Unit 7: Markov Chain Importance Sampling Methods

7.1 Core Principles

In Unit 1, we have introduced the standard importance sampling method, where i.i.d. samples are generated from a reference distribution. The central idea underlying the sampling methods we discuss in this unit is to extend importance sampling to Markov chain samples. The (self-normalized) importance sampling estimator still has the same form. The only difference is that the convergence rate of this estimator depends on the mixing rate of the Markov chain. Below we state this result formally for discrete-space Markov chains; see, e.g., [13, Lemma 2] for the proof and the corresponding central limit theorem. For continuous-space Markov chains, this result would require more technical assumptions (which are typically satisfied in MCMC applications).

Theorem 7.1 (Importance Sampling with Markov Chain Samples)**.** *Let $\pi, \tilde{\pi} > 0$ be two probability distributions defined on a finite state space $\mathcal{X}$. Let $(X_t)_{t\geq 0}$ be an irreducible Markov chain on $\mathcal{X}$ with stationary distribution $\tilde{\pi}$. Define $\pi(h) \coloneqq \sum_{x\in\mathcal{X}} h(x)\pi(x)$, and define*

$$\hat{\pi}_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w(X_i), \text{ where } w(x) = \frac{\pi(x)}{\tilde{\pi}(x)}.$$

Then, $\hat{\pi}_n(h) \overset{a.s.}{\to} \pi(h)$.

Note that importance sampling can be used together with state space augmentation. That is, if (X_t, Y_t) has stationary distribution $\tilde{\pi}(x, y)$, and $\bar{\pi}(x, y)$ is another distribution that satisfies $\int \bar{\pi}(x, y)\mathrm{d}y = \pi(x)$, then the importance weight is given by $\bar{\pi}(x, y)/\tilde{\pi}(x, y)$; see, e.g., Example 7.1.

Exercise 7.1. Assume $\pi, \tilde{\pi} > 0$ are defined on a finite space $\mathcal{X}$. Let $(X_t)_{t\geq 0}$ be an irreducible and aperiodic Markov chain on $\mathcal{X}$ with stationary distribution $\tilde{\pi}$. Let $h\colon \mathcal{X} \to \mathbb{R}$ be such that $\pi(h) = 0$, and define

$$\hat{\pi}_n(h) = \frac{\sum_{i=1}^{n} h(X_i)w(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

where $w(x) = \pi(x)/\tilde{\pi}(x)$. Prove that $\sqrt{n}\hat{\pi}_n(h)$ converges in distribution to a normal random variable. (Hint: use the remark below.)

Remark 7.1. For (X_t) considered in Exercise 7.1, it satisfies the following: for any $h\colon \mathcal{X} \to \mathbb{R}$ with $\tilde{\pi}(h) = 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \overset{\text{a.s.}}{\to} 0, \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(X_i) \overset{\text{D}}{\to} N(0, \sigma_h^2),$$

where $\sigma_h^2 \geq 0$ is called the asymptotic variance.

7.2 Two Examples

Example 7.1 (importance Sampling with Metropolis–Hastings)**.** Consider a Metropolis–Hastings algorithm with transition kernel P reversible with respect to $\pi(x)$. Denote its proposal kernel by Q with density $q(x, y)$ (the density of proposing y given x). Let $(X_t)_{t\geq 0}$ denote the samples generated from this Metropolis–Hastings algorithm, and let Y_t be the proposed state at the t-th iteration. Observe that $(X_t, Y_t)_{t\geq 0}$ is a bivariate Markov chain. Since (X_t) has stationary distribution π and Y_t is generated from $Q(X_t, \cdot)$, $(X_t, Y_t)_{t\geq 0}$ has stationary distribution

$$\tilde{\pi}(x, y) = \pi(x)q(x, y).$$

Define $\bar{\pi}(x, y) = \pi(x)\pi(y)$, which yields the importance weight

$$w(x, y) = \frac{\bar{\pi}(x, y)}{\tilde{\pi}(x, y)} = \frac{\pi(y)}{q(x, y)}.$$

Now consider $\pi(h) = \int h(y)\pi(\mathrm{d}y)$, which we want to estimate. We can treat it as the expectation of a bivariate function $\tilde{h}(x, y) = h(y)$ with respect to $\bar{\pi}$, since $\int \tilde{h}(x, y)\bar{\pi}(\mathrm{d}(x, y)) = \int h(y)\pi(\mathrm{d}x)\pi(\mathrm{d}y) = \pi(h)$. So by importance sampling, we can estiamte $\pi(h)$ by

$$\hat{\pi}(h) = \frac{1}{n}\sum_{t=1}^{n} h(Y_t)\frac{\pi(Y_t)}{q(X_t, Y_t)}.$$

Of course, in most cases π can only be computed up to a normalizing constant, in which case we should use self-normalized importance sampling estimator:

$$\hat{\pi}_{\mathrm{sn}}(h) = \frac{\sum_{t=1}^{n} h(Y_t)\frac{\pi(Y_t)}{q(X_t,Y_t)}}{\sum_{t=1}^{n}\frac{\pi(Y_t)}{q(X_t,Y_t)}}.$$

This estimator was formally proposed in a recent paper [8], but it was already mentioned in the early work of [1, Sec. 5]. Exactly the same reasoning was used to devise a more complicated importance sampling estimator in [9].

Example 7.2 (dynamic weighting)**.** We still let q denote the proposal density and define

$$r(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

In Metropolis–Hastings schemes, we accept a proposal y with probability $1 \wedge r(x, y)$. Now let's simulate a bivariate Markov chain $(X_t, W_t)_{t\geq 0}$ where

- X_{t+1} is generated from the proposal $Q(X_t, \cdot)$;
- $W_{t+1} = W_t\, r(X_t, X_{t+1})$.

That is, we simply accept every proposal but keep track of a weighting variable W_t. Now let's assume that Q is reversible with respect to $\tilde{\pi}(x)$, which means that

$$\tilde{\pi}(x)q(x,y) = \tilde{\pi}(y)q(y,x).$$

Fix $X_0 = x$ and fix W_0 to be a positive constant. So we can write $W_0 = c\,\pi(x)/\tilde{\pi}(x)$ for some $c > 0$. Now we have

$$W_1 = W_0\, r(X_0, X_1) = c\frac{\pi(X_0)}{\tilde{\pi}(X_0)}\frac{\pi(X_1)q(X_1,X_0)}{\pi(X_0)q(X_0,X_1)} = c\frac{\pi(X_1)}{\tilde{\pi}(X_1)}.$$

By induction, this shows that $W_t = c\,\pi(X_t)/\tilde{\pi}(X_t)$ for every t, and thus W_t gives the exact importance weight we need. This method is called dynamic weighting and still applies even if Q is not reversible; see [10, 5] for more sophisticated versions.

7.3 Importance Correction for Informed Proposals

7.3.1 A Simple Algorithm on Discrete Spaces

Let $\mathcal{X}$ be finite and let $\mathcal{N}_x \subset \mathcal{X}$ denote a set of neighboring states of x. Assume that $y \in \mathcal{N}_x$ whenever $x \in \mathcal{N}_y$. Assume $\pi > 0$. Consider the following algorithm.

Algorithm 7.1. Fix $X_0 \in \mathcal{X}$, and fix a function $b\colon (0,\infty) \to (0,\infty)$. In the t-th iteration with $X_{t-1} = x$:

(i) Calculate $w(x,y) = b(\pi(y)/\pi(x))$ for every $y \in \mathcal{N}_x$;

(ii) Draw X_t from $\mathcal{N}_x$ with probability proportional to $w(x,y)$.

Clearly, $(X_t)_{t\ge 0}$ generated from the above algorithm is a Markov chain with transition probability

$$p_b(x,y) = \frac{b\left(\frac{\pi(y)}{\pi(x)}\right)}{Z_b(x)}\mathbb{1}_{\mathcal{N}_x}(y), \qquad \text{where } Z_b(x) = \sum_{x'\in\mathcal{N}_x} b\left(\frac{\pi(x')}{\pi(x)}\right). \tag{1}$$

In most cases, we want to use a monotone nondecreasing function b so that this Markov chain can quickly move to high-posterior regions. The transition matrix described by (1) is also called an "informed proposal scheme" [11] since the transition probabilities depend on the local landscape of π. Of course, $(X_t)_{t\ge 0}$ probably does not have π as the stationary distribution. But it turns out that as long as b satisfies a property, we can use importance sampling to correct for the bias.

Definition 7.1. We say a function $b\colon (0,\infty) \to (0,\infty)$ is a balancing function if

$$b(r) = r\, b(r^{-1}), \quad \forall r > 0.$$

Examples of balancing functions include

$$b(r) = 1 + r, \quad b(r) = \frac{r}{1+r}, \quad b(r) = \sqrt{r}, \quad b(r) = 1 \wedge r, \quad b(r) = 1 \vee r.$$

Theorem 7.2. *Suppose b is a balancing function. Then $(X_t)_{t\geq 0}$ generated from Algorithm 7.1 has stationary distribution*

$$\pi_b(x) \propto \pi(x) Z_b(x).$$

Hence, the importance weight of x is given by $1/Z_b(x)$.

Proof. Try it yourself. □

So to estimate $\pi(h)$, one can pick any balancing function b, run Algorithm 7.1 and then compute the estimator

$$\hat{\pi}_n(h) = \frac{\sum_{i=1}^n h(X_i)/Z_b(X_i)}{\sum_{i=1}^n 1/Z_b(X_i)}. \tag{2}$$

An application of this method to variable selection, named tempered Gibbs sampler, was proposed in [12]; they used $b(r) = 1 + r$ and used "importance tempering" to refer to such an MCMC technique. The work of [7] considered modification of Metropolis–Hastings schemes using Algorithm 7.1, which corresponds to $b(r) = 1 \wedge r$ (see Section 7.3.3). The convergence rate of the estimator given in (2) for general unimodal target distributions was studied in [13], and it was found that $b(r) = \sqrt{r}$ tends to be more robust than other choices such as $b(r) = 1 + r$. More sophisticated versions of this algorithm have also been proposed; see [6, 2, 4], and one example is given below.

7.3.2 Importance Correction for Multiple-try Proposals

We have discussed in Unit 3 the multiple-try Metropolis algorithm. Conceptually a multiple-try proposal can be viewed as an informed proposal applied to a random neighborhood. It turns out that the importance tempering technique can applied as well, which leads to a rejection-free multiple-try algorithm [4] that is applicable to any state space.

Algorithm 7.2 (multiple-try importance tempering)**.** Let $\mathcal{X}$ be arbitrary and the proposal density $q(x, y)$ be given; suppose $q(x, y) > 0$ whenever $q(y, x) > 0$. Fix a balancing function b and a positive integer m (the number of tries). Given X_0, generate a set $\mathcal{S}_0$ consisting of m other states. In the t-th iteration with $X_{t-1} = x, \mathcal{S}_{t-1} = \mathcal{S}$:

(i) For every $y \in \mathcal{S}$, calculate

$$\alpha(x, y) = b\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right).$$

(ii) Calculate $Z(x, \mathcal{S}) = \sum_{y\in\mathcal{S}} \alpha(x, y)$.

(iii) Set the importance weight of $X_{t-1} = x$ to $1/Z(x, \mathcal{S})$.

(iv) Draw $X_t = x'$ from $\mathcal{S}$ with probability proportional to $\alpha(x, x')$.

(v) Draw $y'_1, \ldots, y'_{m-1}$ i.i.d. from the proposal distribution $Q(X_t, \cdot)$.

(vi) Set $\mathcal{S}_t = \{y'_1, \ldots, y'_{m-1}, x\}$.

To prove that this algorithm is correct, we view it as a bivariate Markov chain $(X_t, \mathcal{S}_t)$ and find that it is reversible with respect to

$$\tilde{\pi}(x, \mathcal{S}) \propto \pi(x) Z(x, \mathcal{S}) \prod_{y \in \mathcal{S}} q(x, y). \tag{3}$$

Then we define $\bar{\pi}(x, \mathcal{S}) \propto \pi(x) \prod_{y \in \mathcal{S}} q(x, y)$, which shows that $1/Z(x, \mathcal{S})$ is the importance weight we need. Note that the set $\mathcal{S}$ is used as an auxiliary variable.

Exercise 7.2. Show that Algorithm 7.2 is reversible with respect to (3).

7.3.3 A Remark on Metropolis–Hastings Schemes

Consider a Metropolis–Hastings algorithm with proposal $q(x, y)$ defined on a general state space. Let $(X_t)_{t \geq 0}$ be the samples generated from this sampler and let $(Y_t)_{t \geq 0}$ denote the accepted moves. That is, writing $Y_0 = X_0$ and $\tau(0) = 0$, we define for each $t \geq 1$,

$$Y_t = X_{\tau(t)}, \text{ where } \tau(t) = \min\{i > \tau(t-1) \colon X_i \neq Y_{t-1}\}.$$

If we want to estimate $\pi(h)$, we can use the estimator

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \frac{\sum_{i=1}^{T} h(Y_i) W_i}{\sum_{i=1}^{T} W_i} \tag{4}$$

where $T = \max\{i \colon \tau(i) \leq n\}$ and W_i is the number of iterations the chain has stayed on Y_i. Note that we can express W_i by $W_i = (\tau(i+1) \wedge (n+1)) - \tau(i)$. The second characterization of $\hat{\pi}_n(h)$ is the form of an importance sampling estimator, and we now justify that W_i indeed gives the correct importance weight (which is expected since otherwise Metropolis–Hastings schemes are biased).

Observe that (Y_t) is also a Markov chain with transition density $\tilde{p}$ such that for $y \neq x$,

$$\tilde{p}(x, y) = \frac{q(x, y)\alpha(x, y)}{Z(x)}, \quad Z(x) = \int_{z \neq x} q(x, z)\alpha(x, z) \mathrm{d}z,$$

where of course $\alpha(x, y)$ denotes the acceptance probability. We know that Metropolis–Hastings schemes are invariant with respect to π due to the detailed balance condition:

$$\pi(x) q(x, y) \alpha(x, y) = \pi(y) q(y, x) \alpha(y, x).$$

But this also immediately implies that $\tilde{p}$ is invariant with respect to $\tilde{\pi}(x) \propto \pi(x) Z(x)$. Hence, if we only collect the accepted moves, (Y_t), of the Metropolis–Hastings chain, then the exact importance weight of $Y_t = y$ is equal to $Z(y)^{-1}$. The last observation to make is that, given $Y_i = y$, W_i is a geometric random variable with success probability $Z(y)$, and thus the expectation of W_i equals $Z(y)^{-1}$. Hence, we can think of the Metropolis–Hastings algorithm as a Markov chain with transition density $\tilde{p}$ and stationary distribution $\tilde{\pi}$, and it uses geometric random variables to estimate the importance weight of each sample.

This importance sampling perspective is more general, since we can then improve the estimator (4) by reducing the variance of the importance weight estimate (i.e., W_i). For example, the algorithm we discussed in Section 7.3.1 simply calculates this importance weight exactly (which is possible on finite spaces). More sophisticated schemes on continuous spaces have also been developed; see, e.g., [1, 3].

References

[1] George Casella and Christian P Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[2] Sigeng Chen, Jeffrey S Rosenthal, Aki Dote, Hirotaka Tamura, and Ali Sheikholeslami. Sampling via rejection-free partial neighbor search. *Communications in Statistics-Simulation and Computation*, pages 1–29, 2023.

[3] Randal Douc and Christian P Robert. A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *The Annals of Statistics*, 39(1):261–277, 2011.

[4] Guanxun Li, Aaron Smith, and Quan Zhou. Importance is important: Generalized markov chain importance sampling methods. *arXiv preprint arXiv:2304.06251*, 2023.

[5] Jun S Liu, Faming Liang, and Wing Hung Wong. A theory for dynamic weighting in Monte Carlo computation. *Journal of the American Statistical Association*, 96(454):561–573, 2001.

[6] Samuel Power and Jacob Vorstrup Goldman. Accelerated sampling on discrete spaces with non-reversible Markov processes. *arXiv preprint arXiv:1912.04681*, 2019.

[7] Jeffrey S Rosenthal, Aki Dote, Keivan Dabiri, Hirotaka Tamura, Sigeng Chen, and Ali Sheikholeslami. Jump Markov chains and rejection-free Metropolis algorithms. *Computational Statistics*, pages 1–23, 2021.

[8] Daniel Rudolf and Björn Sprungk. On a Metropolis–Hastings importance sampling estimator. *Electronic Journal of Statistics*, 14(1):857–889, 2020.

[9] Ingmar Schuster and Ilja Klebanov. Markov chain importance sampling — a highly efficient estimator for MCMC. *Journal of Computational and Graphical Statistics*, pages 1–9, 2020.

[10] Wing Hung Wong and Faming Liang. Dynamic weighting in Monte Carlo and optimization. *Proceedings of the National Academy of Sciences*, 94(26):14220–14224, 1997.

[11] Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

[12] Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society Series B*, 81(3):489–517, 2019.

[13] Quan Zhou and Aaron Smith. Rapid convergence of informed importance tempering. In *International Conference on Artificial Intelligence and Statistics*, pages 10939–10965. PMLR, 2022.