$g(u) \geq 0$ for $u > u_0$. Assume $u_0 < 1/2$ and notice that then

$$\begin{aligned}\int_0^{1/2} g(u)g(1-u)\mathrm{d}u &= \int_0^{u_0} g(u)g(1-u)\mathrm{d}u + \int_{u_0}^{1/2} g(u)g(1-u)\mathrm{d}u \\ &\leq \int_0^{u_0} g(u)g(1-u_0)\mathrm{d}u + \int_{u_0}^{1/2} g(u)g(1-u_0)\mathrm{d}u \\ &\leq g(1-u_0)\int_0^1 g(u)\mathrm{d}u = 0.\end{aligned}$$

The case $u_0 = 1/2$ is easy, and if $u_0 > 1/2$, then we may use the proof above with $\tilde{g}(u) := -g(1-u)$. □

5.4 Control variates (*)

Definition 5.15 (Control variates)**.** Suppose $(X_k, W_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} \hat{p}$ with $X_k \sim p$ ($\mathbb{X}$-valued) and W_k is a zero-mean random number. Let $\beta \in \mathbb{R}$, then

$$I^{(n)}_{p,\mathrm{ctrl}}(f) := \frac{1}{n}\sum_{k=1}^{n}\big[f(X_k) + \beta W_k\big].$$

is an unbiased and strongly constent estimator of $\mathbb{E}_p[f(X)]$.

Example 5.16. Suppose that we are interested in estimation of $\mathbb{E}_p[f(X)]$, where p is $N(\mu, \sigma^2)$, but f is a complicated function. Then $X_k \sim N(\mu, \sigma^2)$ and we may use $W_k = X_k - \mu$ as a control variate.

Example 5.17. Suppose that $X_k = F^{-1}(U_k)$, where $U_k \sim \mathcal{U}(0,1)$. We can always use $W_k = U_k - 0.5$ as control variates.

Proposition 5.18. *We have the expression of the variance*

$$\mathrm{Var}\big(I^{(n)}_{p,ctrl}(f)\big) = \frac{1}{n}\big[\mathrm{Var}_p(f(X)) + \beta^2\mathrm{Var}(W_1) + 2\beta\mathrm{Cov}(f(X_1), W_1)\big].$$

If $\mathrm{Cov}(f(X_1), W_1) \neq 0$, it is possible (in principle) to choose β such that the variance is reduced.

Remark 5.19. Theoretically, the best value is

$$\beta_* = -\mathrm{Cov}(f(X_1), W_1)/\mathrm{Var}(W_1),$$

which leads into

$$\mathrm{Var}\big(I^{(n)}_{p,\mathrm{ctrl}}(f)\big) = \frac{1}{n}\big[(1 - \mathrm{Corr}(f(X_1), W_1)^2)\mathrm{Var}_p(f(X))\big].$$

Remark 5.20. The value β_* is often unknown, but β may be chosen as an empirical approximation of β_* based on preliminary simulation of (X_k, W_k). Finding suitable control variates is problem-specific.

6 Markov chain Monte Carlo

Up to this point, we have considered only methods based on i.i.d. random sequences. Sometimes it is useful to construct non-i.i.d. sequence $X_1, X_2, \ldots$ such that we can approximate as before

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \approx \mathbb{E}_p[f(X)].$$

In this section, we will focus on Markov chains like this.

Intuitively, X_k are going to be 'approximately from p' for large k and X_k will be 'approximately independent' of X_j if $|k-j|$ is large.

6.1 Recap of some Markov chain theory

We will restate some concepts and key results related to (time-homogeneous) Markov chains, which you may have seen in earlier courses[8]. We focus here on countable or finite $\mathbb{S}$.

Definition 6.1 (Markov chain)**.** The random variables $(X_k)_{k\geq 0}$ form a Markov chain, if for all $k \in \mathbb{N}$ and $x_0, \ldots, x_k \in \mathbb{S}$,

$$\mathbb{P}(X_k = x_k \mid X_0 = x_0, \ldots, X_{k-1} = x_{k-1}) = \mathbb{P}(X_k = x_k \mid X_{k-1} = x_{k-1}).$$

Definition 6.2 (Transition probability, initial distribution)**.** The *transition probability* or *transition matrix* P of a (time-homogeneous) Markov chain $(X_k)_{k\geq 0}$ on $\mathbb{S}$ consists of

$$P(x,y) = \mathbb{P}(X_{k+1} = y \mid X_k = x) \qquad \text{for all } k \in \mathbb{N} \text{ and } x, y \in \mathbb{S}.$$

The distribution of $(X_k)_{k\geq 0}$ is called *initial distribution* $\lambda(x) = \mathbb{P}(X_0 = x)$ for all $x \in \mathbb{S}$.

Recall that λ and P characterise the distribution of $(X_k)_{k\geq 0}$.

Taking λ as a row vector and P as a matrix (you can think of finite, but the same ideas work with countable case), then

$$(\lambda P)(x) = \sum_{y\in\mathbb{S}} \lambda(y)P(y,x) = \sum_{y\in\mathbb{S}} \mathbb{P}(X_1 = x, X_0 = y) = \mathbb{P}(X_1 = x).$$

This argument can be iterated to find out that $(\lambda \overbrace{P \cdots P}^{k \text{ times}})(x) = (\lambda P^k)(x) = \mathbb{P}(X_k = x)$.

Definition 6.3 (Invariant distribution)**.** If $\pi = \big(\pi(x)\big)_{x\in\mathbb{S}}$ is a p.m.f. on $\mathbb{S}$ taken as a row vector, and if

$$\pi P = \pi, \qquad (\text{that is, } (\pi P)(x) = \pi(x) \text{ for all } x \in \mathbb{S}),$$

then π is the *invariant* or *stationary distribution* of P.

8. MATA271 Stochastic Models.

Definition 6.4 (Irreducibility)**.** Markov chain, or equivalently its transition probability, is *irreducible* if for any $x, y \in \mathbb{S}$ there exists $n = n(x, y) \in \mathbb{N}$ such that

$$\mathbb{P}(X_n = y \mid X_0 = x) > 0.$$

We state the following well-known theorem without proof:

Theorem 6.5 (Markov chain strong law of large numbers)**.** *Suppose π is a p.m.f. on $\mathbb{S}$ and that P is an irreducible transition probability on $\mathbb{S}$ with invariant distribution π.*

Let $(X_k)_{k\geq 0}$ be a Markov chain with transition probability P and with any initial distribution, then for any $f : \mathbb{S} \to \mathbb{R}$ such that $\mathbb{E}_\pi[f(X)]$ is finite,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{n\to\infty} \mathbb{E}_\pi[f(X)] \qquad \textit{almost surely.}$$

For completeness, let us restate also convergence in distribution, which is often of considered instead of Theorem 6.5 in Markov chain theory.

Definition 6.6 (Periodicity, aperiodicity)**.** A Markov chain $(X_k)_{k\geq 0}$ is *periodic* with period $m \in \mathbb{N}$ if there exists a partition $S_0, \ldots, S_{m-1}$ of $\mathbb{S}$, where S_k are non-empty, such that

$$\mathbb{P}(X_n \in S_{(n \bmod m)} \mid X_0 \in S_0) = 1 \qquad \text{for all } n \in \mathbb{N}.$$

The chain is *aperiodic* if it is not periodic with any period $m \geq 2$.

Theorem 6.7. *Suppose P is irreducible and aperiodic, with invariant distribution π. If X_n is a Markov chain with transition probability P with any initial distribution,*

$$\mathbb{P}(X_n = x) \xrightarrow{n\to\infty} \pi(x) \qquad \textit{for any } x \in \mathbb{S}.$$

Remark 6.8. Usually in sampling, we are rather more interested in SLLN in Theorem 6.5, but in some cases Theorem 6.7 may be of interest as well. MCMC chains are rarely periodic, so we usually get Theorem 6.7 automatically. We shall not consider aperiodicity in detail further.

6.2 Reversibility

We shall consider next a Markov chain concept, which may not appear in a general course on Markov chain theory, but proves very useful in checking invariance in the MCMC context.

Definition 6.9 (Reversibility)**.** Suppose P is a transition probability and π is a p.m.f. on $\mathbb{S}$. If

$$\pi(x)P(x, y) = \pi(y)P(y, x) \qquad \text{for all } x, y \in \mathbb{S}, \tag{13}$$

then P is *reversible with respect to* π, or *π-reversible*. (Condition (13) is also known as the *detailed balance*.)

Proposition 6.10. *If P is π-reversible, then π is invariant for P.*

Proof. $(\pi P)(x) = \sum_y \pi(y)P(y,x) = \pi(x)\sum_y P(x,y) = \pi(x)$. □

Remark 6.11. The contrary does not hold true. That is, if π is invariant for P, it *does not imply* π-reversibility.

Remark 6.12. Suppose P is reversible with respect to π and $X_0 \sim \pi$. Then the joint distribution of (X_0, X_1) is symmetric,

$$\mathbb{P}(X_0 = x,\, X_1 = y) = \pi(x)P(x,y) = \pi(y)P(y,x) = \mathbb{P}(X_0 = y,\, X_1 = x).$$

In other words, $(X_0, X_1) \overset{d}{=} (X_1, X_0)$. This generalises to

$$(X_0, X_1, \ldots, X_n) \overset{d}{=} (X_n, X_{n-1}, \ldots, X_0),$$

which can be understood so that the Markov chain *initialised from the stationarity distribution* can be 'time-reversed' without affecting its distribution.

The reversibility can also be understood in terms of the 'backwards' transition probability being equal to the 'forward' transition probability (assuming again $X_0 \sim \pi$),

$$\begin{aligned}\mathbb{P}(X_0 = i \mid X_1 = j) &= \frac{\mathbb{P}(X_0 = i, X_1 = j)}{\mathbb{P}(X_1 = j)} = \frac{\pi(j)P(j,i)}{\pi(j)} \\ &= \mathbb{P}(X_1 = i \mid X_0 = j).\end{aligned}$$

6.3 The Metropolis-Hastings algorithm on discrete $\mathbb{X}$

Assume $\mathbb{X}$ is discrete and p is a p.m.f. on $\mathbb{X}$, and for each $x \in \mathbb{X}$ we have a proposal p.m.f. $q(x, \cdot)$ on $\mathbb{X}$ which we can draw samples from.

Algorithm 6.13 (Metropolis-Hastings)**.** Choose some initial value $X_0 \equiv x_0$ with $p(x_0) > 0$ and iterate for $k = 1, 2, \ldots$

(a) Generate $Y_k \sim q(X_{k-1}, \cdot)$.

(b) Generate $U_k \sim \mathcal{U}(0,1)$, and if $U_k \leq \alpha(X_{k-1}, Y_k)$ *accept* and set $X_k = Y_k$, otherwise *reject* and set $X_k = X_{k-1}$, where the *acceptance probability* α is defined as follows:

$$\alpha(x,y) := \begin{cases} \min\left\{1, \dfrac{p(y)}{p(x)}\dfrac{q(y,x)}{q(x,y)}\right\}, & p(x)q(x,y) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, for some function $f : \mathbb{X} \to \mathbb{R}$, report

$$I^{(n)}_{p,q,\mathrm{MH}}(f) := \frac{1}{n}\sum_{k=1}^{n} f(X_k)$$

as the Metropolis-Hastings approximation of $\mathbb{E}_p[f(X)]$.

Remark 6.14. In Algorithm 6.13,

(i) The distribution p is called the *target distribution*.

(ii) Unnormalised distributions $p_u(x) = Z_p p(x)$ and $q_u(x,y) = Z_q q(x,y)$ can be used, because

$$\frac{p_u(y)}{p_u(x)}\frac{q_u(y,x)}{q_u(x,y)} = \frac{Z_p p(y)}{Z_p p(x)}\frac{Z_q q(y,x)}{Z_q q(x,y)} = \frac{p(y)}{p(x)}\frac{q(y,x)}{q(x,y)}.$$

(iii) The accept-reject step (b) is implemented in practice by drawing $U_k \sim \mathcal{U}(0,1)$ and setting

$$X_k := \begin{cases} Y_k, & \text{if } U_k < \dfrac{p_u(Y_k)}{p_u(X_{k-1})}\dfrac{q_u(Y_k, X_{k-1})}{q_u(X_{k-1}, Y_k)} \\ X_{k-1}, & \text{otherwise.} \end{cases}$$

In many cases, it is easier (and numerically more stable) to compute

$$r_u(x,y) := \log p_u(y) + \log q_u(y,x) - \log p_u(x) - \log q_u(x,y),$$

and then accept if $U_k < \exp\big(r_u(X_{k-1}, Y_k)\big)$ and reject otherwise.

(iv) There is no need to define $\alpha(x,y)$ for $p(x)q(x,y) = 0$ in practice, because $p(X_{k-1})q(X_{k-1}, Y_k) = 0$ never occurs (almost surely).

Proposition 6.15. *The Metropolis-Hastings algorithm:*

(i) Defines a Markov chain on the support of p,

$$\mathbb{S} := \{x \in \mathbb{X} \,:\, p(x) > 0\}.$$

(ii) Has transition probability K given as

$$K(x,y) = q(x,y)\alpha(x,y) + \rho(x)\mathbf{1}\left(y = x\right), \quad x, y \in \mathbb{S},$$

where the probability of rejection $\rho(x)$ *can be given as*

$$\rho(x) = 1 - \sum_{y \in \mathbb{X}} q(x,y)\alpha(x,y).$$

Proof. The transition probability is straightforward to write. Let us then check that $X_n \in \mathbb{S}$. For any $x \in \mathbb{S}$ and $y \in \mathbb{X} \setminus \mathbb{S}$

$$\mathbb{P}(X_{n+1} = y \mid X_n = x) = q(x,y)\alpha(x,y) = 0,$$

because $\alpha(x,y) = 0$. This means $\mathbb{P}(X_{n+1} \in \mathbb{S}) = 1$ if $X_n \in \mathbb{S}$, and by definition, $X_0 = x_0 \in \mathbb{S}$. □

Proposition 6.16. *The Metropolis-Hastings transition probability K is reversible with respect to the target distribution p.*

Proof. Exercise. □

Now, Propositions 6.15 and 6.10 applied with Theorem 6.5 imply the strong consistency.