Monte Carlo Methods

Derek Li

Contents

Pseudorandom Numbers

We first generate an i.i.d. sequence $U_i \sim \text{Uniform}[0, 1]$.

Algorithm (Linear Congruential Generator/LCG)**.**

- Choose large positive integers m, a, and b.
- Start with a seed value x_0, e.g., the current time in milliseconds.
- Recursively, $x_n = (ax_{n-1} + b) \mod m$, i.e., x_n is the remainder when $ax_{n-1} + b$ is divided by m. Hence $0 \leqslant x_n \leqslant m - 1$.
- Let $U_n = \dfrac{x_n}{m}$, $\{U_n\}$ will seem to be approximately i.i.d. Uniform$[0, 1]$.

Note. We need m large so many possible values; a large enough that no obvious pattern between U_{n-1} and U_n; b to avoid short cycles of numbers. We want large period, i.e., number of iterations before repeat. One common choice: $m = 2^{32}, a = 69069, b = 23606797$.

Theorem. The LCG has full period (m) iff both $\gcd(b, m) = 1$, and every "prime or 4" divisor of m also divides $a - 1$.

Once we have $U_i \sim \text{Uniform}[0, 1]$, we can generate other distributions with transformations, using change of variable theorem.

Example. To make $X \sim \text{Uniform}[L, R]$, set $X = (R - L)U_1 + L$.

Example. To make $X \sim \text{Bernoulli}(p)$, set

$$X = \begin{cases} 1, & U_1 \leqslant p \\ 0, & U_1 > p \end{cases}$$

Example. To make $Y \sim \text{Binomial}(n, p)$, either set $Y = X_1 + \cdots + X_n$ where

$$X_i = \begin{cases} 1, & U_i \leqslant p \\ 0, & U_i > p \end{cases}$$

or set

$$Y = \max\left\{ j : \sum_{k=0}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} \leqslant U_1 \right\}$$

Generally, to make $P(Y = x_i) = p_i$ for some $x_1 < x_2 < \cdots$, where $p_i \geqslant 0$ and $\sum_i p_i = 1$, set

$$Y = \max\left\{ x_j : \sum_{k=1}^{j-1} p_k \leqslant U_1 \right\}$$

Example. To make $Z \sim \text{Exponential}(1)$, set $Z = -\ln(U_1)$. Generally, to make $W \sim \text{Exponential}(\lambda)$, set $W = \dfrac{Z}{\lambda} = \dfrac{-\ln(U_1)}{\lambda}$ so that W has density $\lambda e^{-\lambda x}$ for $x > 0$.

Example. If

$$X = \sqrt{2\ln\left(\frac{1}{U_1}\right)}\cos(2\pi U_2)$$
$$Y = \sqrt{2\ln\left(\frac{1}{U_1}\right)}\sin(2\pi U_2)$$

then $X, Y \sim \mathcal{N}(0,1)$ and $X \perp Y$.

Algorithm (Inverse CDF Method)**.**

- We want CDF $P(X \leqslant x) = F(x)$.
- For $0 < t < 1$, set $F^{-1}(t) = \min\{x; F(x) \geqslant t\}$ and $X = F^{-1}(U_1)$.
- $X \leqslant x$ iff $U_1 \leqslant F(x)$ and thus $P(X \leqslant x) = P(U_1 \leqslant F(x)) = F(x)$.

Monte Carlo Integration

We can rewrite an integral as an expectation and compute it with Monte Carlo.

Example. Estimate $I = \int_0^5 \int_0^4 g(x,y)\mathrm{d}y\mathrm{d}x$, where $g(x,y) = \cos(\sqrt{xy})$.

Solution. We have

$$\int_0^5 \int_0^4 g(x,y)\mathrm{d}y\mathrm{d}x = \int_0^5 \int_0^4 5 \cdot 4 \cdot g(x,y) \cdot \frac{1}{4}\mathrm{d}y\frac{1}{5}\mathrm{d}x = \mathbb{E}[20g(X,Y)]$$

where $X \sim \text{Uniform}[0,5]$ and $Y \sim \text{Uniform}[0,4]$. Hence, we let $X_i \sim \text{Uniform}[0,5]$ and $Y_i \sim \text{Uniform}[0,4]$ (all independent) and estimate I by

$$\frac{1}{M}\sum_{i=1}^{M} 20g(X_i, Y_i)$$

with standard error

$$\text{SE} = M^{-1/2}\text{SE}(20g(X_1,Y_1), \cdots, 20g(X_M,Y_M))$$

Example. Estimate $I = \int_0^1 \int_0^\infty h(x,y)\mathrm{d}y\mathrm{d}x$, where $h(x,y) = e^{-y^2}\cos(\sqrt{xy})$.

Solution. We have

$$\int_0^1 \int_0^\infty (e^y h(x,y))e^{-y}\mathrm{d}y\mathrm{d}x = \mathbb{E}[e^Y h(X,Y)]$$

where $X \sim \text{Uniform}[0,1]$ and $Y \sim \text{Exponential}(1)$ are independent.

Hence we estimate I by

$$\frac{1}{M}\sum_{i=1}^{M} e^{Y_i} h(X_i, Y_i)$$

where $X_i \sim$ Uniform$[0,1]$ and $Y_i \sim$ Exponential(1) (all independent).

Alternatively, we could write

$$\int_0^1 \int_0^\infty \frac{1}{5} e^{5y} h(x,y) \cdot 5e^{-5y} \mathrm{d}y \mathrm{d}x = \mathbb{E}\left[\frac{1}{5} e^{5Y} h(X,Y)\right]$$

where $X \sim$ Uniform$[0,1]$ and $Y \sim$ Exponential(5) are independent.

Note. We can choose different λ to estimate I and the one minimizes the standard error is the best choice.

Algorithm (Importance Sampling)**.** Suppose we want to evaluate $I = \displaystyle\int s(y)\mathrm{d}y$.

- We rewrite $I = \displaystyle\int \frac{s(x)}{f(x)} f(x)\mathrm{d}x$, where f is easily sampled from, with $f(x) > 0$ whenever $s(x) > 0$.
- Hence, $I = \mathbb{E}\left[\dfrac{s(X)}{f(X)}\right]$ where X has density f. Thus, we estimate $I \approx \dfrac{1}{M}\displaystyle\sum_{i=1}^M \frac{s(x_i)}{f(x_i)}$ where $x_i \sim f$.

Unnormalized Densities

Suppose $\pi(y) = cg(y)$ where we know g but do not know c or π. Hence,

$$c = \frac{1}{\displaystyle\int g(y)\mathrm{d}y}$$

which might be hard to compute.

Let

$$I = \int h(x)\pi(x)\mathrm{d}x = \int h(x)cg(x)\mathrm{d}x = \frac{\displaystyle\int h(x)g(x)\mathrm{d}x}{\displaystyle\int g(x)\mathrm{d}x}$$

where

$$\int h(x)g(x)\mathrm{d}x = \int \frac{h(x)g(x)}{f(x)} f(x)\mathrm{d}x = \mathbb{E}\left[\frac{h(X)g(X)}{f(X)}\right]$$

with $X \sim f$.

Hence,

$$\int h(x)g(x)\mathrm{d}x \approx \frac{1}{M}\sum_{i=1}^M \frac{h(x_i)g(x_i)}{f(x_i)}$$

if $\{x_i\} \overset{\text{i.i.d.}}{\sim} f$.

Similarly,

$$\int g(x)\mathrm{d}x \approx \frac{1}{M}\sum_{i=1}^M \frac{g(x_i)}{f(x_i)}$$

if $\{x_i\} \overset{\text{i.i.d.}}{\sim} f$.

Therefore,

$$I \approx \frac{\displaystyle\sum_{i=1}^{M} \frac{h(x_i)g(x_i)}{f(x_i)}}{\displaystyle\sum_{i=1}^{M} \frac{g(x_i)}{f(x_i)}}$$

Note. Since we take ratios of unbiased estimates, the resulting estimate is not unbiased, and its standard errors are less clear. But it is still consistent as $M \to \infty$.

Example. Compute $I = \mathbb{E}[Y^2]$ where Y has density $cy^3 \sin(y^4)\cos(y^5)\mathbf{1}_{0<y<1}$ where $c > 0$ is unknown.

Solution. Let $g(y) = y^3 \sin(y^4)\cos(y^5)\mathbf{1}_{0<y<1}$ and $h(y) = y^2$. Let $f(y) = 4y^3\mathbf{1}_{0<y<1}$. Then

$$I \approx \frac{\displaystyle\sum_{i=1}^{M} \sin(x_i^4)\cos(x_i^5)x_i^2}{\displaystyle\sum_{i=1}^{M} \sin(x_i^4)\cos(x_i^5)}$$

where $\{x_i\} \overset{\text{i.i.d.}}{\sim} U^{1/4}$.

Note. It is good to use same sample $\{x_i\}$ for both numerator and denominator since it is easier to compute and leads to smaller variance.

Rejection Sampler

Suppose $\pi(x) = cg(x)$ where we only know g but hard to sample from.

Algorithm (Rejection Sampling)**.** Suppose we want to sample $X \sim \pi$.

- We find easily-sampled density f and known $K > 0$ s.t.

  $$Kf(x) \geqslant g(x)$$

  for all x, i.e., $cKf(x) \geqslant \pi(x)$.

- We sample $X \sim f$ and $U \sim \text{Uniform}[0,1]$ (independent).
  - If $U \leqslant \dfrac{g(X)}{Kf(X)}$, then accept X (as a draw from π).
  - Otherwise, reject X and start over again.

Proof. Conditional on accepting, we have

$$P\left(X \leqslant y \,\middle|\, U \leqslant \frac{g(X)}{Kf(X)}\right) = \frac{P\left(X \leqslant y, U \leqslant \dfrac{g(X)}{Kf(X)}\right)}{P\left(U \leqslant \dfrac{g(X)}{Kf(X)}\right)}$$

for any $y \in \mathbb{R}$. Since $0 \leqslant \dfrac{g(x)}{Kf(x)} \leqslant 1$,

$$P\left(U \leqslant \frac{g(X)}{Kf(X)} \middle| X = x\right) = \frac{g(x)}{Kf(x)}$$

Hence, by the double expectation formula,

$$\begin{aligned} P\left(U \leqslant \frac{g(X)}{Kf(X)}\right) &= \mathbb{E}\left[P\left(U \leqslant \frac{g(X)}{Kf(X)} \middle| X\right)\right] = \mathbb{E}\left[\frac{g(X)}{Kf(X)}\right] \\ &= \int_{-\infty}^{\infty} \frac{g(x)}{Kf(x)} f(x) \mathrm{d}x = \frac{1}{K} \int_{-\infty}^{\infty} g(x) \mathrm{d}x \end{aligned}$$

Similarly, for any $y \in \mathbb{R}$,

$$\begin{aligned} P\left(X \leqslant y, U \leqslant \frac{g(X)}{Kf(X)}\right) &= \mathbb{E}\left[\mathbf{1}_{X \leqslant y} \mathbf{1}_{U \leqslant \frac{g(X)}{Kf(X)}}\right] = \mathbb{E}\left[\mathbf{1}_{X \leqslant y} P\left(U \leqslant \frac{g(X)}{Kf(X)} \middle| X\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}_{X \leqslant y} \frac{g(X)}{Kf(X)}\right] = \int_{-\infty}^{y} \frac{g(x)}{Kf(x)} f(x) \mathrm{d}x = \frac{1}{K} \int_{-\infty}^{y} g(x) \mathrm{d}x \end{aligned}$$

Therefore,

$$P\left(X \leqslant y \middle| U \leqslant \frac{g(X)}{Kf(X)}\right) = \frac{\frac{1}{K} \int_{-\infty}^{y} g(x) \mathrm{d}x}{\frac{1}{K} \int_{-\infty}^{\infty} g(x) \mathrm{d}x} = \int_{-\infty}^{y} \pi(x) \mathrm{d}x$$

□

Note. Probability of accepting may be very small so that we get very few samples.

Auxiliary Variable Approach

Suppose $\pi(x) = cg(x)$ and (X, Y) chosen uniformly under graph of g, i.e.,

$$(X, Y) \sim \text{Uniform}\{(x, y) \in \mathbb{R}^2 : 0 \leqslant y \leqslant g(x)\}$$

then $X \sim \pi$ since for $a < b$

$$P(a < X < b) = \frac{\int_a^b g(x) \mathrm{d}x}{\int_{-\infty}^{\infty} g(x) \mathrm{d}x} = \int_a^b \pi(x) \mathrm{d}x$$

Algorithm (Auxiliary Variable Rejection Sampling)**.** Suppose support of g contained in $[L, R]$ and $|g(x)| \leqslant K$.

- We sample $(X, Y) \sim \text{Uniform}([L, R] \times [0, K])$.
- We reject if $Y > g(X)$; otherwise accept as sample with $(X, Y) \sim \text{Uniform}\{(x, y) : 0 \leqslant y \leqslant g(x)\}$, where $X \sim \pi$.

Example. Suppose $g(y) = y^3 \sin(y^4) \cos(y^5) \mathbf{1}_{0<y<1}$. Then, $L = 0, R = 1, K = 1$. Hence, sample $X, Y \sim \text{Uniform}[0, 1]$ and keep X iff $Y \leqslant g(X)$.

Queueing Theory

Property. Consider a queue of customers and let $Q(t)$ be the number of people in queue at time $t \geqslant 0$. Suppose service times follow Exponential(μ) (mean μ^{-1}) and inter-arrival times follow Exponential(λ) ("M/M/1 queue"). Hence, $\{Q(t)\}$ is a Markov process. Moreover, if $\mu \leqslant \lambda, Q(t) \to \infty$ as $t \to \infty$; if $\mu > \lambda$, then $Q(t)$ converges in distribution as $t \to \infty$:

$$P(Q(t) = i) \to \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i, i = 0, 1, 2, \cdots$$

Markov Chain Monte Carlo (MCMC)

Suppose we have a complicated, high-dimensional density $\pi = cg$. We can define a Markov chain $X_0, X_1, \cdots$ in such a way that for large enough $n, X_n \approx \pi$, then we can estimate $\mathbb{E}_\pi(h) = \displaystyle\int h(x)\pi(x)\mathrm{d}x$ by

$$\mathbb{E}_\pi(h) \approx \frac{1}{M - B}\sum_{i=B+1}^{M} h(X_i)$$

where B is chosen large enough so $X_B \approx \pi$, and M is chosen large enough to get good Monte Carlo estimates.

Algorithm (Metropolis Algorithm/Random Walk Metropolis)**.**

- Choose some initial value X_0 (perhaps random).
- Given X_{n-1}, choose a proposal state $Y_n \sim \mathrm{MVN}(X_{n-1}, \sigma^2 I)$ for some fixed $\sigma > 0$.
- Let $A_n = \dfrac{\pi(Y_n)}{\pi(X_{n-1})} = \dfrac{g(Y_n)}{g(X_{n-1})}$ and $U_n \sim \mathrm{Uniform}[0, 1]$.
- If $U_n < A_n$, set $X_n = Y_n$ (i.e., accept); otherwise, set $X_n = X_{n-1}$ (i.e., reject).
- Repeat for $n = 1, 2, \cdots, M$.

Note. We can choose any X_0, but central ones best. We can also use an overdispersed starting distribution - choose X_0 randomly from some distribution that covers the important parts of the state space.

Example. Suppose $g(y) = y^3\sin(y^4)\cos(y^5)\mathbf{1}_{0<y<1}$ and we want to compute $\mathbb{E}_\pi(h)$ where $h(y) = y^2$. We can use Metropolis algorithm with proposal $Y \sim \mathcal{N}(X, 1)$.

MCMC Standard Error

We want to estimate the standard error from a single run, i.e.,

$$v = \mathrm{Var}\left[\frac{1}{M - B}\sum_{i=B+1}^{M} h(X_i)\right]$$

Let $\overline{h}(x) = h(x) - \mathbb{E}_\pi(h)$ so $\mathbb{E}_\pi(\overline{h}) = 0$. Assume B large enough that $X_i \approx \pi$ for $i > B$. For large $M - B$,

$$\begin{aligned}
v &\approx \mathbb{E}_\pi\left[\left(\frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i) - \mathbb{E}_\pi(h)\right)^2\right] = \mathbb{E}_\pi\left[\left(\frac{1}{M-B}\sum_{i=B+1}^{M} \overline{h}(X_i)\right)^2\right] \\
&= \frac{1}{(M-B)^2}[(M-B)\mathbb{E}_\pi[\overline{h}(X_i)^2] + 2(M-B-1)\mathbb{E}_\pi[\overline{h}(X_i)\overline{h}(X_{i+1})] + \cdots] \\
&\approx \frac{1}{M-B}[\mathbb{E}_\pi[\overline{h}(X_i)^2] + 2\mathbb{E}_\pi[\overline{h}(X_i)\overline{h}(X_{i+1})] + 2\mathbb{E}_\pi[\overline{h}(X_i)\overline{h}(X_{i+2})]\cdots] \\
&= \frac{1}{M-B}\left[\mathrm{Var}_\pi[h] + 2\mathrm{Cov}_\pi(h(X_i), h(X_{i+1})) + 2\mathrm{Cov}_\pi(h(X_i), h(X_{i+2})) + \cdots\right] \\
&= \frac{1}{M-B}\mathrm{Var}_\pi[h][1 + 2\mathrm{Corr}_\pi(h(X_i), h(X_{i+1})) + 2\mathrm{Corr}(h(X_i), h(X_{i+2})) + \cdots] \\
&:= \frac{1}{M-B}\mathrm{Var}_\pi(h)(\text{varfact}) = (\text{i.i.d. variance})(\text{varfact})
\end{aligned}$$

where

$$\text{varfact} = 1 + 2\sum_{k=1}^{\infty}\mathrm{Corr}_\pi(h(X_0), h(X_k)) = 1 + 2\sum_{k=1}^{\infty}\rho_k = 2\sum_{k=0}^{\infty}\rho_k - 1 = \sum_{k=-\infty}^{\infty}\rho_k$$

since $\rho_0 = 1$ and $\rho_{-k} = \rho_k$. We call it integrated autocorrelation time (ACT).

Note. To compute varfact, we do not sum over all k, but set some threshold. We can use R's built-in function `acf` with a good choice of `lag.max` parameter, or write own.

Justification of Metropolis Algorithm

Theorem. If Markov chain is irreducible, with stationarity probability density π, then for π-a.e. initial value $X_0 = x$, (a) if $\mathbb{E}_\pi(|h|) < \infty$, then $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} h(X_i) = \mathbb{E}_\pi(h) = \int h(x)\pi(x)\mathrm{d}x$; (b) if chain aperiodic, then $\lim_{n\to\infty} P(X_n \in S) = \int_S \pi(x)\mathrm{d}x$ for all $S \subseteq \mathcal{X}$.

Notation. $P(i,j) = P(X_{n+1} = j | X_n = i)$ (discrete case), or $P(x, A) = P(X_{n+1} \in A | X_n = x)$ (general case), and $\Pi(A) = \int_A \pi(x)\mathrm{d}x$.

Recall 1. Irreducible means having positive probability of eventually getting from anywhere to anywhere else. In discrete case: $\forall i, j \in \mathcal{X}$ (state space), $\exists n \in \mathbb{N}$ s.t. $P(X_n = j | X_0 = i) > 0$. (We only need to require this for states j s.t. $\pi(j) > 0$.) In general case: $\forall x \in \mathcal{X}, A \subseteq \mathcal{X}$ with $\Pi(A) > 0$, $\exists n \in \mathcal{N}$ s.t. $P(X_n \in A | X_0 = x) > 0$ (π-irreducible). Since usually $P(X_n = y | X_0 = x) = 0$ for all y, irreducibility is usually satisfied for MCMC.

Recall 2. Aperiodic means there are no forced cycles, i.e., there do not exist disjoint non-empty subsets $\mathcal{X}_1, \cdots, \mathcal{X}_d$ for $d \geqslant 2$ s.t. $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i, 1 \leqslant i \leqslant d-1$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$. In discrete case, it is equivalent to $\gcd\{n : p^n(i,i) > 0\} = 1$ for all i.

Note 1. This is true for virtually any Metropolis algorithm, e.g., it is true if $P(x, \{x\}) > 0$ for any one state $x \in \mathcal{X}$, or if positive probability of rejection.

Note 2. This is true if $P(x, \cdot)$ has positive density throughout S for all $x \in S$, for some $S \subseteq \mathcal{X}$ with $\Pi(S) > 0$, e.g., Normal proposals.

Note 3. But it is not quite guaranteed, e.g., $\mathcal{X} = \{0, 1, 2, 3\}, \pi$ uniform on $\mathcal{X}$, and $Y_n = X_{n-1} \pm 1 \mod 4$.

Recall 3. Stationary distribution means that if we start w.p. Π, and then run the Markov chain for one step, that we will still have the probabilities Π. In discrete case: if $X_0 \sim \pi$, i.e., $P(X_0 = i) = \pi(i)$ for all i, then also $\mathcal{X}_i \sim \pi$, i.e., $P(X_1 = j) = \pi(j)$ for all j. Since $P(X_1 = j) = \sum_{i \in S} P(X_0 = i, X_1 = j) = \sum_{i \in S} P(X_0 = i)P(i,j)$, then π is stationary if $\sum_{i \in S} \pi(i)P(i,j) = \pi(j)$ for all j.

Discrete Case

Assume for simplicity that $\pi(x) > 0$ for all $x \in \mathcal{X}$. Let $q(x,y) = P(Y_n = y | X_{n-1} = x)$ be the proposal distribution, e.g., $q(x,x+1) = q(x,x-1) = \frac{1}{2}$. Assume that q is symmetric, i.e., $q(x,y) = q(y,x)$ for all $x, y \in \mathcal{X}$, then if $\alpha(x,y)$ is the probability of accepting a proposed move from x to y, then

$$\alpha(x,y) = P(U_n < A_n | X_{n-1} = x, Y_n = y) = P\left(U_n < \frac{\pi(y)}{\pi(x)}\right) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

We compute for $i, j \in \mathcal{X}$ with $i \neq j$

$$P(i,j) = q(i,j)\alpha(i,j) = q(i,j)\min\left\{\frac{\pi(j)}{\pi(i)}\right\}$$

Hence, using the symmetry of q,

$$\pi(i)P(i,j) = q(i,j)\min\{\pi(i), \pi(j)\} = q(j,i)\min\{\pi(i), \pi(j)\} = \pi(j)P(j,i)$$

which still holds if $i = j$. It follows that the chain is time reversible, i.e., $\pi(i)P(i,j) = \pi(j)P(j,i), \forall i, j \in \mathcal{X}$.

Suppose $X_0 \sim \pi$, i.e., $P(X_0 = i) = \pi(i)$ for all $i \in \mathcal{X}$, then using reversibility, we have

$$P(X_1 = j) = \sum_{i \in \mathcal{X}} P(X_0 = i)P(i,j) = \sum_{i \in \mathcal{X}} \pi(i)P(i,j) = \sum_{i \in \mathcal{X}} \pi(j)P(j,i) = \pi(j)\sum_{i \in \mathcal{X}} P(j,i) = \pi(j)$$

i.e., $X_1 \sim \pi$. So the Markov chain preserves π, i.e., π is a stationary distribution, which is true for any Metropolis algorithm. It then follows from the theorem that as $n \to \infty$, $\mathcal{L}(X_n) \to \pi$, i.e., $\lim_{n\to\infty} P(X_n = i) = \pi(i)$ for all $i \in \mathcal{X}$.

It also follows that if $\mathbb{E}_\pi(|h|) < \infty$, then $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \mathbb{E}_\pi(h) = \int h(x)\pi(x)\mathrm{d}x$ (LLN).

General Continuous Case

Let $\mathcal{X}$ be the state space of all possible values. Usually $\mathcal{X} \subseteq \mathbb{R}^d$. Let $q(x,y)$ be the proposal density for y given x. Let $\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$ be probability of accepting a proposed move from x to y. Let $P(x,S) = P(X_1 \in S | X_0 = x)$ be the transition probability. Then if $x \notin S$,

$$P(x,S) = P(Y_1 \in S, U_1 < A_1 | X_0 = x) = \int_S q(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\mathrm{d}y$$

We write for $x \neq y$, $P(x, \mathrm{d}y) = q(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\mathrm{d}y$. Then for $x \neq y$,

$$\pi(x)P(x,\mathrm{d}y)\mathrm{d}x = q(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\mathrm{d}y\pi(x)\mathrm{d}x = q(x,y)\min\{\pi(x),\pi(y)\}\mathrm{d}y\mathrm{d}x = \pi(y)P(y,\mathrm{d}x)\mathrm{d}y$$

which is symmetric and follows that $\pi(x)P(x, \mathrm{d}y)\mathrm{d}x = \pi(y)P(y, \mathrm{d}x)\mathrm{d}y$ for all $x, y \in \mathcal{X}$. We write $\Pi(\mathrm{d}x)P(x, \mathrm{d}y) = \Pi(\mathrm{d}y)P(y, \mathrm{d}x)$, which is reversible.

Suppose $X_0 \sim \Pi$, i.e., we start in stationarity. Then

$$\begin{aligned} P(X_1 \in S) &= \int_{x\in\mathcal{X}} \pi(x)P(X_1 \in S|X_0 = x)\mathrm{d}x = \int_{x\in\mathcal{X}}\int_{y\in S} \pi(x)P(x, \mathrm{d}y)\mathrm{d}x \\ &= \int_{x\in\mathcal{X}}\int_{y\in S} \pi(y)P(y, \mathrm{d}x)\mathrm{d}y = \int_{y\in S} \pi(y)\mathrm{d}y = \Pi(S) \end{aligned}$$

i.e., $X_1 \sim \Pi$. So the chain preserves Π, i.e., Π is stationary distribution. Since the chain almost always is irreducible and aperiodic, then we can apply the relative theorem.

Examples

Example. $\mathcal{X} = \mathbb{Z}, \pi(x) = \dfrac{2^{-|x|}}{3}$ and $q(x, y) = \dfrac{1}{2}$ if $|x - y| = 1$; otherwise 0.

Solution. It is reversible since it is a Metropolis algorithm. π is stationary following from reversibility. It is aperiodic since $P(x, \{x\}) > 0$. It is irreducible since $\pi(x) > 0$ for all $x \in \mathcal{X}$ so can get from x to y in $|x - y|$steps. By theorem, probabilities and expectations converge to those of π.

Example. Same as above except $\pi(x) = 2^{-|x|-1}$ for $x \neq 0$ with $\pi(0) = 0$.

Solution. It is not irreducible since it cannot go from positive to negative.

Example. Same as above except $q(x, y) = \dfrac{1}{4}$ if $1 \leqslant |x - y| \leqslant 2$; otherwise 0.

Solution. It is irreducible since it can jump over 0 to get from positive to negative and back.

Example. Metropolis algorithm with $\mathcal{X} = \mathbb{R}$, $\pi(x) = Ce^{-x^6}$, and proposals $Y_n \sim \text{Uniform}[X_{n-1} - 1, X_{n-1} + 1]$.

Solution. It is reversible since it is Metropolis, and $q(x, y)$ is symmetric. π is stationary following from reversibility. It is aperiodic since $P(x, \{x\}) > 0$ for all $x \in \mathcal{X}$. It is irreducible since the n-step transitions $P^n(x, \mathrm{d}y)$ have positive density whenever $|y - x| < n$. By theorem, probabilities and expectations converge to those of π.

Example. Same as above except $\pi(x) = C_1e^{-x^6}(\mathbf{1}_{x<2} + \mathbf{1}_{x>4})$.

Solution. It is not irreducible since it cannot jump from $[4, \infty)$ to $(-\infty, 2]$ or back.

Example. Same as above except proposals are $Y_n \sim \text{Uniform}[X_{n-1} - 5, X_{n-1} + 5]$.

Solution. It is irreducible since it can jump from $[4, \infty)$ to $(-\infty, 2]$ or back.

Example. Same as above except proposals are $Y_n \sim \text{Uniform}[X_{n-1} - 5, X_{n-1} + 10]$.

Solution. It does not make sense since proposals are not symmetric, so it it not a Metropolis algorithm.

Metropolis-Hastings Algorithm

Metropolis algorithm works provided proposal distribution is symmetric, i.e., $q(x, y) = q(y, x)$. But if q is not symmetric, we should use Metropolis-Hastings algorithm.

Algorithm (Metropolis-Hastings Algorithm)**.** If we replace A_n by

$$A_n = \frac{\pi(Y_n)q(Y_n, X_{n-1})}{\pi(X_{n-1})q(X_{n-1}, Y_n)}$$

then the algorithm is valid even if q is not symmetric. We accept if $U_n < A_n$; otherwise reject.

Note. It requires $q(x, y) > 0$ iff $q(y, x) > 0$.

Example. Suppose $\pi(x_1, x_2) = C|\cos(\sqrt{x_1 x_2})|I(0 \leqslant x_1 \leqslant 5, 0 \leqslant x_2 \leqslant 4)$ and $h(x_1, x_2) = e^{x_1} + x_2^2$. The proposal distribution is $Y_n \sim \text{MVN}(X_{n-1}, \sigma^2(1+|X_{n-1}|^2)^2 I)$ (larger proposal variance if farther from center). Hence,

$$q(x, y) = C(1 + |x|^2)^{-2} \exp\left(-\frac{|y - x|^2}{2\sigma^2(1 + |x|^2)^2}\right)$$

then we can run Metropolis-Hastings algorithm.

Independence Sampler

We propose $\{Y_n\} \sim q(\cdot)$, i.e., $\{Y_n\}$ are i.i.d. from some fixed density q, independent of X_{n-1}, then we accept if $U_n < A_n$ where $U_n \sim \text{Uniform}[0, 1]$ and

$$A_n = \frac{\pi(Y_n)q(X_{n-1})}{\pi(X_{n-1})q(Y_n)}$$

which is the special case of the Metropolis-Hastings algorithm, where $Y_n \sim q(X_{n-1}, \cdots)$.

Note. If $q(y) = \pi(y)$, i.e., propose exactly from target density π, then $A_n = 1$, i.e., make great proposals and always accept them (i.i.d.).

Langevin Algorithm

We propose

$$Y_n \sim \text{MVN}\left(X_{n-1} + \frac{1}{2}\sigma^2 \nabla \ln \pi(X_{n-1}), \sigma^2 I\right)$$

which is the special case of the Metropolis-Hastings algorithm.

Componentwise (Variable-at-a-Time) MCMC

We propose to move just one coordinate at a time, leaving all the other coordinates fixed, then accept/reject with usual Metropolis rule or Metropolis-Hastings rule.

Note. We need to choose which coordinate to update each time:

1. Systematic-scan: $1, 2, \cdots, d, 1, 2, \cdots$.
2. Random-scan: choose from Uniform$\{1, 2, \cdots, d\}$ each time.

Note that one systematic-scan iteration corresponds to d random-scan iterations.

Bayesian Statistics

Example (Variance Components Model/Random Effects Model)**.** Suppose some population has overall unknown mean μ, population consists of K groups, and we observe $Y_{i1}, \cdots, Y_{iJ_i}$ from group i for $1 \leqslant i \leqslant K$. Assume $Y_{ij} \sim \mathcal{N}(\theta_i, W)$ (conditionally independent), where θ_i and W are unknown. Assume the different θ_i are linked by $\theta_i \sim \mathcal{N}(\mu, V)$ (conditionally independent), with μ and V unknown. We want to estimate some or all of $V, W, \mu, \theta_1, \cdots, \theta_K$. We can use prior distributions, e.g., (conjugate):

$$V \sim \mathrm{IG}(a_1, b_1), W \sim \mathrm{IG}(a_2, b_2), \mu \sim \mathcal{N}(a_3, b_3)$$

where they are independent, a_i, b_i are known constants, and $\mathrm{IG}(a, b)$ is the inverse Gamma distribution with density $\dfrac{b^a}{\Gamma(a)} e^{-b/x} x^{-a-1}$ for $x > 0$.

Combining the above dependencies, we can calculate the joint density and posterior distribution.

Gibbs Sampler

The proposal distribution for ith coordinate is equal to the full conditional distribution of that coordinate (according to π), conditional on the current values of all the other coordinates, which is a special case of componentwise Metropolis-Hastings algorithm.

Note. We can use either systematic or random scan, then we always accept.

Tempered MCMC

Suppose $\Pi(\cdot)$ is multi-modal, i.e., has distinct arts. Define a sequence $\Pi_1, \cdots, \Pi_m$ where $\Pi_1 = \Pi$ (cold) and Π_r is flatter for larger τ (hot). Proceed by defining a joint Markov chain (x, τ) on $\mathcal{X} \times \{1, \cdots, m\}$ with stationary distribution $\overline{\Pi}$ defined by

$$\overline{\Pi}(S \times \{\tau\}) = \frac{1}{m} \Pi_r(S)$$

The Markov chain should have both spatial moves (change x) and temperature moves (change τ). Then, only count those samples where $\tau = 1$.

Parallel Tempering

Parallel tempering is a.k.a. Metropolis-Coupled MCMC, or MCMCMC. Define a sequence $\Pi_1, \cdots, \Pi_m$ where $\Pi_1 = \Pi$ (cold) and Π_r is flatter for larger τ (hot). Use state space $\mathcal{X}^m$ with m chains, i.e., one chain for each temperature. So the state at time n is $X_n = (X_{n1}, \cdots, X_{nm})$ where $X_{n\tau}$ is at temperature τ. Stationary distribution is now $\overline{\Pi} = \Pi_1 \cdots \Pi_m$, i.e., $\overline{\Pi}(X_1 \in S_1, \cdots, X_m \in S_m) = \Pi_1(S_1) \cdots \Pi_m(S_m)$. Then we can update the chain $X_{n-1,\tau}$ at temperature τ (for each $1 \leqslant \tau \leqslant m$), by proposing e.g., $Y_{n,\tau} \sim \mathcal{N}(X_{n-1,\tau}, 1)$ and accepting w.p.

$$\min\left(1, \frac{\pi_\tau(Y_n, \tau)}{\pi_\tau(X_{n-1,\tau})}\right)$$

We can also choose temperatures τ and τ' and propose to swap the values $X_{n,\tau}$ and $X_{n,\tau'}$ and accept w.p.

$$\min\left(1, \frac{\pi_\tau(X_{n,\tau'})\pi_{\tau'}(X_{n,\tau})}{\pi_\tau(X_{n,\tau})\pi_{\tau'}(X_{n,\tau'})}\right)$$

Monte Carlo Optimization - Simulated Annealing

Simulated annealing is a general method to find highest mode of π. Mode of π is same as mode of a flatter or a more peaked version π_τ for any $\tau > 0$. So we ca use tempered MCMC but where $\tau = \tau_n \downarrow 0$, and thus π_{τ_n} becomes more and more concentrated at mode as $n \to \infty$.

MCMC Convergence Rates Theory

Definition. Suppose $\{X_n\}$ is the Markov chain on $\mathcal{X}$ with stationary distribution $\Pi(\cdot)$. Let $P^n(x, S) = P(X_n \in S | X_0 = x)$ be the probabilities for the Markov chain after n steps, when started at x. Let

$$D(x, n) = \|P^n(x, \cdots) - \Pi(\cdot)\| = \sup_{S \subseteq \mathcal{X}} |P^n(x, S) - \Pi(S)|$$

The chain is ***ergodic*** if $\lim_{n \to \infty} D(x, n) = 0$ for Π-a.e. $x \in \mathcal{X}$.

Theorem. If chain is irreducible and aperiodic with $\Pi(\cdot)$ stationary, then chain is ergodic, i.e., converges asymptotically to Π.

Definition. The chain is ***geometrically ergodic*** if there is $\rho > 1$ and $M : \mathcal{X} \to [0, \infty]$ which is Π-a.e. finite s.t. $D(x, n) \leqslant M(x)\rho^n$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$.

Theorem. CLT holds for $\dfrac{1}{n}\sum_{i=1}^{n} h(X_i)$ if chain is geometrically ergodic and $\mathbb{E}_\pi[|h|^{2+\delta}] < \infty$ for some $\delta > 0$.

Definition. The chain is ***uniformly ergodic*** if there is $\rho < 1$ and $M < \infty$ s.t. $D(x, n) \leqslant M\rho^n$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$.

Definition. A ***quantitative bound*** on convergence is an actual number n^* s.t. $D(x, n^*) < 0.01$ (say).

Theorem. If $P(x, \mathrm{d}y) \geqslant \delta\pi(\mathrm{d}y)$ for all $x, y \in \mathcal{X}$, then $D(x, n) \leqslant (1 - \delta)^n$.

Theorem. If state space is finite, and chain is irreducible and aperiodic, then the chain is always ergodic and also geometrically ergodic.

Theorem. RWM is geometrically ergodic essentially iff π has exponentially light tails, i.e., there are $a, b, c > 0$ s.t. $\pi(x) \leqslant ae^{-b|x|}$ whenever $|x| > c$.