just the uniform distribution on $(0, f(x))$. The kernel P_1 can be any kernel that updates x and is invariant with respect to the conditional distribution $\tilde{\pi}_{X \,|\, Y}(\cdot \,|\, y)$, which is the uniform distribution on

$$S_y = \{x \in \mathcal{X} \colon y < f(x)\};$$

the set S_y is called a "slice." For example, if $\mathcal{X} = \mathbb{R}$, we can let $P_1((x, y), \cdot)$ be the distribution corresponding to the following procedure where $w > 0$ is a fixed constant:

(1) Draw u from $\mathrm{Unif}(0, 1)$.

(2) Set $I = (x - wu,\, x + w(1 - u)) \cap S_y$.

(3) Draw x' uniformly from I (e.g. by rejection sampling).

(4) Return (x', y).

More efficient (and complicated) schemes are described in [11]. A similar idea has been utilized in [8] for improving the Hamiltonian Monte Carlo sampler.

Example 6.2 (Swendsen–Wang algorithm)**.** Let $G = (V, E)$ be an undirected graph with node set V and edge set E (since edges are undirected, we can assume that if $(i, j) \in E$, then $(j, i) \notin E$). Define a probability distribution π on $\{0, 1\}^V$ by

$$\pi(x) \propto \exp\left(\beta \sum_{(i,j)\in E} \mathbb{1}_{\{x_i = x_j\}}\right) = \prod_{(i,j)\in E} e^{\beta \mathbb{1}_{\{x_i = x_j\}}},$$

where $\beta > 0$ is a constant. This is known as the Ising model. Introduce an auxiliary variable $U \in \mathbb{R}^E$, and define a joint distribution by

$$\tilde{\pi}(x, u) \propto \prod_{(i,j)\in E} \mathbb{1}\{0 < u_{i,j} < e^{\beta \mathbb{1}_{\{x_i = x_j\}}}\}. \tag{1}$$

This is just a multivariate version of the construction of $\tilde{\pi}$ in slice sampling, and thus $\tilde{\pi}$ has $\pi(x)$ as the marginal distribution.

Define another auxiliary variable $Y \in \{0, 1\}^E$ by $y_{i,j} = \mathbb{1}_{\{u_{i,j} > 1\}}$. Swendeson–Wang algorithm [12] is the Gibbs sampler targeting the joint distribution of X, Y under $\tilde{\pi}$. Given $X = x$, $\{Y_{i,j} \colon (i, j) \in E\}$ are independent, and

$$\tilde{\pi}(Y_{i,j} = 0 \,|\, X = x) = \begin{cases} 1, & \text{if } x_i \neq x_j, \\ e^{-\beta}, & \text{if } x_i = x_j. \end{cases}$$

Consider the conditional distribution of X given $Y = y$. If $y_{i,j} = 1$, we must have $x_i = x_j$. If $y_{i,j} = 0$, which means $u_{i,j} \in (0, 1)$, the indicator function in (1) is 1 regardless of whether $x_i = x_j$, and thus all possible values of x are equally likely. So we can sample X from the conditional distribution given $Y = y$ using the following procedure:

(1) Let $H = (V, E)$ be a copy of G. For every edge $(i, j) \in E$, remove it from H if $y_{i,j} = 0$. Denote the resulting edge set of H by E_H.

(2) Identify the connected components of H, i.e., sub-graphs $H_1 = (V_1, E_1), \ldots, H_n = (V_n, E_n)$ such that $\{V_1, \ldots, V_n\}$ is a partition of V (i.e., V_k's are disjoint and their union equals V) and $\{E_1, \ldots, E_n\}$ is a partition of E_H.

(3) For each connected component $H_k = (V_k, E_k)$, with probability $1/2$, set $x_i = 0$ for every $i \in V_k$; with probability $1/2$, set $x_i = 1$ for every $i \in V_k$.

For generalization of Swendeson–Wang algorithm, see, e.g.,[2].

Example 6.3 (Hamming ball sampler)**.** Let $\mathcal{X} = \{0,1\}^d$. Fix an integer $m > 0$, and for each x, let

$$B_m(x) = \{x' \in \mathcal{X} \colon \|x - x'\|_1 \le m\}$$

denote the Hamming ball centered at x with radius m. Let C denote number of points in $B(x)$, which is a constant independent of x. The Hamming ball sampler [13] is the Gibbs sampling scheme targeting the joint distribution

$$\tilde{\pi}(x, y) = C^{-1}\pi(x)\mathbb{1}_{B_m(x)}(y).$$

Note that $\sum_y \tilde{\pi}(x,y) = C^{-1}\pi(x)\sum_y \mathbb{1}_{B_m(x)}(y) = \pi(x)$. The conditional distribution $\tilde{\pi}_{Y\,|\,X}(\cdot\,|\,x)$ is simply the uniform distribution on $B_m(x)$. The conditional distribution $\tilde{\pi}_{X\,|\,Y}(\cdot\,|\,y)$ is given by

$$\tilde{\pi}_{X\,|\,Y}(x\,|\,y) \propto \pi(x)\mathbb{1}_{B_m(x)}(y) = \pi(x)\mathbb{1}_{B_m(y)}(x).$$

Exact sampling from $\tilde{\pi}_{X\,|\,Y}(\cdot\,|\,y)$ requires us to evaluate $\pi(x)$ for each $x \in B_m(y)$. This is possible as long as d^m is not too large.

Example 6.4 (simulated tempering)**.** Let $\mathcal{Y} = \{0, 1, \dots, K\}$ be finite and choose a sequence of constants $1 = \tau_0 < \tau_1 < \cdots < \tau_K < \infty$. Define

$$\tilde{\pi}(x, y) \propto \kappa_y \pi(x)^{1/\tau_y},$$

where $(\kappa_y)_{y=0}^K$ is another sequence of positive constants. When $y = 1$, the conditional distribution of X coincides with $\pi(x)$. A larger value of τ makes the distribution $\pi(x)^{1/\tau}$ flatter, and we often refer to τ as the "temperature." Simulated tempering is a Metropolis–Hastings algorithm targeting $\tilde{\pi}$ [10]. Let $Q(x, \cdot)$ denote the proposal scheme one would use to sample from $\pi(x)$. Then, we construct the proposal scheme for $\tilde{\pi}$ as follows:

(1) with probability $\rho \in (0,1)$, we propose (x', y) where x' is drawn from the proposal distribution $Q(x, \cdot)$;

(2) with probability $1 - \rho$, we propose (x, y') where $y' = y \pm 1$ with each option having probability 0.5 (if $y' < 0$ or $y' > K$, the proposal is immediately rejected).

When making a proposal of the first type, the acceptance probability is given by

$$\alpha((x,y),(x',y)) = \min\left\{1, \left(\frac{\pi(x')}{\pi(x)}\right)^{1/\tau_y} \frac{q(x',x)}{q(x,x')}\right\}.$$

For the proposal of the second type,

$$\alpha((x,y),(x,y')) = \min\left\{1, \frac{\kappa_{y'}}{\kappa_y}\pi(x)^{\frac{1}{\tau_{y'}} - \frac{1}{\tau_y}}\right\}.$$

So the acceptance probability highly depends on the constants (κ_y), the tuning of which is often a major challenge to the application of simulated tempering. Another important question is how to choose the temperature ladder $(\tau_y)_{y=0}^K$. A popular approach is to use a geometric sequence, i.e., $\tau_y = \beta\tau_{y-1}$ for some fixed $\beta > 1$. There is a large body of literature on how to choose (τ_y) and (κ_y); see, e.g., [5, 1, 9].

Remark 6.1. Let $(X_t, Y_t)_{t\geq 1}$ be samples from the simulated tempering algorithm. To estimate $\pi(f) = \int f(x)\pi(x)\mathrm{d}x$, we can of course just use samples $\{X_t \colon Y_t = 0\}$ since they should asymptotically follow the distribution $\pi(x)$. However, note that the other samples can be utilized as well by importance sampling. Indeed, let $(\eta_y)_{y=0}^K$ be another sequence of positive constants, and let's define

$$\bar{\pi}(x, y) \propto \eta_y \pi(x),$$

which, trivially, has $\pi(x)$ as the marginal. We can then calculate the importance weight of each sample (X_t, Y_t) by

$$\frac{\bar{\pi}(x, y)}{\tilde{\pi}(x, y)} \propto \frac{\eta_y}{\kappa_y}\pi(x)^{-1/\tau_y}.$$

How to choose (η_y), however, is a difficult question, and was studied in [6].

Example 6.5 (parallel tempering)**.** Parallel tempering, also known as Metropolis-coupled MCMC or replica exchange algorithm, is very similar to simulated tempering. The major difference is that we now run $K + 1$ chains in parallel at temperatures $\tau_0, \tau_1, \ldots, \tau_K$ instead of running one single chain with temperature dynamically adjusted. So let's define a target distribution $\tilde{\pi}$ on $\mathcal{X}^{K+1}$ by

$$\tilde{\pi}(x_0, x_1, \ldots, x_K) \propto \prod_{k=0}^{K} \pi(x_k)^{1/\tau_k}.$$

To construct an MCMC algorithm targeting $\tilde{\pi}$, a simple scheme is to update each x_k in turn using a Metropolis–Hastings step with the proposal scheme Q. In this case, the $K+1$ coordinates will just evolve independently of each other. But the motivation of parallel tempering, similar to simulated tempering, is to improve the mixing by letting chains at different temperatures exchange information. To achieve this, let's introduce another updating scheme. Let $(x_0, \ldots, x_K)$ be the current state.

(1) Choose $i < j$ uniformly from $\{0, 1, \ldots, K\}$.

(2) Propose to swap x_i with x_j. Denote the resulting state by

$$x' = (x_0, \ldots, x_{i-1}, x_j, x_{i+1}, \ldots, x_{j-1}, x_i, x_{j+1}, \ldots, x_K).$$

(3) Accept x' with probability

$$\alpha(x, x') = \min\left\{1, \frac{\tilde{\pi}(x')}{\tilde{\pi}(x)}\right\} = \min\left\{1, \left(\frac{\pi(x_i)}{\pi(x_j)}\right)^{\frac{1}{\tau_j}-\frac{1}{\tau_i}}\right\}.$$

Example 6.6 (non-reversible Metropolis–Hastings)**.** The last example is special in the sense that it does not fit within the Metropolis–Hastings or Gibbs frameworks we have discussed so far. It is an example of non-reversible MCMC algorithms, which has gained increasing popularity [14, 3]. Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = \{+,-\}$. Define

$$\tilde{\pi}(x,+) = \tilde{\pi}(x,-) = \frac{1}{2}\pi(x).$$

Define

$$\mathcal{N}_+(x) = \{x' \in \mathcal{X} \colon \|x - x'\| = 1, \|x'\| = \|x\| + 1\},$$
$$\mathcal{N}_-(x) = \{x' \in \mathcal{X} \colon \|x - x'\| = 1, \|x'\| = \|x\| - 1\}.$$

Let $Q((x,+),\cdot)$ be the uniform distribution on $\{(x',+) \colon x' \in \mathcal{N}_+(x)\}$; so $Q_+(x,\cdot)$ flips a coordinate currently equal to 0. Similarly, let $Q((x,-),\cdot)$ be the uniform distribution on $\{(x',-) \colon x' \in \mathcal{N}_-(x)\}$. If there is nothing to propose, then the chain stays at the current state. We now present an algorithm targeting $\tilde{\pi}$ which looks similar to the Metropolis–Hastings algorithm.

(1) Suppose the current state is $(x,+)$. Propose $(x',+)$ from $Q((x,+),\cdot)$.

(2) Calculate the acceptance probability

$$\alpha((x,+),\,(x',+)) = \min\left\{1, \frac{\pi(x')q((x',-),(x,-))}{\pi(x)q((x,+),\,(x',+))}\right\}.$$

(3) With probability $\alpha((x,+),\,(x',+))$, we move to $(x',+)$; with probability $1-\alpha((x,+),\,(x',+))$, we move to $(x,-)$.

The updating scheme at $(x,-)$ follows analogously, with $+$ and $-$ swapped. Such an algorithm is called non-reversible Metropolis–Hastings or lifted Metropolis–Hastings [4]. The term "non-reversible" is self-explanatory: the chain may move from $(x,+)$ to $(x',+)$ for $x' \in \mathcal{N}_+(x)$ but can never move from $(x',+)$ to $(x,+)$.

Theorem 6.1. *The algorithm presented in Example 6.6 has stationary distribution $\tilde{\pi}(x,y)$.*

Proof. Let $p((x,y),\,(x',y'))$ denote the transition density of the algorithm. Fix some $z \in \mathcal{X}$, and it suffices to show that

$$\sum_{x,y} \tilde{\pi}(x,y)p((x,y),(z,+)) = \tilde{\pi}(z,+).$$

By the definition of $\tilde{\pi}$, we need to prove that

$$\sum_x \pi(x)\left\{p((x,+),(z,+)) + p((x,-),(z,+))\right\} = \pi(z).$$

Observe that $p((x,+),(z,+))$ is nonzero only if $x \in \mathcal{N}_-(z)$, and $p((x,-),(z,+))$ is nonzero only if $x = z$. Indeed,

$$\begin{aligned}\sum_x \pi(x)p((x,+),(z,+)) &= \sum_{x\in\mathcal{N}_-(z)} \pi(x)p((x,+),(z,+)) \\ &= \sum_{x\in\mathcal{N}_-(z)} \pi(x)q((x,+),\,(z,+))\alpha((x,+),\,(z,+)),\end{aligned}$$

and

$$\begin{aligned}\sum_x \pi(x)p((x,-),(z,+)) &= \pi(z)p((z,-),(z,+)) \\ &= \sum_{x\in\mathcal{N}_-(z)} \pi(z)q((z,-),\,(x,-))\left[1-\alpha((z,-),\,(x,-))\right] \\ &= \pi(z) - \sum_{x\in\mathcal{N}_-(z)} \pi(z)q((z,-),\,(x,-))\alpha((z,-),\,(x,-)).\end{aligned}$$

It is easy to verify that

$$\pi(x)q((x,+),\,(z,+))\alpha((x,+),\,(z,+)) = \pi(z)q((z,-),\,(x,-))\alpha((z,-),\,(x,-)),$$

from which the conclusion follows. □

Exercise 6.1. Let $S\colon \mathcal{X} \to \mathcal{X}$ be an invertible mapping such that $S(x) = S^{-1}(x)$ for every x and $\pi(S(A)) = \pi(S^{-1}(A)) = \pi(A)$ for every $A \in \mathcal{B}(\mathcal{X})$. Suppose a transition kernel satisfies the skew detailed balance [7]:

$$\pi(x)p(x,y) = \pi(S(y))p(S(y),S(x)).$$

Then π is a stationary distribution of P.

References

[1] Yves F Atchadé, Gareth O Roberts, and Jeffrey S Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, 2011.

[2] Adrian Barbu and Song-Chun Zhu. Generalizing Swendsen–Wang for image analysis. *Journal of Computational and Graphical Statistics*, 16(4):877–900, 2007.

[3] Joris Bierkens. Non-reversible Metropolis-Hastings. *Statistics and Computing*, 26(6):1213–1228, 2016.

[4] Philippe Gagnon and Arnaud Doucet. Nonreversible jump algorithms for Bayesian nested model selection. *Journal of Computational and Graphical Statistics*, 30(2):312–323, 2020.

[5] Charles J Geyer and Elizabeth A Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[6] Robert Gramacy, Richard Samworth, and Ruth King. Importance tempering. *Statistics and Computing*, 20(1):1–7, 2010.

[7] Gregory Herschlag, Jonathan C Mattingly, Matthias Sachs, and Evan Wyse. Non-reversible Markov chain Monte Carlo for sampling of districting maps. *arXiv preprint arXiv:2008.07843*, 2020.

[8] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[9] Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin Monte Carlo. *Advances in neural information processing systems*, 31, 2018.

[10] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

[11] Radford M Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

[12] Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters*, 58(2):86, 1987.

[13] Michalis K Titsias and Christopher Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, 112(520):1598–1611, 2017.

[14] Konstantin S Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.